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

Volume 3 Issue 1

January 2012

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

SAI

www.ijacsa.thesai.org

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

Editorial Preface

From the Desk of Managing Editor…

IJACSA seems to have a cult following and was a humungous success during 2011. We at The Science and Information Organization are pleased to present the January 2012 Issue of IJACSA.

While it took the radio 38 years and the television a short 13 years, it took the World Wide Web only 4 years to reach 50 million users. This shows the richness of the pace at which the computer science moves. As 2012 progresses, we seem to be set for the rapid and intricate ramifications of new technology advancements.

With this issue we wish to reach out to a much larger number with an expectation that more and more researchers get interested in our mission of sharing wisdom. The Organization is committed to introduce to the research audience exactly what they are looking for and that is unique and novel. Guided by this mission, we continuously look for ways to collaborate with other educational institutions worldwide.

Well, as Steve Jobs once said, Innovation has nothing to do with how many R&D dollars you have, it's about the people you have. At IJACSA we believe in spreading the subject knowledge with effectiveness in all classes of audience. Nevertheless, the promise of increased engagement requires that we consider how this might be accomplished, delivering up-to-date and authoritative coverage of advanced computer science and applications.

Throughout our archives, new ideas and technologies have been welcomed, carefully critiqued, and discarded or accepted by qualified reviewers and associate editors. Our efforts to improve the quality of the articles published and expand their reach to the interested audience will continue, and these efforts will require critical minds and careful consideration to assess the quality, relevance, and readability of individual articles.

To summarise, the journal has offered its readership thought provoking theoretical, philosophical, and empirical ideas from some of the finest minds worldwide. We thank all our readers for their continued support and goodwill for IJACSA. We will keep you posted on updates about the new programmes launched in collaboration.

We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJACSA provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

We regularly conduct surveys and receive extensive feedback which we take very seriously. We beseech valuable suggestions of all our readers for improving our publication.

Thank you for Sharing Wisdom and a very Happy New Year!

Managing Editor
IJACSA
Volume 3 Issue 1, January 2012
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Associate Editors

Reviewer Board Members

CONTENTS

Analysis and Selection of Features for Gesture Recognition Based on a Micro Wearable Device

Yinghui Zhou[1], Lei Jing[2], Junbo Wang[2], Zixue Cheng[2]
1. Graduate School of Computer Science and Engineering
2. School of Computer Science and Engineering
University of Aizu
Aizu-Wakamatsu, Japan

***Abstract*—More and More researchers concerned about designing a health supporting system for elders that is light weight, no disturbing to user, and low computing complexity. In the paper, we introduced a micro wearable device based on a tri-axis accelerometer, which can detect acceleration change of human body based on the position of the device being set. Considering the flexibility of human finger, we put it on a finger to detect the finger gestures. 12 kinds of one-stroke finger gestures are defined according to the sensing characteristic of the accelerometer.**

Feature is a paramount factor in the recognition task. In the paper, gestures features both in time domain and frequency domain are described since features decide the recognition accuracy directly. Feature generation method and selection process is analyzed in detail to get the optimal feature subset from the candidate feature set. Experiment results indicate the feature subset can get satisfactory classification results of 90.08% accuracy using 12 features considering the recognition accuracy and dimension of feature set.

Keywords—Internet of Things; Wearable Computing; Gesture Recognition; Feature analysis and selection; Accelerometer.

I. INTRODUCTION

Internet of Things (IoTs) has become a hot topic in the computer science field, which indicates that all objects in the environment like human, home appliances, building, and service equipment can be sensed, identified, even controlled via the internet. IoTs will promote many development of application system, such as health supporting system for elder. Many countries are facing a serious society issue of population ageing. One common trend is more and more elders living alone and less able to benefit from the care and supporting that might be available in a large household. Investigation from World Health Organization indicates, in Japan, the proportion of people living in 3-generation households has fallen from 46% in 1985 to 20.5% in 2006. Health care both physical and mental becomes an important problem in current society.

Health supporting system has been studied widely in recent years [1] [2]. Two kinds of main supporting way focus on speech-based communication and activity-based recognition. The former provides a direct and effective way to know users intension, which has been used in the hospital and household [3] [4]. However, voice signal is sensitive to environment sound such as a TV being on, so that sometimes hard to pick out useful speech signal made by a user. Even under certain circumstance, the user may too weak to make a voice to call for a help. Activity recognition provides an active and undisturbed way for elderly care. For example, if an elder person falls down, the system of falling recognition can send message automatically asking for help.

Among the human activities, finger gestures are the most flexible ones. In our daily life, most of works are performed by hands. Gesture recognition is significant for learning user behavior, realizing for device control, and getting user intention. In the paper, we designed a wearable device with an accelerometer to detect finger gestures. Based on the accelerometer characteristic, a variety of finger gestures are defined in a 3D space.

Particularly, gestures features are studied both in time domain and frequency domain considering the paramount importance of feature generation in recognition task. Each kind of candidate features and their combination are analyzed based on stepwise regression algorithm to form a feature vector for accurate gesture classification and computing complexity control.

The paper is arranged as follows. The section Ⅱ introduces the related work about activity recognition and feature analysis. The section Ⅲ outlines the prototype system of gesture detection, and gives the gestures definition and data collection. The section Ⅳ describes the feature analysis method in detail including features generation process and features selection algorithm based on stepwise regression. The section Ⅴ gives the experiment and evaluation on feature generation and selection. Conclusion is given in the last section.

II. RELATED WORKS

Two types of system are mainly used for activity recognition. One is fixed device-based recognition system, and the other is wearable device-based detection system.

Fixed device-based techniques have been applied widely into various fields by different devices such as camera, computer vision system, and so on [5] [6]. The kind of system provides an application way of no burden to users. However, some people do not like the way of being supervised, and some private spaces are inconvenient to be set a camera like in bathroom. Moreover, some issues have to be considered including whether environment factor is fit for monitoring or

not such as surrounding light and blind corner of a camera, image processing speed and delay, information loss of 3-D object projecting to a 2-D image, and confusion of multiple users in same background and so on [5].

With the development of micro-electrical technology, micro wearable devices are penetrating into our life. They can be attached on human body to obtain user information directly, typically using RFID and sensor. RFID technology is attractive for many applications since it can detect user's situations by simple ID and location information [7]. But it is difficult to detect motion. Wearable sensors have shown their capability for activity recognition. Some research set sensors in different position of body to detect human daily activities. In [8] the system sets five accelerometers on hip, wrist, arm, ankle, and thigh for classifying 20 daily activities such as walking, sitting & relaxing, brushing teeth, bicycling, etc, and got the recognition accuracy ranging from 41% to 97% for different activities. In [9] a large realistic data are collected from many different sensors (accelerometers, physiological sensors, environment sensors, etc.) to recognize 7 activities like lie, row, walk, etc. with accuracy of 80% over. However, to our knowledge, most of the researches seldom focus on the tiny activity recognition like finger gesture.

Finger is one of the most flexible body parts. Most of works in our daily lift are accomplished by it. Therefore, detection of finger gestures not only helps to know user current behavior, but also reflect user intension and carry on some operations. Data glove, as an interactive device worn on the hand to sense gestures, has been applied in the environment of virtual reality [12]. However, the gestural interface required user to wear a cumbersome device to connect with external. It is inflexible and inconvenient for daily operation.

Moreover, most of above researches have no explained on the process of features generation and selection while the features are of paramount importance in any recognition works. Based on different sensing device, some researches get features directly from time-varying signal [10] [11] and with frequency analysis [8] [9]. Some prefer to wavelet analysis to obtain both spectral and temporal information [13] [14]. However, it is not be illustrated why the features are necessary, if they can be substituted on others, what will happen if adding or deleting one of them.

In this paper, we introduce a wearable device in our previous research named Magic Ring. It can be used to detect 12 kinds of predefined one-stroke finger gestures based on a 3-axis accelerometer [15]. A verity of gesture features are extracted for classification evaluation. However, the process of feature selection is a lack. In this paper, we focus on the feature analysis method including feature generation, feature selection, and feature evaluation taking the wearable sensor and its receiver as a prototype of the final target devices.

III. Prototype Introduction and Data Collection

A. Prototype Structure

The system is a ring shape sensing device based on a 3-trial accelerometer, MMA7361L from Freescale Semiconductor, Inc.

In its two sensitive scales of ±1.5g and ±1.8g, ±1.5g is adopted to detect all predefined gestures. Excepting for the sensing unit, data processing unit is used for A/D conversion and simple digital signal processing; and transmitting unit is for acceleration data transmission and communication. The system can be worn a finger with no much disturbing to daily activity as shown in Fig. 1.

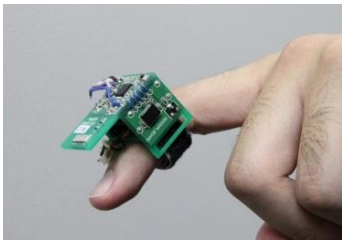


Figure 1. Sensing system on the finger

B. Gestures Definition

In the paper, the purpose of gesture recognition is to learn user simple intension and further to apply the gestures into daily life like controlling home appliances or calling for help. Therefore the gestures should be easy to be done for reducing physical load and easy to be understood and learned for reducing conscious load. Combining the way of controlling appliances and the characteristic of 3-axis accelerometer, 12 kinds of dynamic one-stroke gestures are designed. One-stroke refers to dynamic gestures which are performed no more than one degree of freedom in one direction. For example, "pushing" a button, "turning" a knob, and "pointing to" a picture can be regarded as one-stroke gestures.

The tri-axis accelerometer can detect the acceleration change of three directions in space as shown in Fig.2.

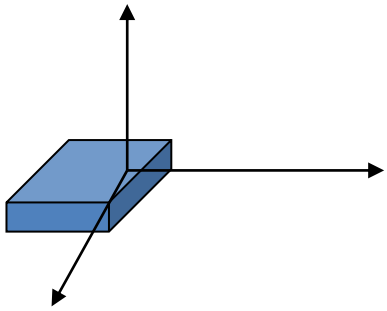


Figure 2. Sensing direction of a tri-axis accelerometer

Considering the tri-axis characteristic of the accelerometer and requirement of finger gestures, 12 kinds of one-stroke finger gestures are defined as shown in Fig. 3.

The 12 kinds of gestures can be divided into 3 pairs of gestures in X, Y and Z axis and 3 pairs of gestures in XY, YZ and XZ axial plane. These gestures are named as Crook and Unbend in X axis, Finger L-Shift and Finger R-Shift in Y axis, Finger Up and Finger Down in Z axis, Wrist L-shift and Wrist R-shift in XY plane, L-Rotate and R-Rotate in YZ plane, Wrist Up and Wrist Down in XZ plane. The modes of motion for the six pairs of gestures are shown in Fig. 3.

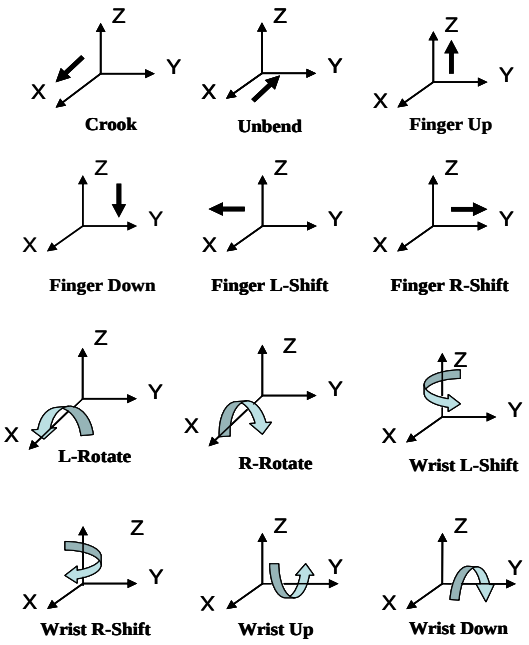


Figure 3. Modes of motion for the six pairs of gestures

C. Data Collection

The system is attached on the middle phalanx of forefinger since it is the most flexible in all fingers. Gestures data is collected as sampling 50Hz. The digital signal is stored in a PC for data analysis, features extraction and gestures classification.

20 students in the university (16 males and 4 females, average age 25.8±7.8) volunteered for the experiment of data collection under the supervision of a researcher. They are required to perform the predefined 12 gestures by forefinger in a natural and relax way. The gestures started in horizontal and static state, ended with static state. Each gesture was repeated 5 times per people and 100 times for 20 people totally.

IV. Features Analysis

For a finger gesture, it can be expressed quantitatively as a digital signal based on the sensing information. The features of the signal can indicate the type of a gesture and is useful for recognizing the gesture. A signal can be identified with various features. Therefore the features analysis is vital for identifying the signal. Roughly speaking, the more features are used, the higher accuracy may be achieved, but higher complexity the recognition is. However, not each feature can be used for distinguishing the gesture from others, e.g. for the two signals in Fig. 4, it is the acceleration change in X axis from two different kinds of finger gestures. The feature of signal energy or mean can distinguish them, but the peak value as a feature is failed to recognize one from the other.

Moreover, the number of features may not have direct relation with identification effects. Although it is different that each feature and their combinations contribute to recognition accuracy, that do not mean the more features are, the better recognition accuracy is.

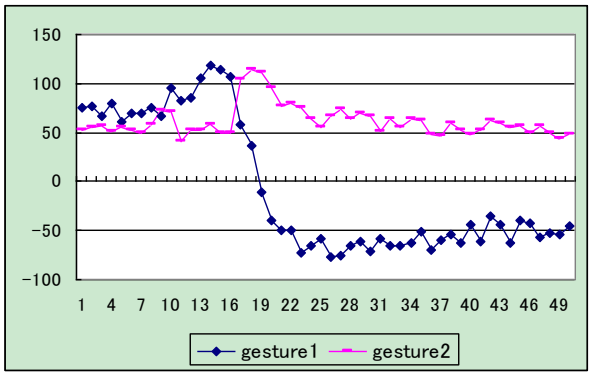


Figure 4. Two signals with same max value

Furthermore, the high dimensionality of features increases computing cost for some recognition algorithm. Therefore, analysis and selection of proper features of gestures to be recognized has to be performed to get the optimal feature vector/set for the balance between acceptable recognition rate and computing complexity.

The feature analysis process is, first, to extract signals of target gestures; second, based on the signals, generate candidate features; and finally, selection of proper features from the candidates to form feature set for recognition task. The paper, taking an accelerometer as an example, gives the feature analysis and selection procedure of finger gestures with an accelerometer.

A. Extraction of Gesture signal

Extraction of gesture signal refers to get the section related to the target gesture from successive signals, namely to detect start and end point of the gesture. Since the gesture signal shows a dynamic change trend from a static state to a dynamic activity, and then back to static, therefore the short time energy (STE) of signal is considered to distinguish the different states.

When STE in a sliding window is higher than a level, we think a gesture start to be performed, until the STE becomes lower than the level. We recorded the duration of each gesture per subject. Results show it roughly ranges from 200 to 800ms. Due to the sampling rate is 50Hz, the window size is compromised within 10 samples with 50% overlap between two continuous windows.

B. Feature Generation

Basically, feature can be divided into two types: features in time domain and frequency domain. Features in time domain show the signal characteristic varying with time. Typical features are shown in Fig. 5.

Frequency features are used to capture the periodic nature of a sensing signal for distinguishing some repetitive activities like walking and running. The typical frequency-domain feature is shown in Fig. 6.

Excepting for the features mentioned above, others can be extracted according to different classification objects like Euclidian distances, similarity, and so on.

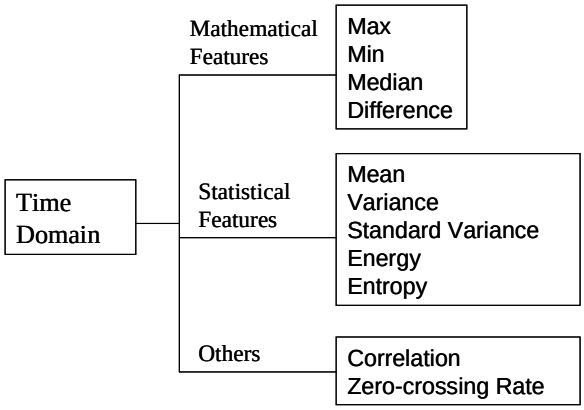


Figure 5. Typical time-domain features

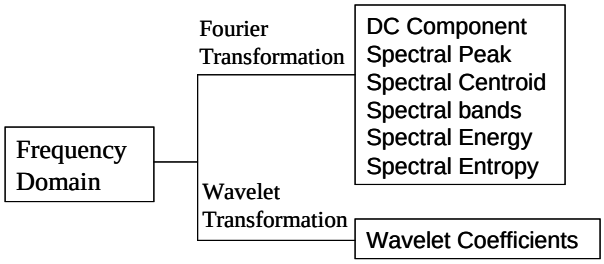


Figure 6. Typical frequency-domain features

All above mentioned features can be generated based on the corresponding mathematical or statistical method. However, not all features are necessary for a classification system. A direct reason is whether a feature may bring good classification significance. Even though both two features have good classification capability, there is maybe little gain when they are collected into a feature vector due to a high mutual correlation [16]. Another reason is the computational complexity. The number of features directly decides the dimensionality of classifier parameter. Thus a feature vector as small as possible is desired both in training process and in classifying process.

C. Feature Selection

The feature selection is very crucial, which helps us use as less as possible of features to find out as much as possible of classification information then to get the optimal recognition performance. Here we try to find an optimal feature vector to reach the balance between the acceptable recognition rate and computational complexity. In practical application, a satisfactory feature vector instead of an optimal vector.

In the paper, we adopt the algorithm of feature selection, stepwise regression. It is a greedy algorithm that includes a regression model in which candidate features are evaluated automatically. Forward selection and backward elimination are two main approaches to achieve the algorithm. The former represents the procedure starting with no features in the model, then trying to add one into the feature vector one by one until them reaching a "satisfactory significance". The latter is contrary to the former, which including all candidate features in the model, and deleting one that is no significant. Here, we used the forward selection way.

Specifically, let $F = \{f_1, f_2, \mathrm{K}, f_n\}$ be the candidate feature set of using for classification design. The elements in the set are descending order by the significance level. Let S_{f_i} denotes the significance level of feature f_i, for the feature set F, $S_{f_1} > S_{f_2} > \mathrm{K} > S_{f_n}$. We select a classifier as the model, then adding the feature one by one from f_1 to test the classification accuracy until reaching a satisfactory result.

V. Experiment and Evaluation

A. Features Generation

20 subjects completed the total 12 kinds of finger gestures under the supervision of researcher. Before each one-stroke finger gesture, it requires finger in a horizontal and static state. When finishing a gesture, finger should maintain ending state and stillness. In other words, the finger is dynamic just during a gesture being performed. Therefore, it is possible to identify if a gesture happening using a threshold-based approach.

Fig. 7 and Fig. 8 shows two finger gesture signals "Finger Left Shift" and "Finger Up", which is composed of three channels and indicates the acceleration change in three axis. From the extracted the gestures signals, various features can be generated both in time domain and frequency domain as described in last section. For example the mean and standard deviation (sd) of each axis in Fig. 7 and Fig. 8 can be computed as

$$mean = \frac{1}{n}\sum_{i=0}^{n-1} x_i, \quad sd = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n-1}(x_i - mean)}$$

where x denotes the sampling, and n denotes the number of sampling in the window. Other features can also be achieved by mathematical or statistical way.

Although any signal features can be regarded as a candidate, in order to reduce the computing load, some obvious insignificant features will be neglected. In our case, the finger gestures are one-stroke type, which means each gesture is aperiodic and instantaneous. Therefore, the features in frequency domain are neglected.

The process of feature generation can be expressed as:

(1) Observe the signals of one gesture from different subjects and try to describe it using some typical features. For example, in Fig. 7, we may consider mean, energy, etc.

(2) Observe the signals of variety of kinds of gestures to find out features with the capability of distinguishing with others like sd of Z axis in above two figures.

(3) Abandon some insignificant features like "amplitude" of each axis in our case, because even if a gesture is performed by same person, amplitude of each time will be great different due to the difference of performance speed.

(4) Collect the features to form a candidate feature set for feature selection.

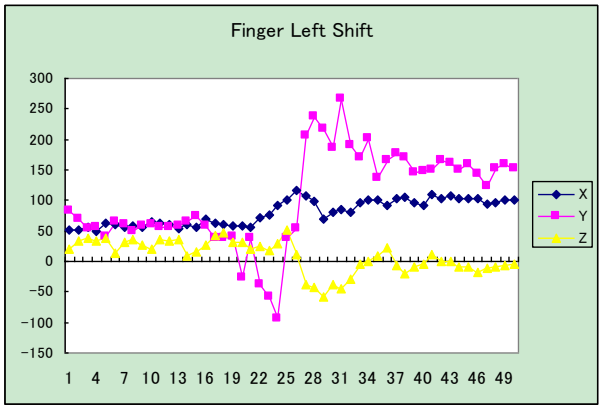


Figure 7. Acceleration curve of a finger gesture "Finger Left Shift"

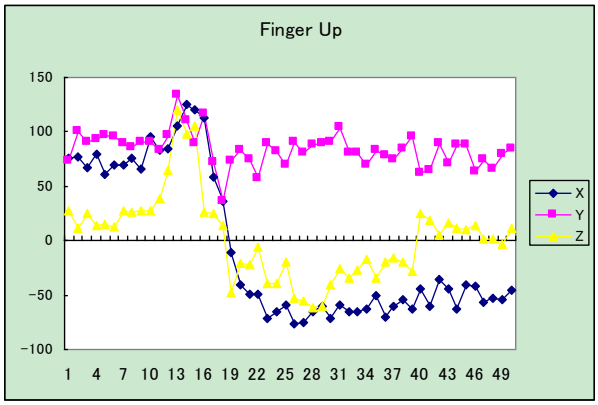


Figure 8. Acceleration curve of a finger gesture "Finger Up"

In our case, time-domain features of each axis are calculated including mean, standard deviation, energy, entropy, correlation of any two axes, difference of peak and valley, and position of peak and valley in the time axis, totally 8 kinds of features to compose a candidate feature set. Each feature in the set consists of three elements in X, Y, and Z axis, such as f_1 = {meanX, meanY, meanZ}.

B. *Features Selection*

Features selection is to find a satisfactory feature subset from the candidate feature set, so that to reach an optimal classification accuracy and computing complexity control. It is crucial since it decides the classification result directly. Forward selection algorithm of stepwise regression is adopted to test each feature and their combinations one by one. In the algorithm, a model is required to evaluate the features. For obtaining an objective evaluation results, here we select three basic classifiers of machine learning, C4.5 decision tree (C4.5), Nearest Neighbor (NN), and Naïve Bayes (NB), as three test models to calculate the classification accuracy of 12 kinds of one-stroke finger gestures mentioned above, and the testing average of three classifiers (Avg) is employed as the final evaluation result.

First, each one in the candidate feature set is tested by three models, the evaluation results are ranged in a descending order as shown in Table 1.

TABLE 1. THE RESULTS OF SINGLE FEATURE EVALUATION

	Features	C4.5	NN	NB	Avg
f_1	meanX, meanY, meanZ	53.67%	56.42%	51.08%	53.72%
f_2	energyX, energyY, energyZ	49.83%	51.92%	49.17%	50.31%
f_3	sdX, sdY, sdZ	48.17%	53.25%	46.50%	49.31%
f_4	diffX, diffY, diffZ	38.92%	41.58%	34.92%	38.47%
f_5	col(X,Y), col(X,Z), col(Z,Y)	34.42%	36.58%	38.08%	36.36%
f_6	posValX, posValY, posValZ	26.58%	28.28%	28.58%	27.81%
f_7	entropyX, entropyY, entropyZ	26.83%	23.25%	29.33%	26.47%
f_8	posPeakX, posPeakY, posPeakZ	26.33%	27.50%	22.75%	25.53%

Second, selecting the optimal feature from the Table 1, f_1, and combining with other features, the combinations will be recomputed based on the models. The results are shown in Table 2.

TABLE 2. THE EVALUATION RESULTS OF COMBINING TWO FEATURES

Features	C4.5	NN	NB	Avg
$f_1 + f_2$	63.83%	71.42%	55.92%	63.72%
$f_1 + f_3$	68.25%	76.83%	66%	70.36%
$f_1 + f_4$	61.17%	72.50%	58.92%	64.20%
$f_1 + f_5$	59.92%	69.58%	59%	62.83%
$f_1 + f_6$	63.50%	68.58%	63.75%	65.28%
$f_1 + f_7$	59.17%	62.75%	58.53%	60.15%
$f_1 + f_8$	63.08%	68.67%	65.25%	65.67%

Third, repeating the above process of selection to get the best one, $f_1 + f_3$, as a basic feature combination, and then combining with rest features to reevaluate the significance of combined features. From Table 3 to Table 7 show the evaluation results with different combination. Each optimal combination in each Table will be the basic of next combination.

TABLE 3. THE EVALUATION RESULTS OF COMBINING THREE FEATURES

Features	C4.5	NN	NB	Avg
$f_1 + f_3 + f_2$	66.92%	77.83%	63%	69.25%
$f_1 + f_3 + f_4$	68.42%	79.25%	63.75%	70.47%
$f_1 + f_3 + f_5$	73.50%	81.92%	72.33%	75.92%
$f_1 + f_3 + f_6$	78.50%	84.50%	78.67%	80.56%
$f_1 + f_3 + f_7$	67.75%	70.25%	66.67%	68.22%
$f_1 + f_3 + f_8$	76.58%	86.08%	74.83%	79.16%

TABLE 4. THE EVALUATION RESULTS OF COMBINING FOUR FEATURES

Features	C4.5	NN	NB	Avg
$f_1 + f_3 + f_6 + f_2$	76.92%	84.33%	76.42%	79.22%
$f_1 + f_3 + f_6 + f_4$	79.33%	85.42%	76.25%	80.33%
$f_1 + f_3 + f_6 + f_5$	80.50%	88.33%	81.75%	83.53%
$f_1 + f_3 + f_6 + f_7$	79.42%	85.58%	78.83%	81.28%
$f_1 + f_3 + f_6 + f_8$	81.08%	87.67%	82.25%	83.67%

TABLE 5. THE EVALUATION RESULTS OF COMBINING FIVE FEATURES

Features	C4.5	NN	NB	Avg
$f_1 + f_3 + f_6 + f_8 + f_2$	80.50%	87.00%	79.45%	82.32%
$f_1 + f_3 + f_6 + f_8 + f_4$	81.17%	87.92%	79.08%	82.72%
$f_1 + f_3 + f_6 + f_8 + f_5$	81.42%	88.92%	84.50%	84.95%
$f_1 + f_3 + f_6 + f_8 + f_7$	82.92%	87.75%	82.92%	84.53%

TABLE 6. THE EVALUATION RESULTS OF COMBINING SIX FEATURES

Features	C4.5	NN	NB	Avg
$f_1 + f_3 + f_6 + f_8 + f_5 + f_2$	82.83 %	85.50 %	82.08 %	83.47 %
$f_1 + f_3 + f_6 + f_8 + f_5 + f_4$	80.92 %	81.50 %	80.00 %	80.81 %
$f_1 + f_3 + f_6 + f_8 + f_5 + f_7$	82.00 %	81.83 %	83.33 %	82.39 %

TABLE 7. THE EVALUATION RESULTS OF COMBINING SEVEN FEATURES

Features	C4.5	NN	NB	Avg
$f_1 + f_3 + f_6 + f_8 + f_5 + f_2 + f_4$	82.75 %	90.00 %	81.42 %	84.72%
$f_1 + f_3 + f_6 + f_8 + f_5 + f_2 + f_7$	83.08 %	90.28 %	84.25 %	85.87%

Finally, the average classification accuracy of combining all features, $f_1 + f_3 + f_6 + f_8 + f_5 + f_2 + f_7 + f_4$, is 85.22%. Fig.9 shows the evaluation results under each kinds of combination.

It can be seen from the Fig. 9, with the increasing of the number of features, the classification accuracy will increase. However, when feature combination reaches to some extent, such as after $f_1 + f_3 + f_6$, the accuracy has no obvious change. That indicates not the more features are, the better classification accuracy is. Even under certain circumstance, large number of features will reduce the accuracy, which is the peaking phenomenon occurring for larger features.

We can find from the Fig. 9, the best classification result is about 85%. To improve the classification, other features are considered to add into the candidate set. By observing the acceleration signals of each kinds of finger gesture, we find, to a gesture signal, the acceleration change in each axis shows great difference. For example the gesture "Finger Left Shift" in Fig. 7, acceleration in Y axis shows intense change than X and Z axis. While for "Finger Up" in Fig. 8, X and Z axes show obvious fluctuation. If taking sd as the fluctuation level, the comparison of sd between two axis is generated adding into the candidate feature, which is sdX>sdY, sdY>sdZ, and sdX>sdZ. we name the features as relative features. In addition, for identifying the gestures with opposite direction, coming sequence of peak and valley in single signal is adopted, which is represented as posPeakX>posValX, posPeakY>posValY, and posPeakZ>posValZ.

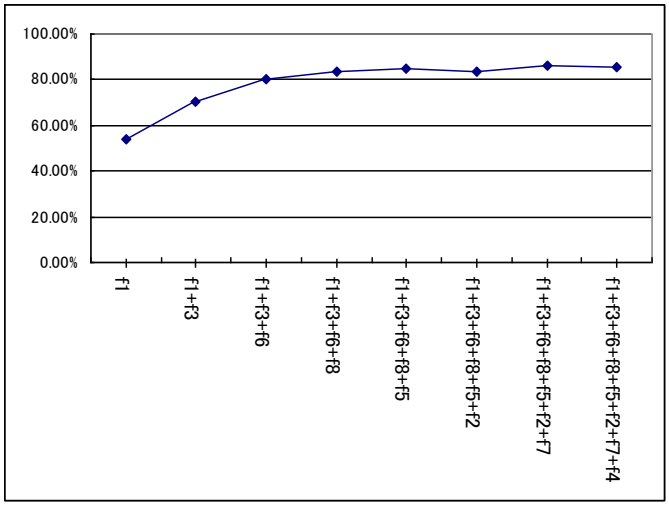


Figure 9. Recognition accuracy under different feature combinations

Using the relative features, candidate feature set is regression evaluated again. Result show the recognition accuracy reach to 90.08% with the feature subset {meanX, meanY, meanZ, sdX, sdY, sdZ, sdX>sdY, sdY>sdZ, sdX>sdZ, posPeakX>posValX, posPeakY>posValY, posPeakZ>posValZ}. These relative features not only improve the recognition accuracy, but also reduce the computing complexity because of their alternative Boolean value. Besides, the system robust can be increased since a relative relationship can prevent classifier from acting of the initial state. The final recognition matrix is shown in Table 8 based on Nearest Neighbor classifier.

VI. CONCLUSION

In the recognition task, features are of paramount importance. In the paper, we focus on the research of feature analysis and selection, which includes how to generate the candidate feature set based on the sensing information, how to evaluate each feature and their combinations, and how to select the optimal feature subset.

Feature generation process need to observe the activity signal and initially select some features and neglect insignificant one. Based on the candidate feature set, forward selection algorithm of stepwise regression is adopted to evaluate each feature and their combinations. The final combination with good classification significance is selected as feature subset for gesture recognition. Experiment result indicates the process of feature analysis and selection is feasible to most of activity recognition research. In the future, we plan to use our method to evaluate other kinds of sensing device and activity data.

ACKNOWLEDGMENT

We would like to thank the subjects who help us with the experiment in the research.

TABLE 8. RECOGNITION MATRIX FOR 12 KINDS OF FINGER GESTURES BASED ON THE SELECTED FEATURE SUBSET

Classified as	a	b	c	d	e	f	g	h	i	j	k	l	90.08%
a=Crook	86	0	3	8	1	1	0	0	0	0	1	0	86%
b=Unbend	0	100	0	0	0	0	0	0	0	0	0	0	100%
c=Finger Down	0	0	91	4	0	1	2	2	0	0	0	0	91%
d=Wrist Down	12	0	8	79	0	0	0	0	0	0	1	0	79%
e=L-Rotate	0	0	0	0	93	1	3	3	0	0	0	0	93%
f=Finger L-Shift	0	0	2	0	0	93	4	0	1	0	0	0	93%
g=Wrist L-Shift	0	0	1	0	2	11	86	0	0	0	0	0	86%
h=R-Rotate	1	0	0	0	3	0	1	89	0	6	0	0	89%
i=Finger R-Shift	0	0	0	0	1	1	0	0	96	2	0	0	96%
j=Wrist R-Shift	0	0	0	0	0	1	1	0	11	87	0	0	87%
k=Finger Up	0	1	0	0	0	0	0	0	3	0	94	2	94%
l=Wrist Up	0	1	0	0	0	0	0	0	0	0	12	87	87%

REFERENCES

[1] S. Consolvo, P. Roessler, et al. "Technology for Care Networks of Elders," IEEE Perv. Comp., vol. 43, no. 2, pp. 22-29, 2004.

[2] Stanford, V. "Using Pervasive Computing To Deliver Elder Care," IEEE Perv. Comp., vol. 1, no. 1, pp. 10-13, 2002.

[3] P. W. Jusczyk, and P. A. Luce, "Speech Perception and Spoken Word Recognition: Past and Present," Ear Hear, vol. 23, no. 1, pp. 2-40, Feb. 2002.

[4] A. Zafar, J. Overhage and C. McDonald, "Continuous Speech Recognition for Clinicians," J. Am. Med. Inform. Assoc., vol. 6, no. 3, pp. 195-204, 1999.

[5] S. Mitra, and A. Tinku, "Gesture Recognition: A Survey," IEEE Trans. on System, Man, and Cybernetics Part C: Applications and Reviews, vol. 37, no. 3, pp. 311-323, 2007.

[6] S. Helal, B. Winkler, et al. "Enabling Location-aware Pervasive Computing Applications for the Elderly," In Proc. of First IEEE International Conference on Pervasive Computing and Communications (PerCom'03), 2003, Fort Worth, Texas, March, pp. 531-536.

[7] J. R. Smith, K. P. Fishkin, et al. "RFID-based Techniques For Human-activity Detection," Communication of ACM, vol. 48, no. 9, pp. 39-44, 2005.

[8] L. Bao, S. S. Intille, "Activity recognition from user-annotated acceleration data," in Pervasive 2004, Springer: Linz, Vienna. pp. 1-17.

[9] P. Juha, M. Ermes, et al., "Activity Classification Using Realistic Data from Wearable Sensors." IEEE Trans. on Information Technology in Biomedicine, vol. 10, no. 1, pp. 119-128, 2006.

[10] U. Maurer, A. Rowe, A. Smailagic, and D. Siewiorek,, "Location and activity recognition using eWatch: A wearable sensor platform," Ambient Intelligence in Everyday Life, Lecture Notes in Computer Science, vol. 3864, pp. 86-102, 2006.

[11] X. Zhang, X. Chen, Y. Li, and etc. "A Framework for Hand Gesture Recognition Based on Accelerometer and EMG Sensors," IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans, Vol. 41, No. 6, pp. 1064-1076, 2011.

[12] T. G. Zimmerman, J. Lanier, C. Blanchard, and etc. "A hand Gesture Interface Divice," in Proc. of the SIGCHI/GI conference on Human factors in computing systems and graphics interface, New York, USA, 1987.

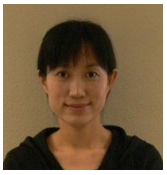
[13] M. Sekine, T. Tamura, T. Togawa, and etc. "Classification of waist-acceleration signals in a continuous walking record," Med Eng Phys. 2000 May;22(4):285-91.

[14] N. Wang, E. Ambikairajah, N. H. Lovell, "Accelerometry Based Classification of Walking Patterns Using Time-frequency Analysis," in Proc. of 29th Annual Conference of IEEE Engineering in Medicine and Biology Sciety, pp. 4899-4902, Lyon, France, 2007.

[15] L. Jing, Y. Zhou, Z. Cheng, and J. Wang, "A Recognition Method for One-Stroke Finger Gestures Using a MEMS 3D Accelerometer," IEICE Trans. on Information and Systems, E94.D(5), pp. 1062-1072, 2011.

[16] S. Theodoridis and K. Koutroumbas, Pattern Recogniton, 4st ed. U.K. Elsevier, 2009, ch. 5, pp. 261-267.

AUTHORS PROFILE



Yinghui Zhou received her B.E. degree and M.E. degree in Computer Science and Engineering from Jiamusi University and Yanshan University, China in 2001 and 2004 respectively. Now she is pursuing her ph.D. Degree in University of Aizu. Her research is concerned with ubiquitous learning, wearable computing and pattern recognition.

Lei Jing received his B.Eng. degree from Dalian University of Technology, China, in 2000, M.Eng. degree from the Yanshan University, China, in 2003, and Ph.D from University of Aizu, Japan, in 2008. Currently he is a special researcher at the University of Aizu. His research interests include sensor networks, wearable computing, and ubiquitous learning.

Junbo Wang received his B.E. in Electrical Engineering and Automation and M.E. in Electric circuits & systems in 2004 and 2007, from the YanShan University, China, and received a Ph.D. degree in Computer Science at the University of AIZU, Japan in 2011. Currently, he is a visiting researcher in the University of Aizu. His current research interests include IoT, ubiquitous computing, context/situation awareness, and WSN.

Zixue Cheng received his B.Eng. degree from Northeast Institute of Heavy Machinery in 1982, his Master degree and Ph.D degree from Tohoku University, Japan in 1990 and 1993, respectively. Currently, he is a full professor the School of Computer Science and Engineering, the University of Aizu, Japan. His current interests include distributed algorithms, ubiquitous learning, context-aware platforms, and functional safety for embedded systems

A Framework for Improving the Performance of Ontology Matching Techniques in Semantic Web

Kamel Hussein Shafa'amri
Princess Sumaya University for Technology
Amman, Jordan

Jalal Omer Atoum
Princess Sumaya University for Technology
Amman, Jordan

***Abstract*—Ontology matching is the process of finding correspondences between semantically related entities of different ontologies. We need to apply this process to solve the heterogeneity problems between different ontologies. Some ontologies may contain thousands of entities which make the ontology matching process very complex in terms of space and time requirements. This paper presents a framework that reduces the search space by removing entities (classes, properties) that have less probability of being matched. In order to achieve this goal we have introduced a matching strategy that uses multi matching techniques specifically; string, structure, and linguistic matching techniques. The results obtained from this framework have indicated a good quality matching outcomes in a low time requirement and a low search space in comparisons with other matching frameworks. It saves from the search space from (43% - 53%), and saves on the time requirement from (38% - 45%).**

Keywords- Ontology matching; RDF statements;Semantic web; Similarity Aggregation.

I. INTRODUCTION

In the current World Wide Web (WWW) computers and machines have no idea about the semantic of the information that are transferred through the web; the transferred information are not machine understandable. The role of computers is only to present the transferred information using web browsers [19].

However, the next generation of the WWW is called a Semantic Web. The role of the computers in the Semantic Web is not only to present the information, but for the computers to read and process the information in the WebPages, and extract knowledge from this information.

The computer can understand the information in the Semantic Web by using a data structure called Ontology. Ontology provides a knowledge representation in a particular domain; it defines concepts (classes and properties) in a given domain, and shows the relationships between the defined concepts [1], [19].

Different people may develop different ontologies that describe a particular domain; this causes heterogeneity problems between ontologies that describe the same domain. In general different ontologies for a specific domain may use different data formats, modeling languages and structures to represent certain knowledge. The heterogeneity problem leads to inability to get accurate search results in semantic web. For example, some ontologies define a car as a "car" and another ontologies define a car as an "automobile", so if we write a keyword "car" in a semantic web search engine then the result of the search engine will be a list of all WebPages that are based on ontologies that define car as a "car", and this list will not contain the WebPages that are based on ontologies that define a car as "automobile".

In general, ontology provides knowledge in a certain domain to help the machines to make intelligent decisions. Ontology consists of four components: concepts, object properties, data properties and Individuals. In ontologies, we can define RDF statements. An RDF statement consists of three elements [1], [19]: Resource (Subject or Domain), Object Property (Predicate or Property), Value (Object or Range).

To solve the heterogeneity problem between ontologies, we must apply a process called ontology matching process. Ontology matching is the process of finding correspondences between semantically related entities of different ontologies. These correspondences stand for different relations such as equivalence, more general, or disjointness, between ontologies entities.

II. BACKGROUND

There are several types of matching techniques that are used to find the correspondences entities between ontologies. These techniques are string-based techniques, language-based techniques and structural techniques [9].

Several ontology matching systems were developed to find matched entities between different ontologies, such as Naive Ontology Mapping (NOM) [7], PROMPT [15], Anchor-PROMPT [14] and GLUE [5]. These previous systems take two ontologies as inputs and use different ontology matching techniques to test all entities of the first ontology with all entities of the second ontology in order to find the matched entities between the input ontologies [6]. So the search space and time requirements of these previous systems are very large.

To reduce the search space and time requirement of the ontology matching process, we present in this paper an ontology matching framework (system). This framework uses a multi matching techniques specifically; string, structure, and linguistic matching techniques, and depends on some important features of ontologies; such as RDF statements and class hierarchies.

III. PROPOSED ONTOLOGY MATCHING FRAMEWORK

A. System Overview

As shown in Fig.1, the proposed matching system (PMS) takes two ontologies as input, and determines the matched entities (e.g. classes, object properties, data properties) between these two ontologies. PMS compares entities of the same type.

Specifically, it compares classes of ontology1 with classes of ontology2, object properties of ontology1 with object properties of ontology2, and data properties of ontology1 with data properties of ontology2. PMS uses three types of matching techniques, string and linguistic techniques in a combined framework called "structure matching".

In order to reduce the search space of the matching process, PMS is dependent on RDF statements and class hierarchies which are the base components in ontologies. This PMS matches RDF statements of ontology1 with RDF statements of ontology2, and matches class hierarchies of ontology1 with class hierarchies of ontology2.

The output of PMS is a set of matched entities with their similarity values (confidence measures) each between 0 and 1. All the entities that have similarity values greater than a pre-defined threshold are considered to be correct matched entities. And the output, also, includes the relationship types between the matched entities (equivalence and subsumption). The matching relationship is dependent on similarity values. Our PMS focuses on one-to-one (1:1) and many-to-many (m: m) match relationships, since they are the most commonly used.

B. Structure of Ontology Matching Framework

The proposed system PMS has three stages: pre-processing, matching process and post-processing.

1) Pre-processing

In this stage, PMS extracts the features of the input ontologies. As shown in Fig.2, the system will read all RDF statements and put them in a list that is called RDF statements list. Each element in the RDF statements list will be one RDF statement (subject, object properties, object).

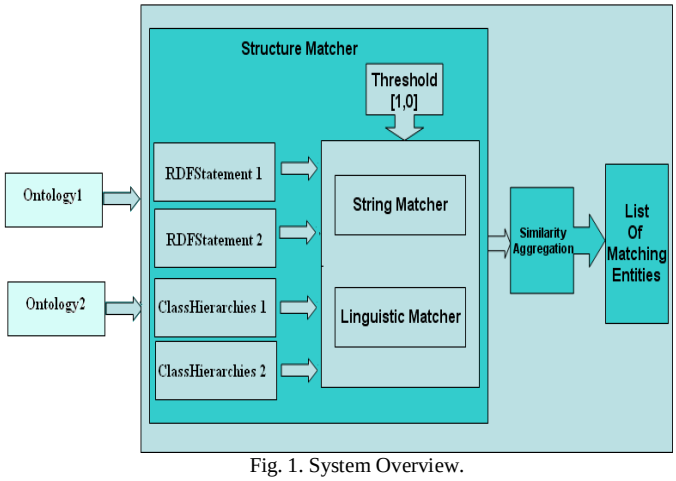


Fig. 1. System Overview.

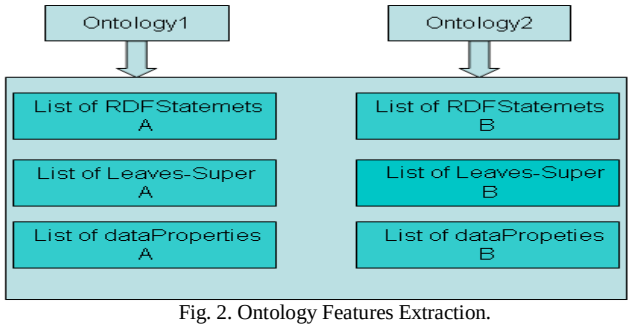


Fig. 2. Ontology Features Extraction.

Then, the system reads all leaves classes and their super-classes and put them in a list that is called leaves-super list. Each element in the leaves-super list will be an object that contains a leaf class and its super classes. Finally, the system reads all ontology classes and their data properties and put them in data properties list. Each element in data properties list will be an object that contains class and its data properties.

2) Matching Process

Structure matching consists of two stages; the first one involves matching of RDF statements and the second one involves matching of class hierarchies. Matching class hierarchies also consists of two sub stages, matching of leaves-super classes and matching of class-data properties.

a) Similarity Aggregation

In order to combine the similarity values of string matcher and linguistic matcher, we use the following similarity aggregation function [4]:

$$Sim_{agg}(e1,e2) = W_S \times Sim_S(e1,e2) + W_L \times Sim_L(e1,e2)$$

Where e1 is an entity of ontology1 and e2 is an entity of ontology2, Simagg( ) is similarity combination of string similarity Sims ( ) and linguistic similarity SimL ( ), Ws is a string weight and WL is a linguistic weight. Ws, WL∈ [0, 1] and Ws + WL = 1. We used Ws = 0.3 and WL = 0.7. This means that the linguistic matcher is more important than the string matcher.

b) Matching RDF Statements

As mentioned earlier, an RDF statement has three components (Subject, Object property, Object). PMS will match every RDF statement in RDF-statements-list-A of ontology1 with every RDF statement in RDF-statements-list-B of ontology2 as illustrated in Fig.3.

PMS computes the similarity aggregation value of the subject of an RDFstatement-A with the subject of an RDFStatement-B, if their similarity aggregation value is greater than the threshold value (matched subject) then it will compute the similarity aggregation value of the object property of an RDFStatement-A with the object property of an RDFStatement-B. If their similarity aggregation value is greater than the threshold value (matched object property) then it will compute the similarity aggregation value of the object of an RDFStatement-A with the object of an RDFStatement-B.

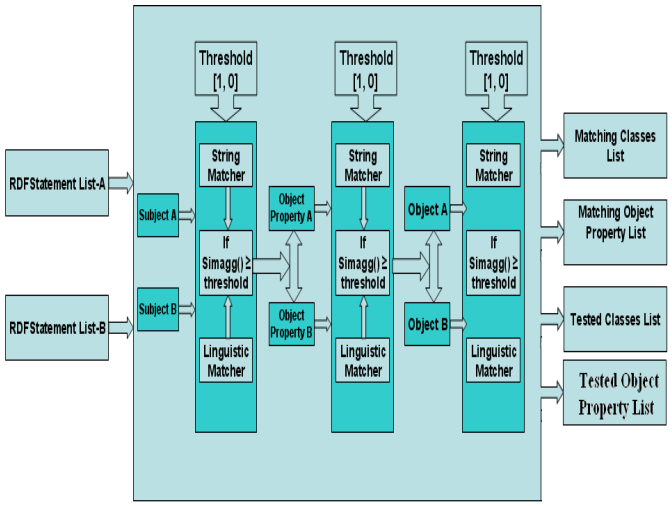


Fig. 3. RDF Statements Matching Process.

Also, PMS adds the tested subjects and objects to the tested classes list, and adds the tested object properties to the tested object properties list. Then, PMS checks if the object, subject classes, and object properties were tested before, by searching the tested corresponding list. These operations are done to prevent computing similarity aggregation values for classes (subjects and objects) and object properties more than once.

Finally, the system will add matched subjects and objects with their similarity aggregation values to the matched classes list, and will add matched object properties with their similarity aggregation values to the matched object properties list.

Matching RDF statements using this scenario will reduce the search space of the matching process for the following reasons:

- PMS ignores the object properties and the objects of RDF statements if the subjects of RDF statements are not matched.
- PMS ignores the object of RDF statements if the subjects or the objects properties of RDF statements are not matched.

The outputs of the matching RDF statement process are the following four lists:

1) Matched classes list.
2) Matched object properties list.
3) Tested classes list.
4) Tested Object Properties List.

c) Matching Class Hierarchies

In PMS, there are two types of class hierarchies, the first one leaves-superClasses and the second one class-data property. Each type has its own matching process.

- *Matching leaves-superClasses*

In this stage, PMS reads four lists; two lists are outputted from the previous process (Matching RDF Statements), which are the matched classes list and the tested classes list. And the other two lists are the leaves-superList-A of ontology1 and the leaves-superList-B of ontology2. Then, PMS will apply the matching process as illustrated in Fig.4.

PMS computes the similarity aggregation value of the leaf in leaves-superList-A with the leaf of leaves-superList-B, if their similarity aggregation value is greater than the threshold value (matched leaves) then it will compute the similarity aggregation value for all super classes of the matched leafs.

Also, PMS adds the tested leaves classes and super-classes to tested classes list. Then, PMS checks if the leaves and super-classes were tested before, by searching the tested corresponding list. These operations are done to prevent computing similarity aggregation values for classes more than once.

Matching leaves-super classes using this scenario will reduce the search space of the matching process because PMS ignores the super-classes if the leaves are not matched.

The outputs of the matching leaves-superClasses process are the following two lists:

1) Matched classes list.
2) Tested classes list.

- *Matching Class-data Property*

In this stage, as illustrated in Fig.5, PMS reads three lists: matched classes list, dataProperties list-A of ontology1 and dataProperties list-B of ontology2. Data Properties list contains objects. Each object presents a class and its data properties. PMS check every pair of matched classes in the matched classes list, if they have data properties.

Also, PMS adds the tested data properties to tested data properties list. Then, PMS checks if the data properties were tested before, by searching the tested corresponding list. These operations are done to prevent computing similarity aggregation values for data properties more than once.

Again, matching class-data property using this scenario will reduce the search space of the matching process because the system will ignore the data properties of non-matching classes, and the system will try to match the data properties of matched classes only.

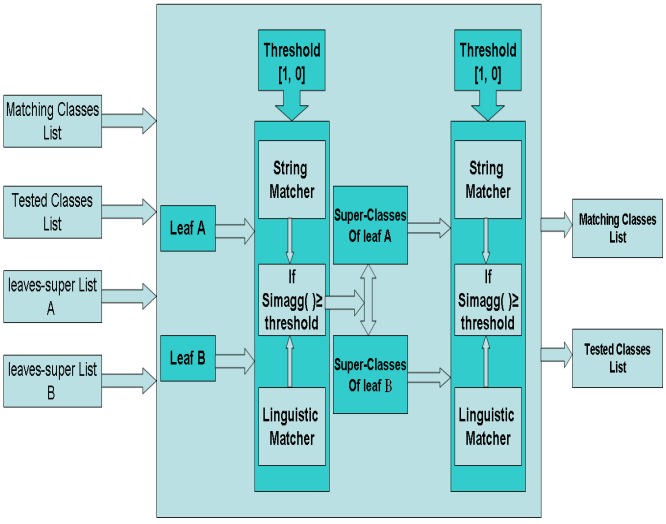


Fig. 4. Leaves-SuperClasses Matching Process.

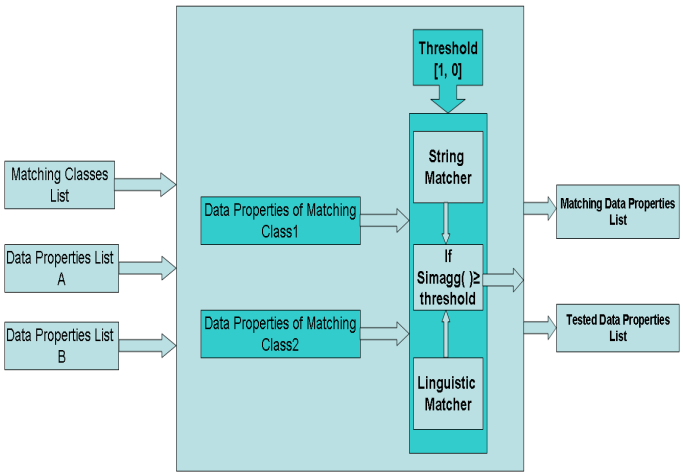


Fig. 5. Data Properties Matching Process.

The outputs of the matching Class-data Property process are the following two lists:

1) Matched data properties list.
2) Tested data properties list.

d) Final Outputs of the Matching Process

The final outputs of the PMS will be three lists as follows:

- Matched classes list.
- Matched object properties list.
- Matched data properties list.

3) Post-Processing

In this stage, the PMS assigns matching relationship R to the matched entities (Classes, Object properties, Data properties) according to their similarity value [9].

- If Simagg (e1, e2) = 1 then R is the equivalence relation.
- If Simagg (e1, e2) ≥ threshold then R is a subsumption relation.

Where Simagg (e1, e2) is the similarity aggregation value between matched entities e1 and e2.

4) String Matcher Implementation

For string matchers, the PMS uses Levenshtein distance similarity measure [17] and soundex similarity measure [17], [22] in combined manner as follows:

$$Sim_S(e1,e2) = \frac{Levenshtein\ distance(e1,e2) + Soundex(e1,e2)}{2}$$

5) Linguistic Matcher Implementation

For linguistic matchers, the PMS uses Wordnet similarity measures of Wu & Palmer similarity measure [18] and path similarity measure [20] in combined manner as follows:

$$Sim_L(e1,e2) = \frac{Wu\ \&\ Palmer(e1,e2) + Path\ (e1,e2)}{2}$$

IV. Proposed Framework Evaluation

There are two types of evaluations that are used to evaluate the PMS. The first evaluation is done by counting the number of tested entities that were tested and by computing the time requirement that are needed by the system to find the matched entities. The second type of evaluation is based on compliance measures to evaluate the quality of the matched results.

A. Compliance Measures

Following the work in [8], Compliance Measures are used to evaluate the degree of compliance of the results of matching algorithms. Compliance measures consist of three measures Precision, Recall and F-measure. These measures are used to evaluate the quality of the matching process and its results. Precision and Recall are based on the comparison of an expected result provided by a reference alignment and the effective result of the evaluated system. Finally, F-measure combines the measures of Precision and Recall as single efficiency measure.

B. Traditional Matching System

For the purpose of evaluating our PMS we have developed a matching system called (Traditional Matching System) TMS that is based on the work of some existing ontology matching systems such as NOM [7], PROMPT [15], Anchor-PROMPT [14] and GLUE [5]. The TMS matches all classes of the first ontology with all classes of the second ontology, and matches all object properties of the first ontology with all object properties of the second ontology, and matches all data properties of the first ontology with all data properties of the second ontology. The goal of developing the TMS is to compare it with the PMS.

C. Conference Dataset

The conference dataset, shown in Table 1, has been proposed in OAEI 2010 and it includes seven ontologies that are dealing with conference organization. These ontologies have been developed within OntoFarm project [21], and are quite suitable for ontology matching task because of their heterogeneous character. Every ontology in this dataset has a number of classes, a number of object properties and a number of data properties. The matching process will be done on each pairs of these ontologies.

TABLE 1. CONFERENCE DATASET [21].

Ontology Name	# of classes	# of object properties	# of data properties
Cmt	29	49	10
Conference	59	46	18
ConfOf	38	13	23
Edas	103	30	20
Ekaw	73	33	0
iasted	140	38	3
sigkdd	49	17	11

D.

E. TMS vs. PMS

The comparison of the PMS and the TMS **was** done on the same computer system ((Intel (R) Core (TM) 2 Duo CPU, 2.4GHz, 3 GB RAM) and Windows 7)). We had applied the matching process using the PMS and the TMS **between** each pair of ontologies of the conference dataset at different threshold values (0.5, 0.7, 0.85, and 1).

1) Search Space and Time Requirement Evaluation

Figures 6 and 7 present the average number of tested entities and the average time requirement at different threshold values (0.5, 0.7, 0.85, and 1) that are needed by both systems TMS and PMS to match all the pairs of Conference dataset ontologies.

We can notice from these two Figures, that the number of tested entities and the time requirement that were needed to find the matched entities in the TMS are larger than the number of tested entities and time requirement that were needed by the PMS, this is due to the fact that the TMS tests more entities than the PMS.

Furthermore, we can notice that the number of tested entities and the time requirement for the TMS remain the same regardless of the threshold values. Whereas, in the PMS they are inversely dependent on the threshold value.

2) Compliance Measures Evaluation

Figures 8, 9 and 10, present the average compliance measures results **(**Precision, Recall and F-measure) of all the matched pairs of Conference dataset ontologies, at different threshold values (0.5, 0.7, 0.85, and 1) for both TMS and PMS.

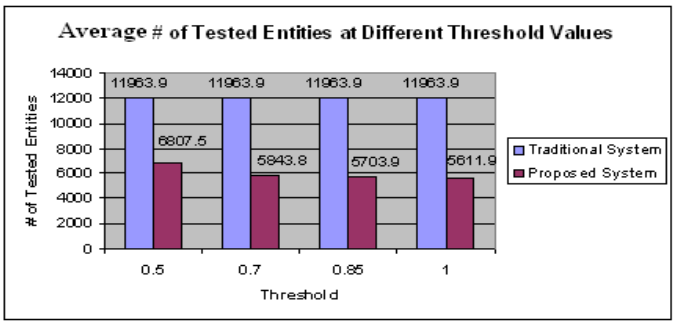


Fig. 6. Average Number of Tested Entities.

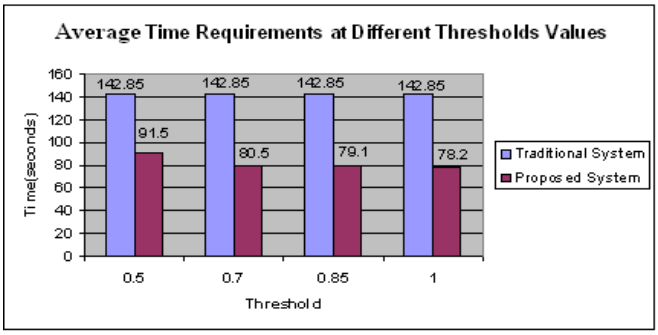


Fig. 7. Average Time Requirement.

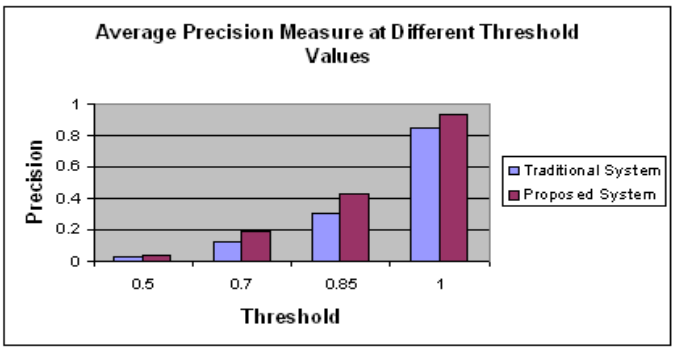


Fig. 8. Average Precision Measure.

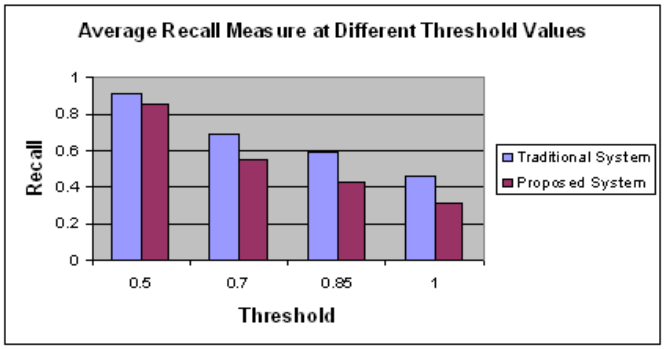


Fig. 9. Average Recall Measure.

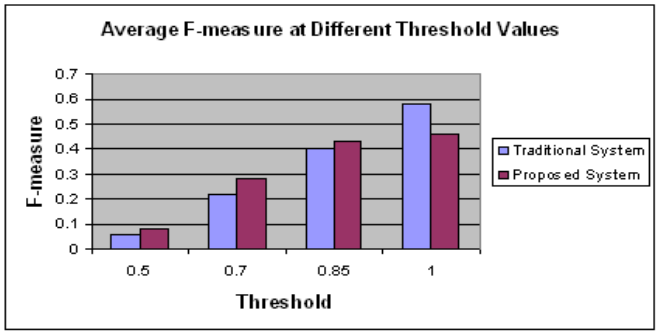


Fig. 10. Average F-measure.

We can notice from these Figures the following:

- At all thresholds values the Precision value of the PMS is better than the Precision value of the TMS. Hence, the PMS returns more accurate matching results than the TMS.
- At all thresholds values the Recall value of the TMS is better than the Recall value of the PMS. This means that the TMS returns more matched entities that are existed in reference alignment R than the PMS;
- At Threshold values (0.5, 0.7, and 0.85) the F-measure value of the PMS is better than the F-measure of the TMS. But at threshold value 1 the F-measure value of the TMS is better than the F-measure of the PMS; this is due to the fact that F-measure is dependent on Recall and Precision values.

F. Comparison with other Existing Matching Systems

We had made a comparison between the PMS and other matching systems that participated in OAEI 2010 in terms of Precision, Recall and F-measure. These systems are AgrMaker [11], AROMA [3], ASMOV [12], CODI [13], Ef2Match [2], Falcon [10] and GeRMeSMB [16]. This comparison is done at threshold values of 0.5 and 0.7. Figures 11 and 12 show the results of this comparison.

We can notice from these Figures the following:

- The PMS at threshold 0.5 and at threshold 0.7 has the lowest Precision value, because the PMS returns the largest number of matched entities but a few of them are existed in the reference alignment.
- The PMS has the highest Recall value at threshold 0.5. This means that the PMS returns the largest number of matched entities that are existed in reference alignment than the other matching systems.
- The PMS has a good Recall value between the Recall values of the other systems at threshold 0.7.
- The PMS has a low F-measure value at threshold 0.5 and at threshold 0.7.

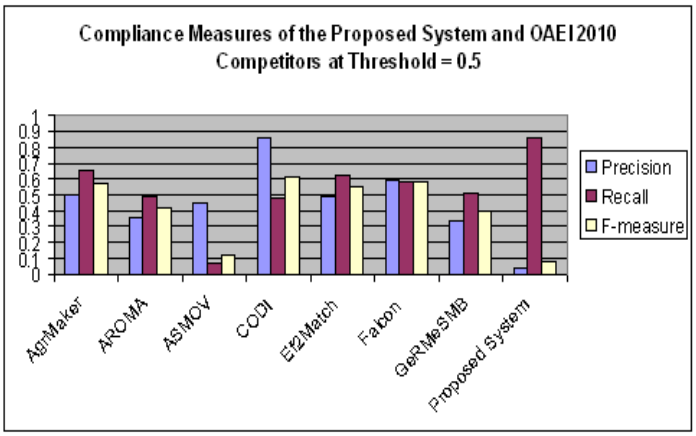


Fig. 11. Comparison with Other Matching Systems at Threshold = 0.5.

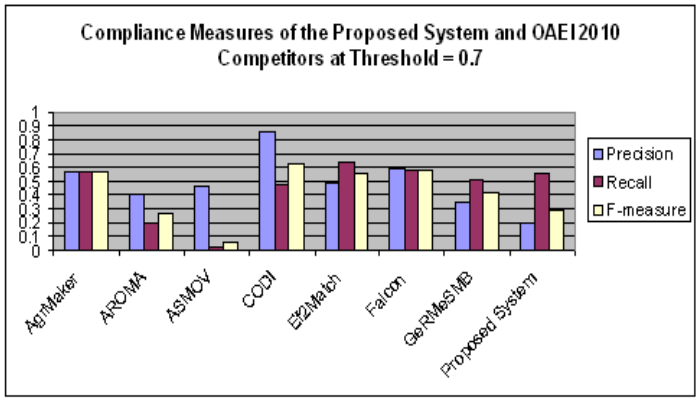


Fig. 12. Comparison with Other Matching Systems at Threshold = 0.7.

V. Conclusion

The main goal of this paper was to reduce the complexity (search space and time requirement) of the ontology matching process. This paper have introduced an ontology matching framework that reduces the search space and the time requirement of the matching process by removing entities (classes, properties) that have less probability of being matched. The proposed ontology matching framework had used a multi matching techniques in order to find the correspondences entities between ontologies.

The proposed matching framework saves (43% - 53%) from the number of tested entities (search space). Furthermore, the proposed matching framework saves on time requirement of the matching process from (38% - 45%) in comparisons with other matching frameworks.

The drawback of the proposed matching framework is that it can't find all possible alignments entities between ontologies, due to the fact that the PMS doesn't test all entities of the matching ontologies. Hence, the PMS is recommended to be used in matching large ontologies, since it will produce a huge number of matching entities that could be enough for web searching using semantic web.

References

[1] G. Antoniou and F. Harmelen , "A semantic web prime", (2nd ed). London: Massachusetts Institute of Technology, 2008.

[2] W. Chua and J. Kim, "Eff2Match results for OAEI 2010", proceedings of the 5th international workshop on ontology matching, Shanghai, China, 2010.

[3] J. David F. Guillet and H. Briand , "Matching directories and OWL ontologies with AROMA", proceedings of the 15th ACM international conference on Information and knowledge management, ACM, New York, USA, pp. 831-831, 2006.

[4] H. Do and E. Rahm, "COMA - A system for flexible combination of schema matching approaches", proceedings of the very larged data bases conference (VLDB), Hong Kong, China, pp610-621, 2002.

[5] A.Doan, J. Madhavan, P. Domingos and A.Halevy, "Ontology matching: a machine learning approach" In: S. Staab, and R. Studer (Ed), handbook on ontologies in information system, (pp.385–404), Berlin: Springer-Verlag, 2003.

[6] M. Ehrig and S.Staab, "QOM: quick ontology mapping", Proceedings of the international semantic web conference (ISWC), Hiroshima, Japan, pp. 683–697, 2004.

[7] M. Ehrig and Y. Sure. "Ontology mapping - an integrated approach", proceedings of the 1st european semantic web symposium (ESWS), vol. 3053, pp.76–91, Heraklion: Springer-Verlag, 2004.

[8] J. Euzenat, "Semantic precision and recall for ontology alignment evaluation", proceedings of international joint conference on artificial intelligence (IJCAI), Hyderabad, India, pp. 248-253, 2007.

[9] J.Euzenat and P. Shvaiko, "Ontology matching", (1st ed.). Berlin: Springer-Verlag, 2007.

[10] W. Hu, J. Chen, G. Cheng and Y. Qu , "ObjectCoref & Falcon-AO: results for OAEI 2010", proceedings of the 5th international workshop on ontology matching, Shanghai, China, 2010.

[11] C. Isabel. S. Cosmin, C. Michele, F. Caimi, M. Palmonari, F. Antonelli and K. Ulas, "Using agreementmaker to align ontologies for OAEI 2010", proceedings of the 5th international workshop on ontology matching, Shanghai, China, 2010.

[12] Y. Jean-Mary P. Shironoshita and M. Kabuka, "ASMOV results for OAEI 2010", proceedings of the 5th international workshop on ontology matching, Shanghai, China, 2010.

[13] J. Noessner and M. Niepert, "CODI: Combinatorial Optimization for Data Integration – results for OAEI 2010", proceedings of the 5th international workshop on ontology matching, Shanghai, China, 2010.
[14] N. Noy and M. Musen, "Anchor-PROMPT: using non-local context for semantic matching" ,proceedings of international joint conference on artificial intelligence (IJCAI), Seattle, US, pp. 63–70, 2001.
[15] N. Noy and M. Musen, "The PROMPT suite: interactive tools for ontology merging and mapping". international journal of human-computer studies, vol. 59(6), pp. 983–1024, 2003.
[16] C. Quix, A. Gal, T. Sagi and D. Kensche, "An integrated matching system: GeRoMeSuite and SMB Results for OAEI 2010", proceedings of the 5th international workshop on ontology matching, Shanghai, China, 2010.
[17] M. Taye, "Ontology alignment mechanisms for improving web-based searching", unpublished doctoral dissertation, De Montfort University, United Kingdom, England, 2009.
[18] Z. Wu and M. Palmer, "Verb semantics and lexical selection". proceedings of 32nd annual meeting of the association for computational linguistics (ACL), Las Cruces, US, pp. 133–138, 1994.
[19] L. Yu, "Introduction to the semantic web and semantic web services", (1st ed.). Florida: Taylor & Francis Group, 2007.
[20] Pedersen, Ted. "Wordnet similarity", from, http://search.cpan.org/dist/WordNet Similarity/lib/WordNet/Similarity/path.pm .
[21] Ontology alignment evaluation initiative, http://oaei.ontologymatching.org/ .
[22] "Soundex", from http://en.wikipedia.org/wiki/Soundex.

Fingerprint Image Enhancement:
Segmentation to Thinning

Iwasokun Gabriel Babatunde
Department of Computer Science
Federal University of Technology,
Akure, Nigeria

Akinyokun Oluwole Charles
Department of Computer Science
Federal University of Technology,
Akure, Nigeria

Alese Boniface Kayode
Department of Computer Science
Federal University of Technology,

Olabode Olatubosun
Department of Computer Science
Federal University of Technology,
Akure, Nigeria

***Abstract*— Fingerprint has remained a very vital index for human recognition. In the field of security, series of Automatic Fingerprint Identification Systems (AFIS) have been developed. One of the indices for evaluating the contributions of these systems to the enforcement of security is the degree with which they appropriately verify or identify input fingerprints. This degree is generally determined by the quality of the fingerprint images and the efficiency of the algorithm. In this paper, some of the sub-models of an existing mathematical algorithm for the fingerprint image enhancement were modified to obtain new and improved versions. The new versions consist of different mathematical models for fingerprint image segmentation, normalization, ridge orientation estimation, ridge frequency estimation, Gabor filtering, binarization and thinning. The implementation was carried out in an environment characterized by Window Vista Home Basic operating system as platform and Matrix Laboratory (MatLab) as frontend engine. Synthetic images as well as real fingerprints obtained from the FVC2004 fingerprint database DB3 set A were used to test the adequacy of the modified sub-models and the resulting algorithm. The results show that the modified sub-models perform well with significant improvement over the original versions. The results also show the necessity of each level of the enhancement.**

Keyword- AFIS; Pattern recognition; pattern matching; fingerprint; minutiae; image enhancement.

I. INTRODUCTION

In the world today, fingerprint is one of the essential variables used for enforcing security and maintaining a reliable identification of any individual. Fingerprints are used as variables of security during voting, examination, operation of bank accounts among others. They are also used for controlling access to highly secured places like offices, equipment rooms, control centers and so on. The result of the survey conducted by the International Biometric Group (IBG) in 2004 on comparative analysis of fingerprint with other biometrics is presented in Fig. 1. The result shows that a substantial margin exists between the uses of fingerprint for identification over other biometrics such as face, hand, iris, voice, signature and middleware [1]. The following reasons had been adduced to the wide use and acceptability of fingerprints for enforcing or controlling security [1]-[4]:

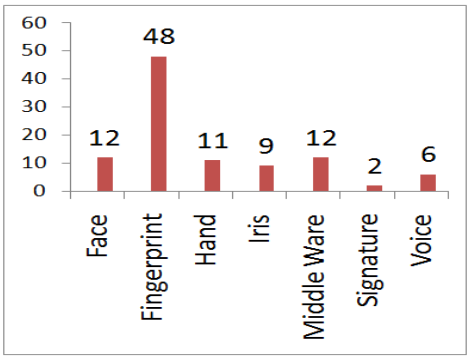


Figure 1: Comparative survey of fingerprint with other biometrics

a) Fingerprints have a wide variation since no twopeople have identical prints.

b) There is high degree of consistency in fingerprints. A person's fingerprints may change in scale but not in relative appearance, which is not the case in other biometrics.

c) Fingerprints are left each time the finger contacts a surface.

d) Availability of small and inexpensive fingerprint capture devices

e) Availability of fast computing hardware

f) Availability of high recognition rate and speed devices that meet the needs of many applications

g) The explosive growth of network and Internet transactions

h) The heightened awareness of the need for ease-of-use as an essential component of reliable security.

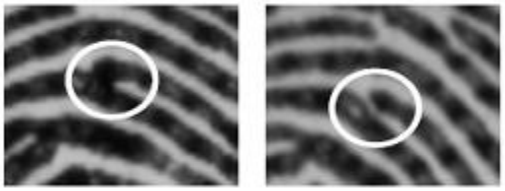
The main ingredients of any fingerprint used for identification and security control are the features it possesses. The features exhibit uniqueness defined by type, position and orientation from fingerprint to fingerprint and they are

classified into global and local features [5]-[7]. Global features are those characteristics of the fingerprint that could be seen with the naked eye. They are the features that are characterized by the attributes that capture the global spatial relationships of a fingerprint. Global features include ridge pattern, type, orientation, spatial frequency, curvature, position and count. Others are type lines, core and delta areas.

The Local Features are also known as Minutia Points. They are the tiny, unique characteristics of fingerprint ridges that are used for positive identification. Local features contain the information that is in a local area only and invariant with respect to global transformation. It is possible for two or more impressions of the same finger to have identical global features but still differ because they have local features (minutia points) that are different. In Fig. 2, ridge patterns (a) and (b) are two different impressions of the same finger (person). A local feature is read as bifurcation in (a) while it appears as a ridge ending in (b).

II. FINGERPRINT IMAGE ENHANCEMENT

Reliable and sound verification of fingerprints in any AFIS is always preceded with a proper detection and extraction of its features. A fingerprint image is firstly enhanced before the features contained in it could be detected or extracted. A well enhanced image will provide a clear separation between the valid and spurious features. Spurious features are those minutiae points that are created due to noise or artifacts and they are not actually part of the fingerprint. This paper adopts with slight modifications, the algorithm implemented in [8]-[9] for fingerprint image enhancement. The overview of the algorithm is shown in Fig. 3. Its main steps include image segmentation, local normalization, filtering and binarization/thinning.



(a) Bifurcation (b) Ridge ending

Figure 2: Different minutiae for different impressions of the same finger

A Image Segmentation

There are two regions that describe any fingerprint image; namely the foreground region and the background region. The foreground regions are the regions containing the ridges and valleys. As shown in Fig. 4, the ridges are the raised and dark regions of a fingerprint image while the valleys are the low and white regions between the ridges. The foreground regions often referred to as the Region of Interest (RoI) is shown for the image presented in Fig. 5. The background regions are mostly the outside regions where the noises introduced into the image during enrolment are mostly found. The essence of segmentation is to reduce the burden associated with image enhancement by ensuring that focus is only on the foreground regions while the background regions are ignored.

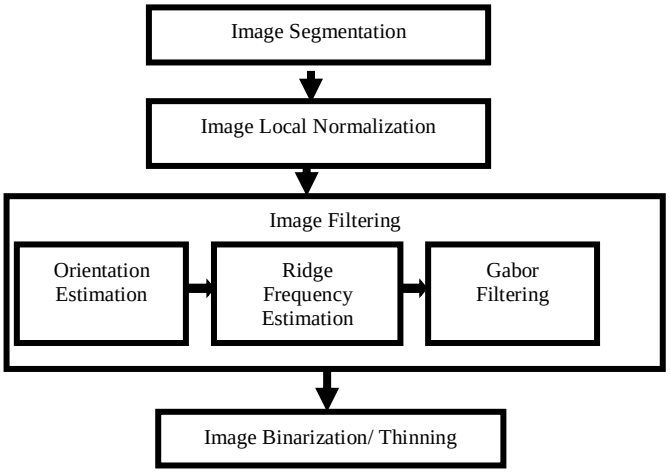


Figure 3: The conceptual diagram of the fingerprint enhancement algorithm

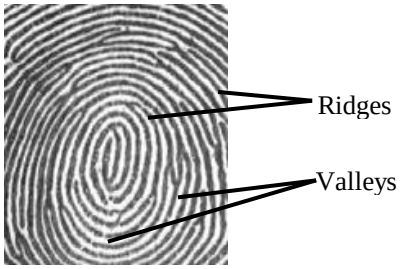


Figure 4: Ridges and valleys on a fingerprint image

The background regions possess very low grey-level variance values while the foreground regions possess very high grey-level variance values. A block processing approach used in [8]-[9] is adopted in this research for obtaining the grey-level variance values. The approach firstly divides the image into blocks of size W x W and then the variance V(k) for each of the pixels in block k is obtained from:

$$V(k) = \frac{1}{W^2}\sum_{i=1}^{W}\sum_{j=1}^{W}(I(i,j) - M(k))^2 \quad (1)$$

$$M(k) = \frac{1}{W^2}\sum_{a=1}^{W}\sum_{b=1}^{W}J(a,b) \quad (2)$$

I(i,j) and J(a,b) are the grey-level value for pixel i,j and (a,b) respectively in block k.

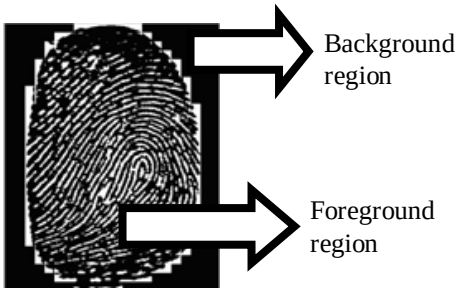


Figure 5: A fingerprint image and its foreground and background regions

A. Image Local Normaliuzation

Normalization is performed on the segmented fingerprint image ridge structure so as to standardize the level of variations in the image grey-level values. By normalization, the grey-level values are made to fall within certain range that is good enough for improved image contrast and brightness. The first of the tasks of image normalization implemented in [8]-[9] and adopted for this research is the division of the segmented image into blocks of size S x S. The grey-level value for each pixel is then compared with the average grey-level value for the host block. For a pixel I(i,j) belonging to a block of average grey-level value of M, the result of comparison produced a normalized grey-level value N(i,j) defined by the formula:

$$N(i,j) = \begin{cases} M_0 + \sqrt{\dfrac{V_0(I(i,j)-M)^2}{V}} & if\ \ I(i,j) > M \\ M_0 - \sqrt{\dfrac{V_0(I(i,j)-M)^2}{V}} & otherwise \end{cases} \quad (3)$$

where an assumed value of M_0 is set for the desired mean and an assumed value of V_0 is set for the desired variance.

B. Image Filtering

Normalized fingerprint image is filtered for enhancement through removal of noise and other spurious features. Filtering is also used for preserving the true ridge and valley structures. The fingerprint image filtering structure adopted for this research is in the following phases:

1) Orientation Estimation: Orientation estimation is the first of the prerequisites for fingerprint image filtering. In every image, the ridges form patterns that flow in different directions. The orientation of a ridge at location x,y is the direction of its flow over a range of pixels as shown in Fig. 6.

The Least Square Mean (LSM) fingerprint ridge orientation estimation algorithm proposed and implemented in [8]-[9] was slightly modified and used in this research. The modified algorithm involves the following steps:

a) Firstly, blocks of size S x S were formed on the normalized fingerprint image.

b) For each pixel, (p,q) in each block, the gradients $\partial_x(p,q)$ and $\partial_y(p,q)$ were computed as the gradient magnitudes in the x and y directions, respectively.

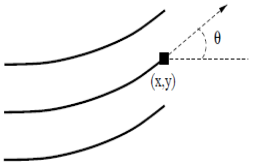


Figure 6: The orientation of a ridge pixel in a fingerprint

∂_x(p,q) was computed using the horizontal Sobel operator while ∂_y(p,q) was computed using the vertical Sobel operator.

$$\begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

Horizontal Sobel Operator Vertical Sobel Operator

c) The local orientation of a pixel in a fingerprint image was computed by using its S x S neighborhood in [8]-[9]. This was slightly modified in this research by dividing the image into S x S blocks and the local orientation for each block centered at pixel I(i,j) was then computed from:

$$V_x(i,j) = \sum_{p=i-\frac{S}{2}}^{i+\frac{S}{2}} \sum_{q=j-\frac{S}{2}}^{j+\frac{S}{2}} 2\partial_x(p,q)\partial_y(p,q) \quad (4)$$

$$V_y(i,j) = \sum_{p=i-\frac{S}{2}}^{i+\frac{S}{2}} \sum_{q=j-\frac{S}{2}}^{j+\frac{S}{2}} \partial_x^2(p,) - \partial_y^2(p,q) \quad (5)$$

$$\theta(i,j) = \frac{1}{2}\tan^{-1}\frac{V_y(i,j)}{V_x(i,j)} \quad (6)$$

where $\Theta(i, j)$ is the least square estimate of the local orientation at the block centered at pixel (i, j).

d) The orientation image is then converted into a continuous vector field defined by:

$$\varphi_x(i,j) = \cos\big(2\theta(i,j)\big), \quad (7)$$
$$\varphi_y(i,j) = \sin\big(2\theta(i,j)\big), \quad (8)$$

where φ_x and φ_y are the *x* and *y* components of the vector field, respectively.

e) Gaussian smoothing is then performed on the vector field as follows:

$$\varphi_X'(i,j) = \sum_{p=-\frac{S_\varphi}{2}}^{\frac{S_\varphi}{2}} \sum_{q=-\frac{S_\varphi}{2}}^{\frac{S_\varphi}{2}} G(p,q)\varphi_x(i-ps, j-qs), \quad (9)$$

$$\varphi_y'(i,j) = \sum_{p=-\frac{S_\varphi}{2}}^{\frac{S_\varphi}{2}} \sum_{q=-\frac{S_\varphi}{2}}^{\frac{S_\varphi}{2}} G(p,q)\varphi_y(i-ps, j-qs), \quad (10)$$

where G is a Gaussian low-pass filter of size $S_\varphi \; X \; S_\varphi$.

f) The orientation field O of the block centered at pixel (i,j) is finally smoothed using the equation:

$$O(i,j) = \frac{1}{2}\tan^{-1}\frac{\varphi_y'(i,j)}{\varphi_x'(i,j)} \quad (11)$$

2) Ridge Frequency Estimation: The second prerequisite for fingerprint image filtering is the ridge frequency estimation. In any fingerprint image, there is a local frequency of the ridges that collectively form the ridge frequency image. The ridge frequency is obtained from the extraction of the ridge map from the image. The extraction of the ridge map involves the following steps:

a) Compute the consistency level of the orientation field obtained from the first prerequisite in the local neighborhood of a pixel (p,q) with the following formula:

$$C_o(p,q) = \frac{1}{n^2}\sqrt{\sum_{(i,j)\in W} |\theta(i,j)-\theta(p,q)|^2} \quad (12)$$

$$|\theta(i,j)-\theta(p,q)| = \begin{cases} d & if\ d < 180 \\ d-180 & otherwise \end{cases} \quad (13)$$

$$d = (\theta(i,j)-\theta(p,q)+360)\ mod\ 360 \quad (14)$$

where W represents the local neighborhood around (p,q), which is an n x n local window, θ(i,j) and θ(p,q) are local ridge orientations at pixels (i,j) and (p,q) respectively.

b) If the consistency level is below a certain threshold F_c, then the local orientations in this region are re-estimated at a lower image resolution level until the consistency is above F_c. After the orientation field is obtained, the following two adaptive filters are applied to the image:

$$h_t(p,q,i,j) = \begin{cases} \frac{-1}{\sqrt{2\pi\delta}} e^{\frac{-1}{\delta^2}}, & if\ i = l(j) - d,\ j \in \rho \\ \frac{1}{\sqrt{2\pi\delta}} e^{\frac{-1}{\delta^2}}, & if\ i = l(j),\ j \in \rho \\ 0, & otherwise, \end{cases} \quad (15)$$

$$h_b(p,q,i,j) = \begin{cases} \frac{-1}{\sqrt{2\pi\delta}} e^{\frac{-1}{\delta^2}} & if\ i = l(j) + d,\ j \in \rho \\ \frac{1}{\sqrt{2\pi\delta}} e^{\frac{-1}{\delta^2}}, & if\ i = l(j),\ j \in \rho \\ 0, & otherwise, \end{cases} \quad (16)$$

$$l(j) = j\ tan(\theta(p,q)); \quad (17)$$

$$d = \frac{Y}{2\cos(\theta(p,q))}, \quad (18)$$

$$\rho = Y\left[\left|\frac{\sin(\theta(p,q)}{-2}\right|, \left|\frac{\sin(\theta(p,q)}{2}\right|\right] \quad (19)$$

The two filters are capable of stressing under different condition the local maximum grey level values along the normal direction of the local ridge orientation. The normalized image is first convolved with these two masks, h_t (p,q, i, j) and h_b (p,q, i, j). If both the grey level values at pixel (p,q) of the convolved images are larger than a certain threshold F_{ridge}, then pixel (p,q) is labeled as a ridge.

3) Orientation Estimation: Having obtained the prerequisites, Gabor filtering is then used to improve or enhance the fingerprint image to a finer structure. It involves the removal of noise and artifacts. The general form of Gabor filter is:

$$G(x,y:f,\theta) = exp\left\{\frac{1}{2}\left[\frac{a^2}{\delta_x^2}+\frac{b^2}{\delta_y^2}\right]\right\} cos(2\pi f a) \quad (20)$$

where f is the frequency of the cosine wave along the direction θ from the *x*-axis, and δ_x and δ_y are the space constants along *x* and *y* axes respectively. a= xsinΘ + ycosΘ and b= xcosΘ + ysinΘ.

The values of the space constants δ_x and δ_y for the Gabor filters were empirically determined as each is set to about half the average inter-ridge distance in their respective direction. δ_x and δ_y are obtained from $\delta_x = k_x F$ and $\delta_y = k_y F$ respectively. F is the ridge frequency estimate of the original image, and k_x *and* k_y are constant variables. The value of δ_x determines the degree of contrast enhancement between ridges and valleys while the value of δ_y determines the amount of smoothing applied to the ridges along the local orientation.

C. Image Binarization/Thinning

The image obtained from the Gabor filtering stage is binarized and thinned to make it more suitable for feature extraction. The method of image binarization proposed in [10] is employed. The Method sets the threshold (T) for making each cluster in the image as tight as possible, thereby minimizing their overlap. To determine the actual value of T, the following operations are performed on set of presumed threshold values:

a) The pixels are separated into two clusters according to the threshold.
b) The mean of each cluster are determined.
c) The difference between the means is squared.
d) The product of the number of pixels in one cluster and the number in the other is determined.

The success of these operations depends on the difference between the means of the clusters. The optimal threshold is the one that maximizes the between-class variance or, conversely, the one that minimizes the within-class variance. The *within-class* variance of each of the cluster is then calculated as the weighted sum of the variances from:

$$\sigma^2_{within}(T) = n_B(T)\sigma_B^2(T) + n_O(T)\sigma_O^2(T) \quad (21)$$

$$n_B(T) = \sum_{i=0}^{T-1} p(i) \quad (22)$$

$$n_O(T) = \sum_{i=T}^{N-1} p(i) \quad (23)$$

$$\sigma_B^2(T) = the\ variance\ of\ the\ pixels\ in\ the\ background\ (below) threshold$$

$$\sigma_O^2(T) = the\ variance\ of\ the\ pixels\ in\ the\ foreground\ (above) threshold$$

p(i) is the pixel value at location i, N is the intensity level and $[0, N-1]$ is the range of intensity levels. The between-class variance, which is the difference between the within-class variance and the total variance of the combined distribution, is then obtained from:

$$\sigma^2_{between}(T) = \sigma^2 - \sigma^2_{within}(T) \quad (24)$$
$$= n_B(T)[A] + n_O(T)[B] \quad (25)$$
$$A = (\mu_B(T) - \mu)^2 \quad (26)$$
$$B = (\mu_O(T) - \mu)^2 \quad (27)$$

where σ^2 is the combined variance, μ_B(T) is the combine mean for cluster T in the background threshold, μ_O(T) is the combine mean for cluster T in the foreground threshold and μ is the combined mean for the two thresholds. The between-class variance is simply the weighted variance of the cluster means themselves around the overall mean. Substituting $\mu = n_B(T)\mu_B(T) + n_O(T)\mu_O(T)$ into Equation 25, the result is:

$$\sigma^2_{between}(T) = n_B(T)n_O(T)[\mu_B(T) - \mu_O(T)]^2 \quad (28)$$

Using the following simple recurrence relations, the between-class variance is successfully updated by manipulating each threshold T using a constant value p as follows:

$$n_B(T+1) = n_B(T) + p \quad (29)$$
$$n_O(T+1) = n_O(T) - p \quad (30)$$
$$\mu_B(T+1) = \frac{\mu_B(T)n_B(T) + pT}{n_B(T+1)} \quad (31)$$
$$\mu_O(T+1) = \frac{\mu_O(T)n_O(T) - pT}{n_O(T+1)} \quad (32)$$

III. Experimental Results

A slightly modified version of the fingerprint enhancement algorithm used in [8]-[9] was implemented in this research by using MATLAB Version 7.2 on the Windows Vista Home Basic operating system. The experiments were performed on a Pentium 4 – 1.87 GHz processor with 1024MB of RAM. The purpose of the fingerprint enhancement experiments is to analyze the performance of the modified algorithm under different conditions of images as well as generate the metrics that could serve the basis for the comparison of the results from the research with results from related works. Two sets of experiment were conducted for the performance analysis. The first set of experiments was on synthetic images. The orientation estimation, ridge frequency estimation and Gabor filtering experiments all employed the *circsine* function [11] to generate the synthetic images. The major arguments passed into the *circsine* function include a number for the size of the square image to be produced, a number for the wavelength in pixels of the sine wave and an optional number specifying the standard pattern of behaviour to use in calculating the radius from the centre. This defaults to 2, resulting in a circular pattern while large values give a square pattern. Where necessary, the MATLAB *imnoise* function was also used to generate noise and artifacts on synthetic images. The arguments passed into the *imnoise* function include the image on which noise is to be generated, noise type and noise level. The second set of experiments was on the FVC2004 fingerprint database DB3 set A.

Fig. 7(a), 7(b) and 7(c) are synthetic images of size 200 x 200 and wavelength 8. They were obtained with the *imnoise* function using the salt and pepper noise level of 0, 0.2 and 0.4 respectively. The results of the ridge orientation estimation experiments on each of the three images are shown in Fig. 7(d), 7(e) and 7(f) respectively. These results show that for noise levels of 0 and 0.20, the modified ridge orientation algorithm effectively generated ridge orientation estimates that are very close to the actual orientations. However, for image with noise level of 0.4, the result shows a substantial number of ridge orientation estimates that significantly differ from the actual orientations. These results show that the performance of the algorithm depends on the image noise level. When the noise level on the image is within reasonable range, the algorithm does well while it fails when the noise level rises beyond the threshold which was found to be 0.29.

The efficiency of the ridge orientation estimation algorithm was quantitatively measured by estimating the Mean Square Error (MSE) which represents the difference between the estimated and actual ridge orientation values in radians. Mean square error estimation results for the synthetic image shown in Fig. 7(a) under different conditions of noise for both the pixel processing approach in [8] and the block processing approach formulated for this research are shown in Table 1. The increasing mean square errors in both cases indicate that the accuracy of the two approaches decreases with increase in the noise level. It is also revealed that the orientation estimate is closer to the actual value for the block processing approach at any noise level. With lower standard deviation of 0.0393 for its MSE values, it is established that the block processing approach performs better than the pixel processing approach with MSE values of standard deviation of 0.1686. This higher performance is attributed to the fact that in the block processing approach, the degree of variation in the orientation estimates for pixels in a block is reduced to zero as each pixel assumes the orientation estimate for the center pixel of its host block. This significantly increases the ability of the algorithm to estimate the ridge orientation close to the actual value.

Fig. 7(g), 7(h) and 7(i) present the results of the ridge frequency estimation experiments on the synthetic images of different noise levels shown in Fig. 7(a), 7(b) and 7(c) respectively. In the ridge frequency estimates shown in Fig. 7(g) and Fig. 7(h), there is uniformity between most of the estimated frequency values for each 32 *x* 32 block as against Fig. 7(i) which shows irregular patterns. Fig. 7(g), 7(h) and 7(i) produced MSE value of 0.0006, 0.0017 and 0.0077 respectively. The ridge frequency estimates and the increasing MSE values reveal that the performance of the ridge frequency estimation algorithm decreases with increase in the image noise level. Generally, below the noise level of 0.30, it was discovered that the algorithm produced relatively uniform ridge frequency estimates as shown in Fig. 7(g) and 7(h).

When the noise level equals or exceeds 0.3, the orientation estimation algorithm produced non uniform results as shown in Fig. 7(i). Since the performance of the ridge frequency estimation algorithm depends significantly on the performance

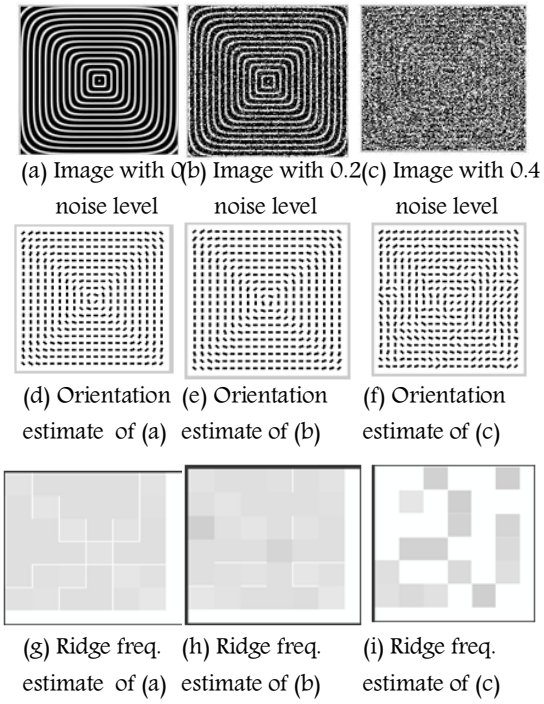


Figure 7: Orientation and ridge frequency estimates for synthetic images of different noise levels

of the ridge orientation estimation algorithm, the failure of the orientation estimation algorithm explains the failure of the ridge frequency estimation algorithm for noise level above 0.29.

The MSE results of the implementation of the ridge frequency estimation algorithm on the synthetic image shown in Fig. 7(a) under different conditions of noise in [8] and the current study are shown in Table 2. It is revealed that there is increase in the MSE values as the noise level increases for the two cases. This translates to diminishing performance due to increasing margin between the actual and the estimated ridge frequency values. It is also shown that the MSE value is significantly lower in all cases for the current study than in [8]. With lower standard deviation of 0.0032 for its MSE values, it is established that the implementation of the ridge frequency estimation algorithm in the current study yields better results than its implementation in [8] which yielded a standard deviation of 2.8514 for its MSE values. This improvement is attributed to the superior performance of the block processing approach to ridge orientation estimation over the pixel processing approach.

The performance of the Gabor filtering algorithm on a zero, medium and high quality synthetic image of size 410 x 410 and wavelength 15 is presented in Fig. 8(d), 8(e) and 8(f) respectively. Parameter values of k_x =0.45 and k_y =0.45 were used to obtain these results. The results presented in Fig. 8(d) and 8(e) reveal that with zero or medium level noise density, the filter effectively removed the noise from the image and enhanced it to a level that is comparable with the original image. This effectiveness is partly due to the previous accurate estimation of the ridge orientation and the ridge frequency for zero or medium noise level images.

However, the experimental result presented in Fig. 8(f) reveals that when the filter is applied to images with higher noise level, the filter is unable to remove the noise effectively as it produced a significant amount of spurious features. This ineffectiveness is due to the inaccurate estimation of the ridge orientation and the ridge frequency at higher noise levels. When experiments were performed on real fingerprint images, like in [8], the best results were obtained for image segmentation with variance threshold of 100. This threshold value provided the best segmentation results in terms of differentiating between the foreground and the background regions as shown in Fig. 9(b). The results of segmentation using threshold value of 95 and 105 are presented in Fig. 9 (c) and 9(d) respectively.

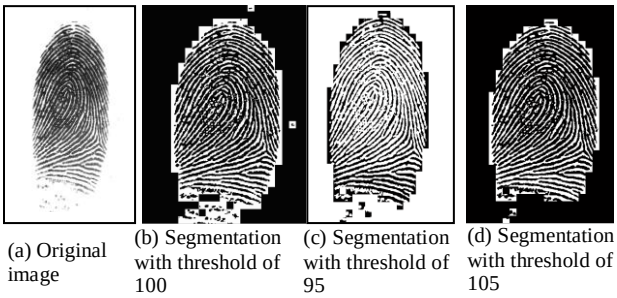


Fig. 9: Results of segmentation with different threshold

TABLE 1: COMPARISON OF MEAN SQUARE ERROR FOR RIDGE ORIENTATION ESTIMATES

Noise level	Mean Square Error	
	Raymond [8]	Current study
0.00	0.0003	0.0000
0.05	0.0009	0.0006
0.10	0.0032	0.0008
0.15	0.0102	0.0019
0.20	0.0246	0.0015
0.25	0.0691	0.0026
0.30	0.1722	0.0405
0.35	0.2330	0.0521
0.40	0.3041	0.0661
0.45	0.4124	0.0803
0.50	0.4262	0.1085

TABLE 2: COMPARISON OF MEAN SQUARE ERROR FOR RIDGE FREQUENCY ESTIMATES

Noise level	Mean Square Error	
	Raymond[8]	Current Study
0.00	0.0100	0.0006
0.05	0.0211	0.0009
0.10	0.0465	0.0011
0.15	0.0820	0.0012
0.20	0.1702	0.0017
0.25	1.1149	0.0033
0.30	2.0229	0.0041
0.35	4.1149	0.0060
0.40	5.8543	0.0077
0.45	6.1098	0.0080
0.50	7.2616	0.0090

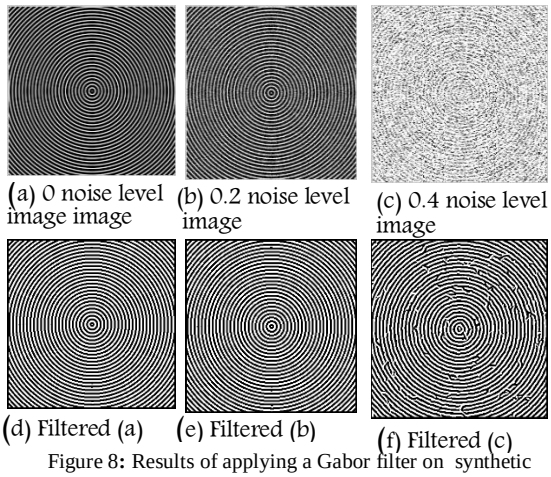


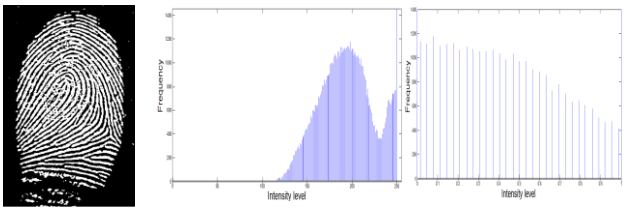
Figure 8: Results of applying a Gabor filter on synthetic images of different noise levels.

These results show inappropriate segmentation due to inaccurate variance thresholds. With lower threshold of 95, some of the background regions have been segmented to the foreground while some foreground regions are also segmented to the background under higher threshold of 105.

The result of the normalization experiment on the fingerprint images shown in Fig. 9(a) is presented in Fig. 10(a). The desire mean of zero and variance of one used in [8] were adopted and used to normalize the ridges in the images. During normalization, all positions are evenly shifted along the horizontal axis, which makes the structure of the ridges and valleys to become well and suitably positioned.

The histogram plots of the original and the normalized images are shown in Fig. 10(b) and 10(c) respectively. The histogram plot of the original image shows that all the intensity values of the ridges show irregular frequency values and also fall within the right hand side of the 0–255 scale, with no pixels in the left hand side. This leads to an image with a very low contrast.

The histogram plot of the normalized image shows that the range of intensity values for the ridges has been adjusted between 0-1 scale such that there is a more evenly and balanced distribution between the dark and light pixels and that the ridge frequencies fall within close values. The normalized image histogram plots also show that the normalization process does distribute evenly the shape of the original image. The positions of the values are evenly shifted along the *x*-axis, which means the structure of the ridges and valleys are now well and suitably positioned. This shifted and improved positioning lead to images with a very high contrast shown in Fig. 10(a).



(a)Normalized image (b) Histogram of original image (b) Histogram of normalised image

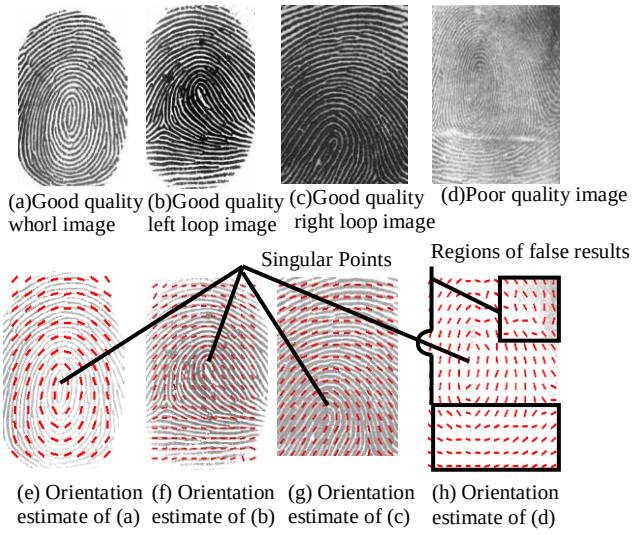
Figure. 10: Normalized image and the histogram plots

The orientation fields for the real fingerprint images were obtained around their singular points since they are prominent features used by any AFIS for fingerprint classification and matching. Good quality images are shown in Fig. 11(a), 11(b) and 11(c). Their orientation estimates are shown in Fig. 11(e), 11(f) and 11(g) respectively.

At the singular points, the orientation field is discontinuous and unlike the normal ridge flow pattern, the ridge orientation varies significantly. From these results, it is observed that there exists no deviation between the actual fingerprint ridge orientation and the estimated orientation of the vectors. In both cases, the algorithm produces accurate estimates of the orientation vectors such that they flow smoothly and consistently with the direction of the ridge structures in the images. In the superimposed version of images in Fig. 11(e), 11(f), 11(g) and 11(h), the contrast of the original image is lowered in each case. This was done to improve the visibility of the orientation vectors against the background.

The ridge orientation estimate for poor quality image shown in Fig. 11(d) is presented in Fig. 11(h). The estimate indicates a fairly smooth orientation field in some well-defined regions while it gives misleading results in areas of very poor quality as evident in the top-left and bottom regions of the estimate. The orientation estimates resulting from the pixel processing approach in [8] and the block processing approach of the current study for the image shown in Fig. 12(a) are presented in Fig. 12(b) and 12(c) respectively.

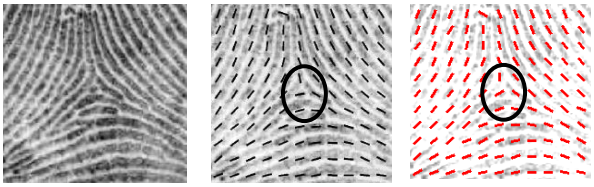
Visual inspection of these results reveals that the two methods did well in the ridge orientation estimation. However, the orientation is observed to be closer to the actual orientation in block processing than pixel processing in some regions especially the core areas represented by the inserted circles. The reason adduced to this is that assigning equal orientation estimate for pixels in a block rather than maintaining different values is better and able to take the estimates closer to their actual values.



(a)Good quality whorl image (b)Good quality left loop image (c)Good quality right loop image (d)Poor quality image

(e) Orientation estimate of (a) (f) Orientation estimate of (b) (g) Orientation estimate of (c) (h) Orientation estimate of (d)

Figure 11: Fingerprint images and their orientation estimates

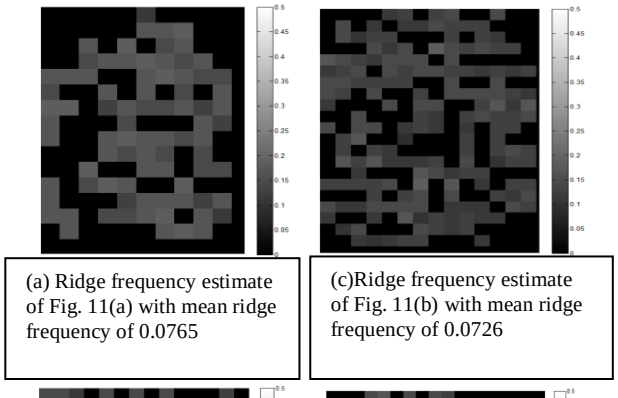


(a) Original Image (b) Orientation estimate obtained in [8] (c) Orientation estimate for current study

Fig. 12: Orientation estimates for pixel and block processing approaches

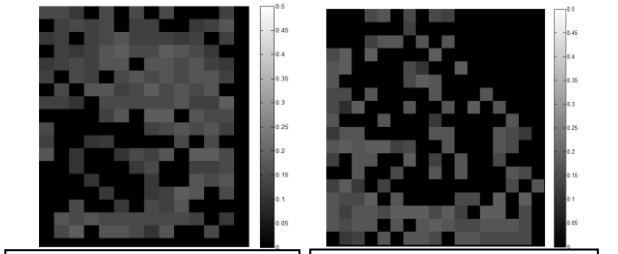
Visual inspection of the results for the ridge frequency estimates for real fingerprint images shown in Fig. 11(a), 11(b), 11(c) and 11(d) are shown in Fig. 13(a), 13(b), 13(c) and 13(d) respectively. The mean ridge frequency (MRF), which is the average of the image ridge frequencies, is also presented for each image. It is noted that the MRF values differ for all the images. This difference is attributed to the fact that fingerprints exhibit variation in their average ridge frequency characteristics and contrast levels. The intensities of frequency differ for blocks or regions within same image. Some blocks or regions exhibit high contrast while others exhibit low contrast. Based on these, the synthetic images are more appropriate for the evaluation of the accuracy of the ridge frequency estimation algorithm.

Fig. 14 reveals the extent to which the filtering algorithm was able to remove noise from the images shown in Fig. 11 for different values of k_x and k_y. The results shown in Fig. 14(a), 14(b), 14(c) and 14(d) were obtained using parameter values of k_x = 0.45 and k_y = 0.45. With these values, it is shown that the contrast level between ridges and valleys for each of the images is neither too high nor too low. Appropriate level of smoothing is also applied to the ridges along the local orientation.



(a) Ridge frequency estimate of Fig. 11(a) with mean ridge frequency of 0.0765

(c)Ridge frequency estimate of Fig. 11(b) with mean ridge frequency of 0.0726



(d) Ridge frequency estimate of Fig. 11(c) with mean ridge frequency of 0.0821

(e) Ridge frequency estimate of Fig. 11(d) with mean ridge frequency of 0.0637

Fig. 13: Ridge frequency estimate of selected images

With lower parameter values of k_x = 0.40 and k_y = 0.40, it is shown in Fig. 14(e), 14(f), 14(g) and 14(h) that the contrast level between ridges and valleys is too low and this explains why there is a number of dark regions. The degree of smoothing is also poor as there is a good number of overlapping ridges in the filtered images. When k_x=0.5 and k_y=0.5 were used, the results of the filtering experiments shown in Fig. 14(i), 14(j), 14(k) and 14(l) reveal that there is no significant difference from the results obtained for k_x=0.45 and k_y=0.45 except that some regions described by circles appear to be excessively smoothened.

It is therefore stated that based on the modified algorithm, parameter values of k_x=0.45 and k_y=0.45 are most appropriate for image filtering as against k_x=0.50 and k_y=0.50 proposed in [8]-[9]. This reduction in parameter values is due to the better performance of the block processing approach in the orientation and ridge frequency estimations as attested to by the results in Table 1 and Table 2. Results presented in Fig. 14 show that the filtering algorithm is able to smoothen to a fine level with appropriate parameter values for good quality fingerprint images.

When the quality degrades like the one shown in Fig. 11(d), the performance of the algorithm diminishes as it produces images with inappropriately filtered regions as shown in Fig. 14(d). This is also corroborated with reasons adduced to the values presented in Table 1 and Table 2.

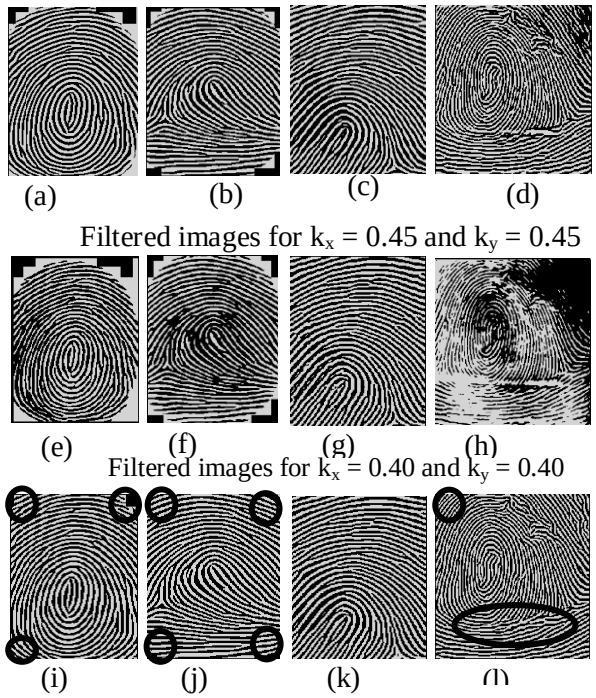


Fig. 14: Filtered images with different values for k_x and k_y

The results of the binarization experiments for the images shown in Fig. 11 are presented in Fig. 15. Visual inspection of the results shows that the binarization algorithm perfectly separated the ridges (black pixels) from the valleys (white pixels).

To obtain these results, the grey-level value of each pixel in the filtered image is examined and, if the value is greater than the threshold value 1, then the pixel value is set to a binary value one; otherwise, it is set to zero. The threshold value successfully made each cluster as tight as possible and also eliminated all overlaps. The threshold value of 1 was taken after a careful selection from a series of within and between class variance values ranging from 0 to 1 that optimally supported the maximum separation of the ridges from the valleys. The clear separation of the ridges from the valleys verifies the correctness of the algorithm as proposed in [9] and implemented in this research.

The results of the thinning experiment on each of the images shown in Fig. 11 are presented in Fig. 16(a-d). The MATLAB's *bwmorph* operation using the 'thin' option was used to generate the thinned images. These results show that the ridge thickness in each of the images has been reduced to its smallest form or skeleton (one pixel wide). It is also shown that the connectivity of the ridge structures is well preserved. Fig. 16(e-h) shows the results of performing binarization experiments on the raw images without the enhancement stages. In contrast to Fig. 15(a-d), the binary images in Fig. 16(e-h) are not well connected and contain significant amount of noise and corrupted elements. Consequently, when thinning is applied to these binary images, the results in Fig. 16(i-l) show that the accurate extraction of minutiae would not be possible due to the large number of spurious features produced. Thus, it is shown that employing the image enhancement stages prior to thinning is effective for accurate and speedy extraction of minutiae. Consequently, when thinning is applied to these binary images, the results in Fig. 16(i-l) show that the accurate extraction of minutiae would not be possible due to the large number of spurious features produced.

Thus, it is shown that employing the image enhancement stages prior to thinning is effective for accurate and speedy extraction of minutiae.

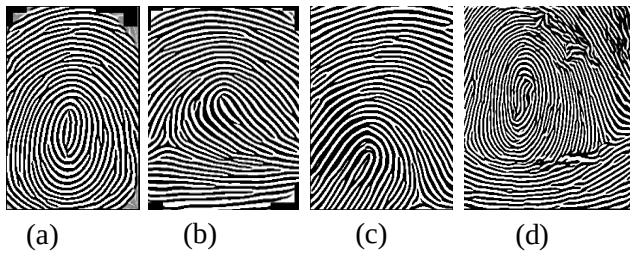


Fig. 15: Results of Binarization for images shown in Fig. 11(a), 11(b), 11(c) and 11(d)

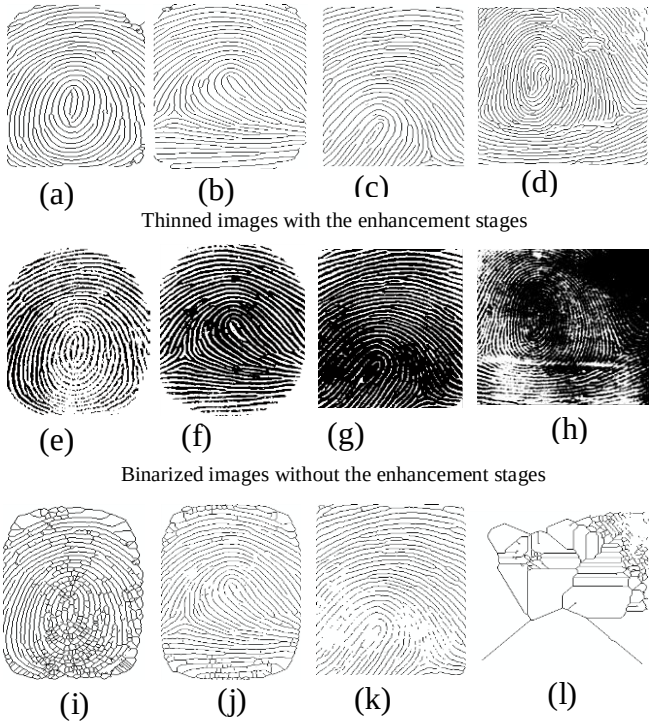


Fig. 16: Thinned and Binarized images

IV. CONCLUSION

This paper discusses the results of the modification and verification of the fingerprint enhancement algorithm developed and implemented in [8]-[9]. Some stages of the algorithm were slightly modified for improved performance. For instance, block processing approach was introduced into the orientation estimation algorithm in place of the pixel processing approach.

While the pixel processing approach subjects each pixel in the image to orientation estimation, the block processing approach firstly divides the image into S x S blocks and obtains the orientation estimate for the center pixel. This resulted in higher performance as attested to by the Tables of MSE for ridge orientation and frequency estimates. Parameter values of k_x =0.45 and k_y =0.45 were found to perform well in the image filtering experiments as against k_x =0.5 and k_y =0.5.

The results of the experiments conducted for image segmentation, normalization, ridge orientation estimation, ridge frequency estimation, Gabor filtering, binarization and thinning on synthetic and real fingerprint images reveal that with free or minimal noise level, the algorithms perform well. Improved performance is specifically noticed for the modified ridge orientation estimation algorithm. It is also established that each stage of the enhancement process is important for obtaining a perfectly enhanced image that is acceptable and presentable to the features extraction stage. The results obtained from the final stage of thinning show that the connectivity of the image ridge structure has been preserved and improved at each stage.

REFERENCES

[1] C. Roberts: 'Biometrics': (http://www.ccip.govt.nz/newsroom/informoationnotes/2005/biometrics.pdf, Accessed: July, 2009

[2] C. Michael and E. Imwinkelried: 'Defence practice tips, a cautionary note about fingerprint analysis and reliance on digital technology', Public Defense Backup Centre Report., 2006

[3] M. J. Palmiotto: 'Criminal Investigation'. Chicago: Nelson Hall, 1994

[4] Salter D.: 'Fingerprint – An Emerging Technology', Engineering Technology, New Mexico State University, 2006

[5] J. Tsai-Yang, and V. Govindaraju: A minutia-based partial fingerprint recognition system, Center for Unified Biometrics and Sensors, University at Buffalo, State University of New York, Amherst, NY USA 14228, 2006

[6] O. C Akinyokun, C. O. Angaye and G. B. Iwasokun: 'A Framework for Fingerprint Forensic'; Proceeding of the First International Conference on Software Engineering and Intelligent System, organized and sponsored by School of Science and Technology, Covenant University, Ota, Nigeria, 2010, pages 183-200.

[7] O. C. Akinyokun and E. O. Adegbeyeni: Scientific Evaluation of the Process of Scanning and Forensic Analysis of Fingerprints on Ballot Papers', Proceedings of Allied Academies International Conference on Legal, Ethics and Regulatory Measures in New Orleans, April 8 - 10, 2009; Vol. 13; No. 1; pages 1 - 13.

[8] L. Hong, Y. Wan and A. Jain: 'Fingerprint image enhancement: Algorithm and performance evaluation'; Pattern Recognition and Image Processing Laboratory, Department of Computer Science, Michigan State University, 2006, pp1-30

[9] R. Thai: Fingerprint Image Enhancement and Minutiae Extraction, PhD Thesis Submitted to School of Computer Science and Software Engineering, 2003, University of Western Australia.

[10] X. Xu : 'Image Binarization using Otsu Method'. Proceedings of NLPR-PAL Group CASIA Conference, 2009, pp345-349

[11] P. Kovesi: 'MATLAB functions for computer vision and image analysis', School of Computer Science and Software Engineering, University of Western Australia, http:/www.cs.uwa.edu.au/~pk/Research/MatlabFns/Index.html, Accessed: 20 February 2010.

Data Warehouse Requirements Analysis Framework: Business-Object Based Approach

Anirban Sarkar
Department of Computer Applications
National Institute of Technology, Durgapur
West Bengal, India

***Abstract*—Detailed requirements analysis plays a key role towards the design of successful Data Warehouse (DW) system. The requirements analysis specifications are used as the prime input for the construction of conceptual level multidimensional data model. This paper has proposed a Business Object based requirements analysis framework for DW system which is supported with abstraction mechanism and reuse capability. It also facilitate the stepwise mapping of requirements descriptions into high level design components of graph semantic based conceptual level object oriented multidimensional data model. The proposed framework starts with the identification of the analytical requirements using business process driven approach and finally refine the requirements in further detail to map into the conceptual level DW design model using either Demand-driven of Mixed-driven approach for DW requirements analysis.**

***Keywords- Requirements analysis; Business objects; Conceptual Model; Graph Data Model; Data Warehouses*.**

I. INTRODUCTION

Complex, online and multidimensional analysis of data is done by fetching just-in-time information from subjective, integrated, consolidated, non–volatile, historical collection of data. Data Warehouse (DW) and On Line Analytical Processing (OLAP) in conjunction with multidimensional database are typically used for such analysis. DW facilitates data navigation, analysis, and business oriented visualization of data using multidimensional cube and OLAP query processing. DW systems are used mainly by decision makers to analyze the status and the development of an organization [1], based on large amounts of data integrated from heterogeneous sources into a multidimensional data model. As in other information systems, requirement analysis plays a key role within DW system development to reduce the risk of failure.

DW projects are long-term projects, so it is very difficult to anticipate future requirements for the decision-making process in the scenario of evolving business processes over time [2]. Further, information requirements for DW systems are difficult to specify at the early stage because decision processes are flexibly structured and shared across the different sectors of the large organization. In this scenario, requirement analysis for DW project must support reuse of domain level abstractions and step wise refinement mechanisms strictly for mapping the requirements in high level design. Hence, for the DW project, requirements analysis must start focusing with early requirements analysis and further it should move on detailed requirements analysis. Analyzing early requirements will significantly decrease the possibility of misunderstanding the user's requests and consequently reduce the risk of failure for the DW project.

Several studies [2, 3, 4] indicate that improper analysis of user requirements or avoidance of requirements analysis phase leads to unsuccessful design of DW System. Further they recommended that, there must be a dedicated phase of requirement analysis for the purpose of DW system design. The primary deliverables of the DW requirements analysis phase is conceptual level multidimensional data model [5]. In this course, the schemata of the available operational data sources are also compared with the user driven information requirements. Indeed, the approaches to DW design are usually classified in three categories:

a) *Supply-driven / Data-driven:* In these approaches [6, 7, 8], the DW design starts from a detailed analysis of the operational data sources. In order to determine the structure of conceptual multidimensional data model the user requirements have less impact in this approach.

b) Demand-driven / Requirements-driven: These approaches [2, 9, 10] start from determining the information requirements of business users or stakeholders. The problem of mapping these requirements onto the available data sources is faced only at the late design phase of DW system. This approach is the only alternative whenever a deep analysis of data sources is unfeasible, or data sources reside on legacy systems whose inspection and normalization is not recommendable. However, in this case, conceptual level multidimensional data model design can be directly based on mapping of requirements.

c) Mixed-Driven (Supply/Demand): In these approaches [9, 11, 12], requirements analysis and source inspections are carried out in parallel. The Supply-driven and mixed framework is recommended when source schemata are well known, and their size and complexity are substantial.

Majority of these approaches mainly have focused on requirement analysis and does not provide any definite guidelines to move from requirements model to high level conceptual design. In [9] goal oriented approaches has been described to support both demand driven and mixed framework for DW system requirements analysis. It comprehensively supports early requirement analysis phase for DW system and

provides a mechanism to map the requirements into the conceptual model. But author has not clearly explained about the detailed requirements analysis for DW system.

In this paper, a requirement analysis framework for DW system has been proposed based on concept of Common Business Objects [13]. A business object (BO) captures information about a real world (business) concept, operations on that concept, constraints on those operations, and relationships between those elements and other business concepts. The set of related BOs express an abstract view of the business's "real world". The advantage of using this concept is that, the set of BOs can be reusable in the context of business domain and it ensures the resultant system will be scalable, reliable, secure and interoperable [13].

In the perspective of requirements analysis for DW system, our proposed framework consist of three phases, namely, (i) Early Requirements Analysis Phase, (ii) Detailed Requirements Analysis Phase and (iii) Mapping Phase. The early requirements analysis phase allows for modeling and analyzing the contextual setting of the business domain, in which the DW will operate. In detailed requirements analysis phase, the early requirements specifications are refined with the structural, functional and nonfunctional features of the domain that is relevant to the participants and their role related to the analytical tasks. Moreover, with the aim of the user centric requirements analysis this phase provides a guideline to identification of materialized views [19] for the target DW system. The refinement process is largely influenced by the concepts of Feature Oriented Domain Analysis (FODA) [16]. The mapping phase used to map the DW requirement specifications to the conceptual design model and can starts just after the early requirements analysis phase. The framework supports both demand driven and mixed demand / supply driven requirements analysis approaches for DW system. Also it is supported with abstraction mechanism; reuse capability and mapping facility for requirements descriptions into high level design components.

In the context, the preliminary version of this work has been published in [17]. A BO based requirements analysis framework for generic large scale information system also has been published in [18]. Moreover, the proposed framework has used the graph semantic based conceptual level object oriented multidimensional data model proposed in [14, 15] for its mapping phase. In this context, Graph Object Oriented Multidimensional Data (GOOMD) model provides a novel graph based semantic and simple but powerful algebra to conceptualize the multidimensional data visualization and operational model for OLAP, based on object oriented paradigm.

II. GOOMD Model with Example

In this section, we will summarize the basic concepts of GOOMD model [14, 15]. The GOOMD model is the core of the comprehensive object oriented model of a DW containing all the details that are necessary to specify a data cube, a description of the dimensions, the classification hierarchies, a description fact and measures.

The GOOMD model allows the entire multidimensional database to be viewed as a Graph (V, E) in layered organization. At the lowest layer, each vertex represents an occurrence of an attribute or measure, e.g. product name, day, customer city etc. A set of vertices semantically related is grouped together to construct an Elementary Semantic Group (ESG). So an ESG is a set of all possible instances for a particular attribute or measure. On next, several related ESGs are group together to form a Contextual Semantic Group (CSG) – the constructs to represent any context of business analysis. A set of vertices of any CSG those determine the other vertices of the CSG, is called Determinant Vertices of said CSG. The most inner layer of CSG is the construct of highest level of granularity of fact in Multidimensional database formation.

This layered structure may be further organized by combination of two or more CSGs as well as ESGs to represent next upper level layers and to achieve further lower level granularity of contextual data. From the topmost layer the entire database appears to be a graph with CSGs as vertices and edges between CSGs as the association amongst them. Dimensional Semantic Group (DSG) is a type of CSG to represent a dimension member, which is an encapsulation of one or more ESGs along with extension and / or composition of one or more constituent DSGs. Fact Semantic Group (FSG) is a type of CSG to represent a fact, which is an inheritance of all related DSGs and a set of ESG defined on measures. Two types of edges has been used in GOOMD model, (i) directed edges from DSGs to FSG or constituent DSG to determinant vertex of parent DSG to represent the one – to – many associations and (ii) undirected edges between constituent ESGs and determinant ESGs to represent the association within the members of any CSG.

Since, In order to materialize the cube, one must ascribe values to various measures along all dimensions and can be created from FSG. The cube will also obey a functional constraint f: D1 x D2 x … x Dp → MI. Where any Di is a member of all related top level DSGs and MI is instances of set of measures M. For schema containing multiple FSGs with shared DSGs, the DSG set {D1, D2, … Dp} are the common set of DSGs for all FSGs of the schema.

Let consider an example, based on Sales Application with Sales Amount as measure and with four dimensions – Customer, Model, Time and Location. Model, Time and Location dimensions have upper level hierarchies as Product, QTR and Region respectively. Then in the notation of GOOMD model, there will be four DSGs DSales = {DCustomer, DModel, DLocation, DTime} with hierarchies. Each DSG will be comprised of either a set of ESGs EX ⊆ ESales or a combined set of ESGs and DSGs. As described above, the lower layer DSG will be comprised of ESGs only. The Product DSG DProduct is comprised of only ESGs like EP_ID, EP_NAME and EP_DESC and will be represented as the inner layer of the graph. In the example DModel DSG is an extension of DProduct DSG as well as encapsulation of EM_ID and EM_NAME. The DProduct and DModel DSG graphically can be represented as Figure 1. The FSG for the database can be described as FSales = {DET(DCustomer), DET(DModel), DET(DLocation), DET(DTime), EAMOUNT}. Where

EAMOUNT is the ESGs defined on the measure. The schema from the topmost layer has shown in Figure 2.

GOOMD model also provides algebra of OLAP operators those will operate on different semantic groups. The dSelect (π) operator is an atomic operator and will extract vertices from any CSG depending on some predicate P. The Retrieve (σ) operator extracts vertices from any Cube using some constraint over one or more dimensions or measures. The Retrieve operator is helpful to realize slice and dice operation of OLAP. The Aggregation (α and $+\alpha$) operators perform aggregation on Cube data based on the relational aggregation function like SUM, AVG, MAX etc. on one or more dimensions. Aggregation operators are helpful to realize the roll-up and drill down operations of OLAP. GOOMD model also provides the definitions of the operators like Union ($\cup$), Intersection ($\cap$), Difference (–), Cartesian Product (x) and Join (| x|), which are operated on any CSG or Cube.

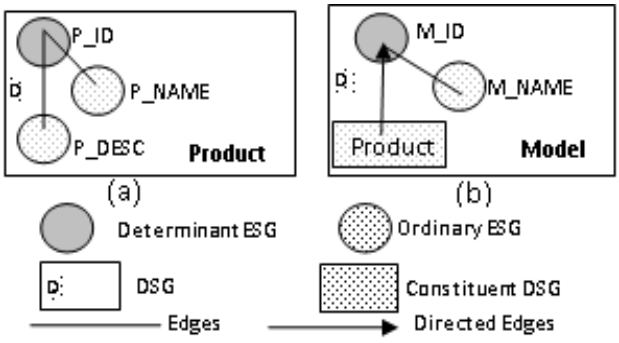


Figure 1. Lowest Level DSG, (b) Higher Level DSG

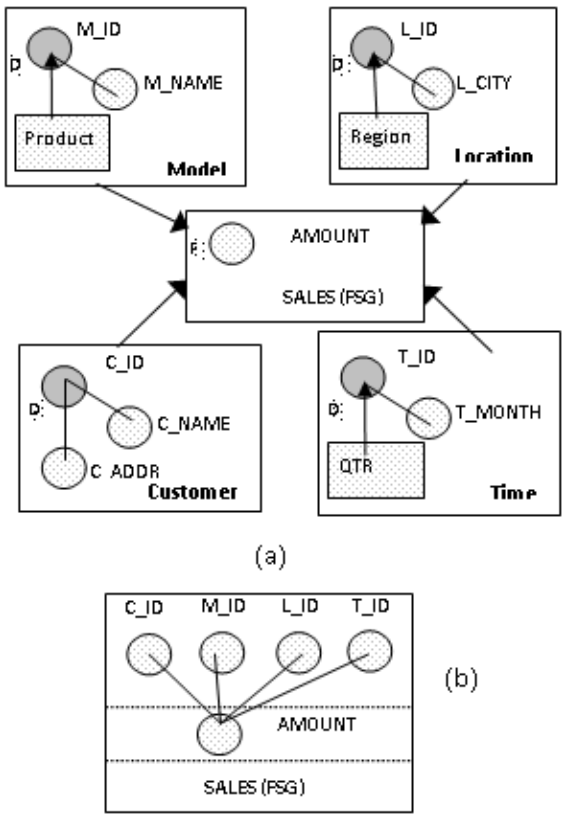


Figure 2. (a) Schema for Sales Application in GOOMD Model
(b) SALES FSG construct after inheritance

III. PROPOSED REQUIREMENTS ANALYSIS FRAMEWORK

The core of the proposed Requirements Analysis Framework for DW system is Business Object Model. A business object (BO) is a conceptual object that has been specified for the purpose of directly describing and representing a business concept with a well-defined boundary and identity. A BO encapsulates identity, domain specific features and behavioral features.

The domain specific features are comprised of structural, functional and nonfunctional features of the domain of interest. The behavior is given by the activities that the business object is capable of performing to fulfill its purpose including the collaboration with other BOs. The primary issue is capturing business semantics (for processes, events and stakeholders) having a common idea or concept that is usable by different parts of a business and by different participants of that business. In the context of DW system, Business Object Modeling is an abstraction technique that consists of identifying the set of concepts and their contexts that belong to some business domain. Such model is comprised of set of BOs to characterize the set of processes and activities of that specific business.

A. Components of Proposed Framework

The part of business object model taxonomies [13], relevant to DW system requirements analysis are as follows,

a) *Business Object (BO):* This is an abstraction that describes a concept of interest in the business itself and capable of being specialized through inheritance mechanism. A BO is the super type of all objects that represent business concepts either entities or activities involved in such specific business domain.

b) Entity BO: This is a specialize form of BO that describes basic business concepts those are engaged in the conduct of business processes. For example stakeholders, products, locations etc. Two entity BOs based on some specific role can collaborate with each other in the context of some business process.

c) Process BO: This is a specialized form of BO that describes a business process or workflow and is comprised of a specified collection of Entity BOs, a pattern of interactions and business events. For example, order fulfillment, procurement, payment etc. Interactions represent and implement activities. The entity BO instances are the actors with specific role and subjects of action. A Process BO can be further refined as collection of related sub process BOs.

d) *Event BO:* This is a specialized form of BO that describes a business event, which may trigger and result from interactions between entity BOs in the context of a process BO. For example, inventory threshold, account overdrawn, end of fiscal year etc. Event BOs are used to capture the business constraints on the interactions.

Besides the above described taxonomies, DW context requires some new taxonomy to be introduced and are as follows,

a) *Measure Attribute:* This is an attribute that describes a quantitative aspect of business process that is relevant for decision making.

b) Fact BO: This is a specialized form of Process BO to capture the concept of subject of analysis specific to some business process. It encapsulates the Measure Attributes along with other features.

c) Dimension BO: This is a specialized form of Entity BO to capture the concept of parameters over which the Fact BO will be analyzed using other Entity BO having relevant role. Any behavioral features will not be included in the Dimension BO. Further a specialized Dimension BO can be formed from generalized one in the proposed framework.

d) *Relations:* several relationship types have been used in the proposed framework like, (i) Encapsulation, (ii) Inheritance, (iii) Collaboration and (iv) Interaction.

The graphical notations for above taxonomies have been summarized in Table 1.

TABLE I. GRAPHICAL NOTATIONS FOR BO-BASED REQUIREMENTS ANALYSIS FRAMEWORK

Taxonomy	Graphical Notation
Entity BO	
Process BO	
Event BO	
Measure Attribute	
Fact BO	F
Dimension BO	D
Encapsulation	
Inheritance	
Collaboration	
Interaction Relation	

B. Early Requirements Analysis Phase

The focus of early requirements analysis phase is to analyze the target business domain. It includes identification along with high level of abstraction of the business processes, different stakeholders, the interactions and collaborations between them, and the events relevant for analytical task specific to the business domain. This phase is important to understand how a business is perceived by its stakeholders mainly decision makers a as set of related business processes and events. This phase is also the basis to identify the possible analytical requirements of decision maker or other stakeholders from the business process driven approach. The phase consists of five steps and are as follows,

a) *Identification of Process and Entity Level Business Objects:* In the first step the relevant business processes with their context, possible stakeholders related to those business processes and interactions between them are represented with high level of abstraction. The business processes and stakeholders can be represented in the proposed framework using Process BOs and Entity BOs respectively. Interactions can be represented between Process BOs and Entity BOs. Interactions represent and implement the activities. In between one Process BO and one Entity BO, more than one activity may exist and this will result in more than one interaction respectively. The interactions between the set of Entity BOs and Process BOs of some specific business can be expressed using business domain level Interaction Diagram. The Interaction Diagram of each interested Process BO can be achieved from business domain level Interaction Diagram.

For example in a Retail Organization, entire business is comprised of several business processes like, (i) Procurements – for procuring the products for sale, (ii) Sales – to handle the customer orders and to sale the products as per order, and (iii) Accounting – to handle the bills, order payment, salaries etc. Several stakeholders may be involved with these business processes, e.g. Sales Manager and Customer may interact with the Sales and Accounting processes with activities like Place Order, Receive Product, Payment, Raise Bill, Bill adjustment etc. Now, the business process will be mapped into the Process BOs and the Stakeholders will be mapped into the Entity BOs. Several other Entity BOs can be involved with each Process BO without having any specific interactions, but they are use to supply important information on the interactions between the business processes and related stakeholders. For example Time, Location, Product, Transaction Type etc. The BOs and interactions relevant to the example can be captured using Interaction Diagram as shown in Figure 3. In the diagram Interactions are labeled with the activities and with direction of initiator and receiver of the interaction. The important point to note that, in this step all Process BOs and related Entity BOs are the representation of high level abstraction for the set of involved business process.

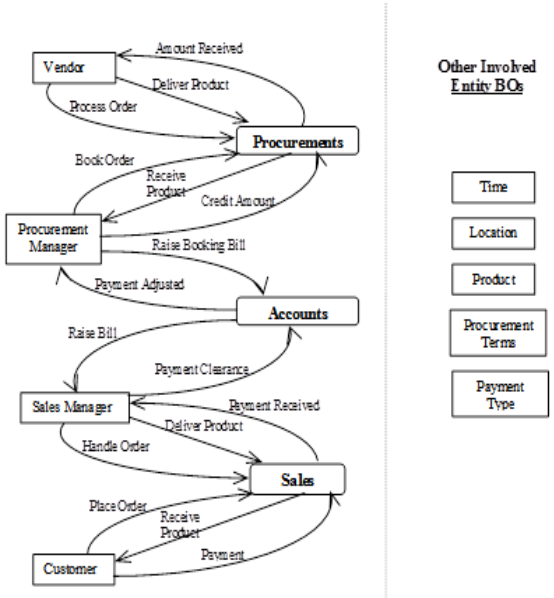


Figure 3. Business Domain Level Interaction Diagram

Now let, the decision makers' requirements include analyze of the Sales of the product. So the Sales Process BO is one of the interesting candidates of target DW system. Two stakeholders may be involved with the process and represented as Entity BOs namely, Sales Manager and Customer. The Interaction Diagram related to Sales Process BO has been shown in Figure 4.

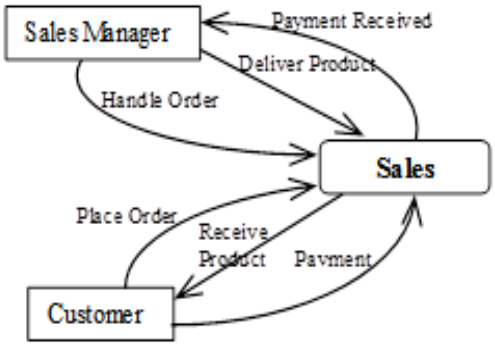


Figure 4. Interaction Diagram for Sales Process BO

b) *Identification of Collaborations:* In this step the collaborations are identified between Entity BOs in the context of some interested Process BO relevant for the purpose of analytical processing. Collaborations are occurred based on some specific roles played by participant Entity BOs. In software engineering term role can be defined as separation of concerns i.e. separation of behavioral characteristic of some Entity BO. The purpose of collaboration is to fulfill some set of activities concerned to the specific Process BO, with which the participant Entity BOs can interact. The roles and collaborations in the context of some Process BO can be identified from Interaction Diagram. This will lead towards developing the Collaboration Diagram in the context of some specific Process BO. Further, for a specific collaboration one can model the activities performed in the context of some Process BO, which will result the Collaboration Interaction Network.

In the above example Sales Manager can play the role as Supplier to fulfill the activities like Handle Order and Deliver Product and also can play the role as Payee for the purpose of activity like Payment Received. Similarly two roles like Client and Payer can be defined for Entity BO Customer. The Collaboration Diagram and Collaboration Interaction Networks are represented in Figure 5.

c) Identification of Measure Attributes and Fact Business Objects: Measure attributes are the quantitative aspects of some Process BO and are relevant for performing analytical tasks on that specific BO. Measure attributes related to each interested Process BO are identified in this step. This step is also the basis of Fact BO construction. Each Process BO relevant to analytical task can be specialized into a Fact BO by encapsulating the related measure attributes.

For example in the running example of Retail Organization for the Sales Process BO two possible measure attributes may

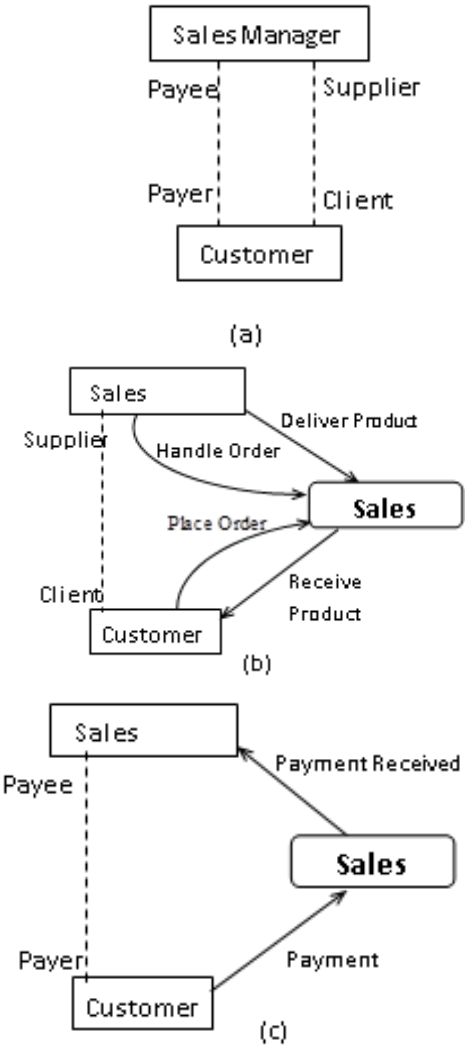


Figure 5. Collaboration Diagram, (b) Collaboration Interaction Network for Supplier / Client Collaboration, (c) Collaboration Interaction Network for Payee / Payer Collaboration

be Quantity and Amount of sales. A Fact BO for Sales can be described from the Process BO Sales by encapsulating the measure attributes Quantity and Amount.

d) Identification of Dimension Business Objects: In this step Entity BOs other than representing stakeholders are identified in the context of some specific Process BO. The set of Entity BOs identified in this step, in general captured the parameters over which the measure attributes related to the specific Process BO are dependent and can be analyzed. Both the set of Entity BOs that captured the stakeholders and other Entity BOs that provides the important information to the Interactions are used to describe the abstraction of Dimension BO in the context of some Fact BO. The set of related Entity BOs which are used to describe the dimension abstraction can be identified from the business domain level interaction diagram of the first step of this phase.

In the Retail Organization example in the context of Sales Fact BO, the possible Dimension BOs can be Time, Location,

Product and Payment Type including Sales Manager and Customer.

e) Identification of Event Business Objects: In this step the Event BOs are identified in the context of some specific Process BO and related Entity BOs. Identification of Event BOs is important to understand the rationale for recording the process BO instances in the DW system for the analytical task. Entity BOs are also important to realize how the specific Process BO will react on some interaction. In the system level view, an Event BO can be realized as a trigger which may fire as a result of specific set of interactions between related Entity BOs and is important in the creation of the DW system.

For example, in the Sales Process BO the possible Event BOs may be Order Processed and Full Payment Received. This step provides the basis of mapping the instances in DW system.

C. Detailed Requirements Analysis Phase

The focus of detailed requirements analysis phase is to refine the Fact BOs and related Dimension BOs identified in early requirements analysis phase to satisfy the analytical requirements of decision makers or other stakeholders. The Fact BOs and Dimension BOs are refined by adding the domain specific features like structural, functional and nonfunctional feature to capture the activities, interactions, collaborations and other stakeholders' requirements. Further this phase provide the guidelines to select materialized views for the target DW system. This phase is also capable to implement the Supply-Driven part of the framework by further refining the identified Fact BOs and Dimension BOs with the comparison with existing operational schemas. The refined Fact BOs and Dimension BOs produced in this case will be more realistic and appropriate for mixed analysis approach of DW system requirements analysis. But it is important to note that the detailed analysis of source operational schemas must be available a prior for this purpose. This phase consists of *four steps* and are as follows,

a) *Refinement of Fact Business Objects:* In this step the Process BOs identified in early requirements phase are refined through two tasks. In the first task, Process BOs are refined as possible collection of related sub Process BOs. The measure attributes identified for top level Process BOs are placed with appropriate sub Process BOs. This task may be iterative in nature to achieve further details sub Process BOs. The Interaction Diagram concerned to the specific Process BO acquired from earlier phase is extended accordingly.

Further the refined Fact BOs can be realized from the refined Process BOs to satisfy stakeholders' expectations. For example, the Sales Process BO may be thought as the composition of two sub-Process BOs Order Processing and Payment Processing. Measure attributes Quantity and Amount can be associated with Order Processing and Payment Processing Process BOs respectively. Henceforth two possible Fact BOs can be realized based on the sub-Process BOs. In fact, each of them encapsulates at least one measure attribute. It will result Extended Interaction Diagram for the Sales Process BO and has been represented in Figure 6.

In the second task, each refined Process BO can be further refined by adding the domain level feature like structural, functional and nonfunctional along with the constraint specifications. The features are the attributes of the system which directly affect the stakeholders. Structural features describe the object level properties of some BO and may also include constraint specification over some features. Functional features describe the operational capabilities of some Process BO. In the context of DW system requirements analysis, functional features of any Process BO include the set of possible analytical operations. Thus functional features are the basis of identification of the possible set of OLAP operations those are required to perform by the decision makers, over the related Fact BO. The prime focus of Nonfunctional features of some Process BO is related with the expected QoS requirements of stakeholders concerned to the specific BO. This may include the features for Security, Performance, Usability etc., specific to some Process BO. Nonfunctional features may be optional for certain Process BO. The features of some Process BO can be represented using Feature Tree Diagram.

Also, Process BO may have specialized Process BOs.

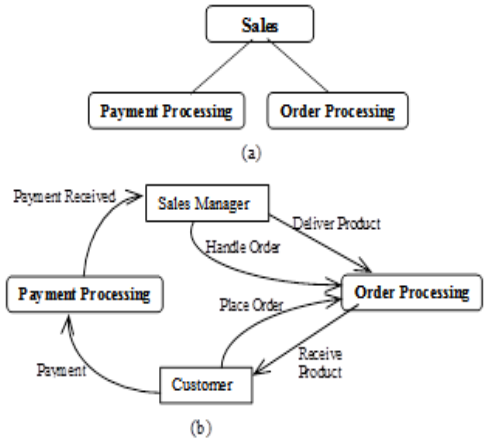


Figure 6. (a) Refined Sales Process BO, (b) Extended Interaction Diagram.

Henceforth a Process BO can contain both basic and derived features. Besides representing the features of some business process as per the stakeholders' perspective, a Feature Tree is capable to represent the logical grouping (AND, OR and EXOR grouping) of the features to satisfy the decision makers or other stakeholders need. Several features of the specific Process BO are the basis of the set of attributes that can be analyzed along with the parameter over which those can be analyzed, in the context of the related Fact BO. Henceforth, the feature tree exhibits the set of features using which the related Fact BO can be associated with concerned Dimension BOs for the formation of conceptual level multidimensional schema for the DW system. The concerned high level Dimension BOs are already identified in the first phase of the framework. Further the feature tree may be accompanied with set of composition rules to express the existing semantics between the subset of features. An example of partial Feature tree for the Process BO, Payment Processing has been shown in Figure 7.

b) Refinement of Dimension Business Objects: In this step, the set of Dimension BOs identified in the early

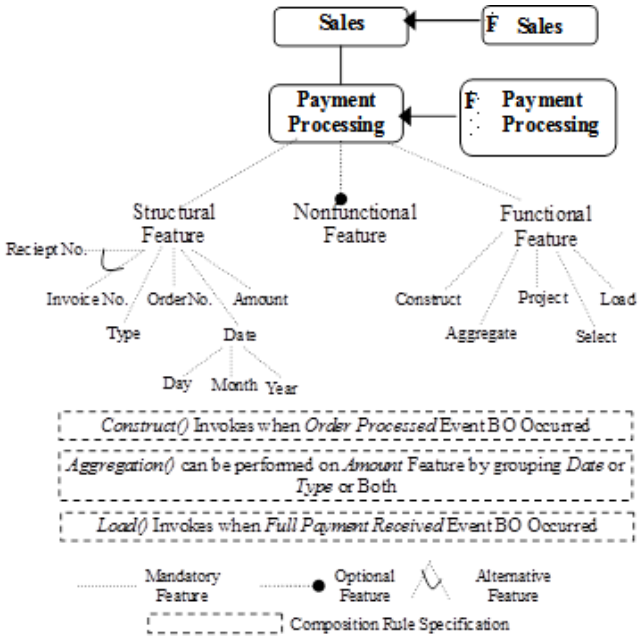


Figure 7. Partial Feature Tree for the Process BO Payment Processing

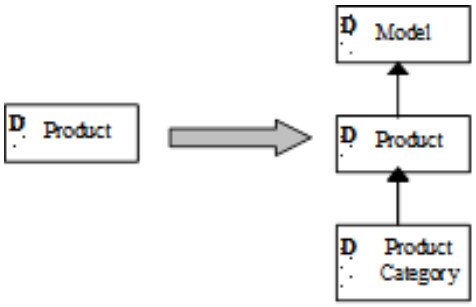


Figure 8. Vertical Refinement of Dimension BO Product

requirements analysis phase are refined. The refinement process involves two tasks. Firstly, the different levels of granularity required for the multidimensional information are identified for some specific business processes. Basis on that, the set of related Dimension BOs need to refine starting from the lowest level granularity to the highest one. The Dimension BO with highest level granularity will exhibit lowest level abstraction in Dimension BO instances. The Dimension BOs with highest level granularity in the context of some Fact BO already have been identified in early requirements analysis phase. This task facilitates the vertical refinement of Dimension BOs and focus on the creation of dimension hierarchy in the context of some fact for the DW system. As an example, the possible vertical refinement of Product Dimension BO in the context of Sales Fact BO has been represented in Figure 8.

In the second task, the refined set of Dimension BOs (each of all hierarchies) are further refined by adding the domain level features in the same way as it has been done in case of Process BOs in the previous step of this phase. It will result the Feature Tree diagram for each Dimension BO. This task facilitates the horizontal refinement of Dimension BOs. The only difference here is, the accompanied composition rules will express the semantics between the subset of features from the specific Dimension BO or in the context of the concerned Fact BO. This will facilitate to realize the different constraint requirements for the DW system.

c) Compare with Operational Schemas: This step is to implement the Supply-Driven part to the framework. The refined Fact BOs and Dimension BOs resulted from the last two steps can further be modified by comparing with the existing operational schemas of the organization. To perform this step a prior knowledge of detailed analysis of operational schemas are required. The set of Fact BOs and Dimension BOs can be refined or modified or also can be filtered by navigating the source operational schemas. Several literatures suggest algorithmic approaches mostly based on the stakeholders' requirements using navigation through the path of many-to-one associations from the attributes of facts.

In fact the steps can be performed by drawing the dependencies between each of Fact BOs and related Dimension BOs, and related operational source schemas. The feature tree diagram of each BO on next can be modified as per the related operational schemas features. The steps are as followed,

1) Draw the dependencies between the Measure Attributes of concerned Fact BO and the Appropriate Attribute of existing source Operational Schemas. It is important to note that the Measure Attribute names may be decided by the decision makers or designers and henceforth attribute names may not match.

2) From each of such Operational Schemas identified in previous step, start navigation using the available foreign keys to other associated Operational Schemas. The each of associated set of Operational Schemas is required to map as Dimensional BO. Dependency can be drawn between the Dimension BOs achieved from the previous steps of the framework to the Operational Schemas achieved on navigation.

3) If dependencies have not been drawn from some Dimension BOs then those BOs are purely demanded by the decision makers and can be labeled as "Demanded Dimension BO". On other hand, where dependencies have drawn from some Dimension BOs, those are labeled as "Supplied Dimension BO".

4) For each Supplied Dimension BO, the attributes of the appropriate operational schemas are compared with Feature Tree of the specific Dimension BO. On comparison structural features in the feature tree can be refined or modified as per the attribute details of the related operational schemas.

5) For each operational schema, identified in step (ii) but related Dimension BO does not exist, are required to map as New Dimension BO. The feature tree of new Dimension BO can be drawn from the attributes descriptions of the related operational schemas.

6) The feature tree of each Fact BO can be modified further from the available Dimension BOs and newly achieved Dimension BOs from the step (v).

d) *Identification of Materialized Views:* For user oriented DW requirements engineering, it is also important to analyze that how user will efficiently interact with the DW system to perform the necessary analysis activities. Materialized views are the central issue for the usability of the DW system. DW data are organized multi dimensionally to support OLAP. A DW can be seen as a set of materialized views defined over the source relations. Those views are frequently evaluated by the user queries. The materialized views need to be updated when the source relations change. During the DW analysis and design, the initial materialized view need to be selected to make the user's interactions simple and efficient in terms of accomplishing user analysis objectives. In the proposed requirements engineering framework, the domain boundary has been drawn through identifications of Fact BOs, Dimension BOs, Actor BOs, and interactions between them in the first phase and which have been further refined in this phase. The list of analysis activities may be performed by Actor BOs based on their roles and also Event BOs have been identified in the same phase. Moreover, the feature tree concept explores the constraint requirements for the interest of domain. Based on those identifications, the different materialized views can be identified in this step. In this step the materialized views are used to represent semantically in the context of some Fact BO and in terms of actor along with their roles, analysis activities those may be performed, events those may be occurred, related Dimension BOs involved and the related constraints. Related to one Fact BO, there may exist several materialized views to minimize the views level dependency and to meet the analytical evaluation requirements of the stake holders. Semantically, a materialized view will be represented using View Template. The Interface Template will contain the information of View name, identification, analysis objectives, target Fact BO, Actor BO, roles, related activities, related Dimension BOs to realize the source relations, related Event BOs and related constraints. Any view template is reusable and modifiable through iterative process to accommodate the updatable materialized view.

In case of *Retail Organization*, to interact with the *Sales* Fact BO, one example view template related to *Customer* has been shown in Figure 9.

View Name and ID: Customer View 1 ***Target Fact BO:*** Sales ***Requirements Objective:*** Analysis of Order Placing of Location XXX ***Target Measures:*** Quantity					
Actor BO: Customer	***Role:*** Client	***Activities:*** Place Order; Receive Products;	***Dimension BOs:*** Time; Product; Procurements Term ;	***Event BOs:*** Order Status	***Constraint:*** Construct() invoke.

Figure 9. View Template related to Sales Fact BO

IV. Requirements Analysis to Conceptual Model

This section will describe the Mapping Phase of our framework. In this phase the DW requirements specifications achieved from the early requirements analysis phase and consequently from the detailed requirements analysis phase are mapped into conceptual level multidimensional design model. In this phase, we will map the requirements specifications in the constructs and concepts of GOOMD model described in Section 2. But there is no binding on mapping the proposed requirements framework to any multidimensional conceptual model. This phase can starts just after the early requirements analysis phase and consists of two levels, namely, Early Mapping and Detailed Mapping.

a) *Early Mapping:* Early mapping can be done just after the early requirements analysis phase for DW system. The early requirements analysis phase is business process driven approach and is used to identify the set of Fact BOs and related Dimension BOs relevant to analytical requirements of decision makers and other stakeholders. Each identified Fact BO and encapsulated Measure Attributes will be mapped as Fact Semantic Group (FSG) and Elementary Semantic Group (ESG) for measures respectively as described in GOOMD model concept. Each Dimension BO related to the specific Fact BO can be mapped as Dimension Semantic Group (DSG). On next, each DSG need to be connected with the FSG using Link. The early mapping will yield the topmost layer of the GOOMD model Schema which exhibit high level abstraction.

b) *Detailed Mapping:* In detailed mapping steps, the multidimensional schema achieved from the early mapping can be further refined to deliver the full-fledged conceptual schema for DW system. The topmost layer GOOMD model schema will be further modified in this step, from the specifications available from detailed requirements analysis phase. Firstly, the Dimension BOs of different granularity related to each top level Dimension BO are mapped as separate DSGs and are connected using Link to form the dimension hierarchies. On next, the basic structural features and identity from the Feature Tree of each Dimension BO are mapped as ESG and Determinant ESG respectively. The related ESGs and Determinant ESGs are connected using Association and need to be encapsulated in the specific DSG. Finally, each FSG can be modified from the feature tree of related FO by extracting the relevant features into set of ESGs and by encapsulating those ESGs. This step will exhibit all the inner layers of GOOMD model schema.

The descriptions of possible OLAP operations for the refined GOOMD model schema can be guided from the Functional features of related set of Fact BO and Dimension BOs. The DW System QoS requirements can be realized from nonfunctional features of Fact BO.

V. Features of Proposed Requirements Framework

The proposed requirements analysis framework for DW system has been drawn from concepts of business object model. Besides the advantages of using the business object concept, the proposed framework facilitates several other features and are as follows,

a) Process Driven Approach: The proposed requirements analysis framework for DW system starts with

business process driven approach. The early requirements analysis phase in the framework identify the concerned business processes, events, stakeholders and interactions between them possibly relevant to analytical task. The phase is not biased from perspective of stakeholders' understanding about the system. Rather it ensures the modeling of the business concept in high level of abstraction, how it is exist in reality. So it is more realistic approach.

b) *Object Orientation:* The core of the proposed framework is Business Object model, which supports the general concepts and characteristics of object oriented paradigm. Further, the requirements specification resulted from the proposed framework can be mapped into any object oriented conceptual level multidimensional data model (GOOMD model as an example).

c) *Abstraction:* The proposed requirements analysis framework is capable to represent the different business concepts and stakeholders' requirements in different level of abstraction. Most abstract description of the analytical requirements in the context of DW system is available from the early requirements analysis phase of the framework. Moreover the early mapping, which can be performed just after the early requirements analysis phase, will produce the conceptual multidimensional data model schemas with high level of abstraction. In the detailed requirements framework the early requirements descriptions are refined stepwise to lower the abstraction level (see Table 2).

TABLE II. ABSTRACTION AND REUSE POTENTIAL IN EACH PHASE OF PROPOSED REQUIREMENTS ANALYSIS FRAMEWORK

Base Concepts and Phases for proposed Requirement analysis framework	**Level of**		
	Abstraction	**Reuse Potential**	**Productivity**
Business Object (Base Concept)	High	High	Low
Early Requirements Analysis Phase	↓	↓	↓
Early Mapping Phase	↓	↓	↓
Detailed Requirements Analysis Phase	↓	↓	↓
Detailed Mapping Phase	Low	Low	High

d) *Reusability:* One of the major advantages of the proposed framework is that, it supports reuse of domain level abstractions and step wise refinement mechanisms for mapping the DW requirements in high level design. This is important because the anticipation of future requirements for the decision-making process is very difficult in large system like DW. The support of abstraction mechanism and feature oriented stepwise refinement of the BO based requirements descriptions in detailed requirements analysis phase enable the capability of reuse (see Table 2) of different types of BOs specific to some business. A new BO can be formed from the existing BO either at high level of abstraction in early requirements analysis phase or by adding new features in detailed analysis phase. The refinement processes of BOs are iterative in nature.

e) *Support for Multiple Analysis Approaches:* The proposed framework supports both Demand-driven and Mixed analysis approaches for the requirements analysis of DW system. Also it is capable to map the requirements to object oriented multidimensional conceptual model by transforming different detailed BOs into the relevant high level components of the design model.

In the proposed framework, the early requirements analysis phase and subsequent early mapping step are basically business process driven and independent of any specific DW requirements analysis approaches. But in detailed requirements analysis, the step for comparison of refined Fact BO and Dimension BO with the operational schemas [subsection 3.C] is to support supply-driven part in the proposed framework. By omitting this step from the detailed requirements analysis phase, the proposed framework is fully compatible with *Demand-Driven* DW requirements approach. But with the presence of that step, the resultant GOOMD model schema from the detailed mapping phase from refined Process BOs and related Dimension BOs, are largely influenced by the source operational schemas specifications. In that case the proposed framework will support *Mixed* Analysis approach towards DW requirements analysis.

VI. CONCLUSION

Requirements analysis plays a key role within DW system development with the aim to reduce the risk of failure. A good DW design method should be preceded by the requirements elicitation and their analysis methodologies by considering both user requirements and operational data sources for data warehouse development. For the purpose, the Business Object based requirements analysis framework for DW system has been devised. The framework is comprised of three phases namely, Early Requirements Analysis phase, Detailed Requirements Analysis phase and Mapping phase. It starts from understanding the set of business processes, events and stakeholders in terms of set of well-defined BOs concerned to some business domain in which DW system will operate. In the framework, several requirements modeling elements like Fact BO, Dimension BO, Measure attributes, Interaction Diagram, Collaboration Diagram, Interaction Collaboration Network, Feature Tree etc. have been described to express different business concepts of the domain, relevant to DW system and in the real business scenario. Finally the framework results the mapping of DW requirements descriptions in high level design components of conceptual level object oriented multidimensional data model. The proposed framework supports abstraction mechanism and reuse of different well defined elements those have been used to realize the different business concepts of the domain and useful for analytical task. These features enable the framework to be used efficiently in the evolving business processes over time. Further the framework supports both demand-driven and mixed analysis approach of DW requirements analysis.

Future work will include developing of a prototype tool in the support of the proposed DW requirements analysis framework. Moreover, quality evaluation of the proposed

requirements analysis framework also is a prime objective of the future work.

REFERENCES

[1] R. Kimball, M. Ross, "The Data Warehouse Toolkit", Book Wiley & Sons (2002).

[2] R. Winter and B. Strauch, "A method for demand-driven information requirements analysis in data warehousing projects", In Proc. HICSS, PP 1359–1365, 2003.

[3] F. R. S. Paim and J. B. Castro, "DWARF: An approach for requirements definition and management of data warehouse systems", Proc. of Int.. Conf. on Requirements Engineering, 2003.

[4] Robert Winter, Bernhard Strauch, "Information requirements engineering for data warehouse systems", Proceedings of the ACM Symposium on Applied Computing, 2004.

[5] Rizzi, S., Abell´o, A., Lechtenb¨orger, J., Trujillo, J., "Research in data warehouse modeling and design: dead or alive?" Proc. of the 9th ACM Int. Workshop on Data warehousing and OLAP, PP 3–10, 2006.

[6] M. Golfarelli, D. Maio, S. Rizzi, "The dimensional fact model: A conceptual model for data warehouses", Intl' Journal of Cooperative Information Systems, Vol. 7(2-3), PP 215–247, 1998.

[7] B. H¨usemann, J. Lechtenb¨orger, and G. Vossen. "Conceptual data warehouse design", Proc. 2nd Int. Workshop on Design and Management of Data Warehouse, PP 3–9, 2000.

[8] D. Moody and M. Kortink, "From enterprise models to dimensional models: A methodology for data warehouse and data mart design", Proc. 2nd Int. Workshop on Design and Management of Data Warehouse, 2000.

[9] Paolo Giorgini, Stefano Rizzi, Maddalena Garzetti, "GRAnD: A goal-oriented approach to requirement analysis in data warehouses", Decision Support Systems, Vol. 45(1), PP 4-21, 2008.

[10] N. Prakash and A. Gosain, "Requirements Driven Data Warehouse Development", Proc. of CAiSE Short Paper, 2003.

[11] Jose-Norberto Mazón, Juan Trujillo, Jens Lechtenbörger, "Reconciling Requirement-driven Data warehouses with Data Sources via Multidimensional Normal Forms", Data & Knowledge Engineering, Vol. 63(3), PP 725-751, 2007.

[12] Jose-Norberto Mazón and Juan Trujillo, "An MDA approach for the development of data warehouses", Decision Support Systems, Vol. 45(1), PP 41-58, 2008.

[13] OMG, Business Object DTF – Common Business Objects, OMG Document bom/97-11-11, ftp://ftp.omg.org/pub/docs/bom/97-11-11.pdf, 1997.

[14] A. Sarkar, S. Bhattacharya, "The Graph Object Oriented Multidimensional Data Model: A Conceptual Perspective", 16th Int. Conf. on Software Engineering and Data Engineering (SEDE 2007), PP 165 – 170, 2007.

[15] A. Sarkar, S. Choudhury, N. Chaki, S. Bhattacharya, "Conceptual Level Design of Object Oriented Data Warehouse: Graph Semantic Based Model", International Journal of Computer Science (INFOCOMP), Vol. 8(4), PP 60 – 70, December 2009.

[16] K. C. Kang, S. G. Cohen, J. A. Hess, W. E. Novak, A. Spencer Peterson, "Feature-Oriented Domain Analysis (FODA) Feasibility Study", Technical Report, Software Engineering Institute, Carnegie Mellon University (USA), 1990.

[17] A. Sarkar, S. Choudhury, N. Chaki, S. Bhattacharya, "Business-Object Oriented Requirements Analysis Framework for Data Warehouses", 22nd International Conference on Software Engineering and Knowledge Engineering (SEKE 2010), PP 34 – 37, July 2010.

[18] A. Sarkar, N. C. Debnath, "Business Object Oriented Requirements Analysis for Large Scale Information System", 20th International Conference on Software Engineering and Data Engineering (SEDE 2011), PP 103 - 108, June 2011.

[19] A. C. Dhote, M. S. Ali, "Materialized View Selection in Data Warehousing", 4th International Conference on Information Technology (ITNG '07), PP 843 – 847, 2007.

AUTHORS PROFILE

Anirban Sarkar is presently a faculty member in the Department of Computer Applications, National Institute of Technology, Durgapur, India. He received his PhD degree from National Institute of Technology, Durgapur, India in 2010. His areas of research interests are Database Systems and Software Engineering. His total numbers of publications in various international platforms are about 30.

A new graph based text segmentation using Wikipedia for automatic text summarization

Mohsen Pourvali
Department of Electrical & Computer Engineering at
Qazvin Branch Islamic Azad University
Qazvin, Iran

Ph.D. Mohammad Saniee Abadeh
Department of Electrical & Computer Engineering at
Tarbiat Modares University
Tehran, Iran

***Abstract*—The technology of automatic document summarization is maturing and may provide a solution to the information overload problem. Nowadays, document summarization plays an important role in information retrieval. With a large volume of documents, presenting the user with a summary of each document greatly facilitates the task of finding the desired documents. Document summarization is a process of automatically creating a compressed version of a given document that provides useful information to users, and multi-document summarization is to produce a summary delivering the majority of information content from a set of documents about an explicit or implicit main topic. According to the input text, in this paper we use the knowledge base of Wikipedia and the words of the main text to create independent graphs. We will then determine the important of graphs. Then we are specified importance of graph and sentences that have topics with high importance. Finally, we extract sentences with high importance. The experimental results on an open benchmark datasets from DUC01 and DUC02 show that our proposed approach can improve the performance compared to state-of-the-art summarization approaches.**

Keywords- text Summarization; Data Mining; Word Sense Disambiguation.

I. INTRODUCTION

The technology of automatic document summarization is maturing and may provide a solution to the information overload problem. Nowadays, document summarization plays an important role in information retrieval (IR). With a large volume of documents, presenting the user with a summary of each document greatly facilitates the task of finding the desired documents. Text summarization is the process of automatically creating a compressed version of a given text that provides useful information to users, and multi-document summarization is to produce a summary delivering the majority of information content from a set of documents about an explicit or implicit main topic [14]. Authors of the paper [10] provide the following definition for a summary: "A summary can be loosely defined as a text that is produced from one or more texts that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that. Text here is used rather loosely and can refer to speech, multimedia documents, hypertext, etc. The main goal of a summary is to present the main ideas in a document in less space. If all sentences in a text document were of equal importance, producing a summary would not be very effective, as any reduction in the size of a document would carry a proportional decrease in its in formativeness. Luckily, information content in a document appears in bursts, and one can therefore distinguish between more and less informative segments. Identifying the informative segments at the expense of the rest is the main challenge in summarization". Assume a tripartite processing model distinguishing three stages: source text interpretation to obtain a source representation, source representation transformation to summary representation, and summary text generation from the summary representation. A variety of document summarization methods have been developed recently.

The paper [4] reviews research on automatic summarizing over the last decade. This paper reviews salient notions and developments, and seeks to assess the state of-the-art for this challenging natural language processing (NLP) task. The review shows that some useful summarizing for various purposes can already be done but also, not surprisingly, that there is a huge amount more to do. Sentence based extractive summarization techniques are commonly used in automatic summarization to produce extractive summaries. Systems for extractive summarization are typically based on technique for sentence extraction, and attempt to identify the set of sentences that are most important for the overall understanding of a given document. In paper [11] proposed paragraph extraction from a document based on intra-document links between paragraphs. It yields a text relationship map (TRM) from intra-links, which indicate that the linked texts are semantically related. It proposes four strategies from the TRM: bushy path, depth-first path, segmented bushy path, augmented segmented bushy path.

In our study we focus on sentence based extractive summarization. We express that the lexical cohesion structure of the text can be exploited to determine the importance of a sentence. Eliminate the ambiguity of the word has a significant impact on the inference sentence. In this article we will show that the separation text into the inside issues by using the correct concept Noticeable effect on the summary text is created. We have used Word Sense Disambiguation [8] for select correct concept. The experimental results on an open benchmark datasets from DUC01 and DUC02 show that our proposed approach can improve the performance compared to state-of-the-art summarization approaches.

II. Related Work

Generally speaking, the methods can be either extractive summarization or abstractive summarization. Extractive summarization involves assigning salience scores to some units (e.g. sentences, paragraphs) of the document and extracting the sentences with highest scores, while abstraction summarization (e.g.http://www1.cs.columbia.edu/nlp/newsblaster/) usually needs information fusion, sentence compression and reformulation [14]. Sentence extraction summarization systems take as input a collection of sentences (one or more documents) and select some subset for output into a summary. This is best treated as a sentence ranking problem, which allows for varying thresholds to meet varying summary length requirements. Most commonly, such ranking approaches use some kind of similarity or centrality metric to rank sentences for inclusion in the summary – see, for example [1].The centroid-based method [3] is one of the most popular extractive summarization methods.

MEAD (http://www.summarization.com/mead/) is an implementation of the centroid-based method for either single-or-multi-document summarizing. It is based on sentence extraction. For each sentence in a cluster of related documents, MEAD computes three features and uses a linear combination of the three to determine what sentences are most salient. The three features used are centroid score, position, and overlap with first sentence (which may happen to be the title of a document). For single-documents or (given) clusters it computes centroid topic characterizations using tf–idf-type data. It ranks candidate summary sentences by combining sentence scores against centroid, text position value, and tf–idf title/lead overlap. Sentence selection is constrained by a summary length threshold, and redundant new sentences avoided by checking cosine similarity against prior ones. In the past, extractive summarizers have been mostly based on scoring sentences in the source document. In paper [12] each document is considered as a sequence of sentences and the objective of extractive summarization is to label the sentences in the sequence with 1 and 0, where a label of 1 indicates that a sentence is a summary sentence while 0 denotes a non-summary sentence. To accomplish this task, is applied conditional random field, which is a state-of-the-art sequence labeling method.

In paper [15] proposed a novel extractive approach based on manifold–ranking of sentences to query-based multi-document summarization. The proposed approach first employs the manifold–ranking process to compute the manifold–ranking score for each sentence that denotes the biased information-richness of the sentence, and then uses greedy algorithm to penalize the sentences with highest overall scores, which are deemed both informative and novel, and highly biased to the given query.

The summarization techniques can be classified into two groups: supervised techniques that rely on pre-existing document-summary pairs, and unsupervised techniques, based on properties and heuristics derived from the text. Supervised extractive summarization techniques treat the summarization task as a two-class classification problem at the sentence level, where the summary sentences are positive samples while the non-summary sentences are negative samples. After representing each sentence by a vector of features, the classification function can be trained in two different manners [7]. One is in a discriminative way with well-known algorithms such as support vector machine (SVM) [16]. Many unsupervised methods have been developed for document summarization by exploiting different features and relationships of the sentences – see, for example [3] and the references therein. On the other hand, summarization task can also be categorized as either generic or query-based. A query-based summary presents the information that is most relevant to the given queries [2] and [14] while a generic summary gives an overall sense of the document's content [2] , [4] , [12] , [14].

The QCS system (Query, Cluster, and Summarize) [2] perform the following tasks in response to a query: retrieves relevant documents; separates the retrieved documents into clusters by topic, and creates a summary for each cluster. QCS is a tool for document retrieval that presents results in a format so that a user can quickly identify a set of documents of interest. In paper [17] are developed a generic, a query-based, and a hybrid summarizer, each with differing amounts of document context. The generic summarizer used a blend of discourse information and information obtained through traditional surface-level analysis. The query-based summarizer used only query-term information, and the hybrid summarizer used some discourse information along with query-term information. The article [18] presents a multi-document, multi-lingual, theme-based summarization system based on modeling text cohesion.

III. Create Graph and Text Segmentation

The algorithm presented in this paper, at first the input text is pre-processing and the stop words is removed. Then stem of words is found and its (POS) is tagged.

Only verbs and nouns are used in the text, in the way we have presented. The algorithm starts from the beginning of the main text, and take the word, and using Wikipedia knowledge base provides a two-level tree from the links of the word' abstract. So that root of the word is the same word and tree Children are related words (links) to the target word in the abstract of its web page. Then it searches the children of the target word in the input text and it creates a graph using target word and the words that both are in the children of the previous step tree and input text.

Let s = {s_1, s_2, ..., s_n} is the set of sentences and $w = \{w_{11}, w_{21}, \dots, w_{ij}, \dots, w_{kn}\}$ is the set of all words are nouns or verbs in the input text. So that w_{ij} shows i-th word in the j-th sentence. Since the goal is to extract sentences with high importance. The sentences are considered as nodes. The relationships between words within a sentence with other sentences words are considered to be edges in the graph. The algorithm is shown in Figure 1.

```
For n=0 to EndOfSentenc
   For i=0 to EndOfSentenc_n
      Child[] = CreateTreeInWiki(w_in)
         For r=0 to EndOfChild
            For k=0 to EndOfWord
                  If Child[r] == AnyWordOf_W
                    Graph[] = Create_Or_Update_Graph(S_r, S_k)
               Endif
            EndFor
         EndFor
   EndFor
EndFor
```

Figure 1. Base algorithm for create the tree and graph

In the above algorithm, ***Child*** is the children of target word tree in the Wikipedia, and ***Graph*** is the constructed graph from the sentences that target word is in them. This algorithm is implemented for all target words in the input text. Finally, we have several independent graphs, that according to the relationship between its nodes, each graph implies a topic in the input text. Figure 2 shows related sentences in the text.

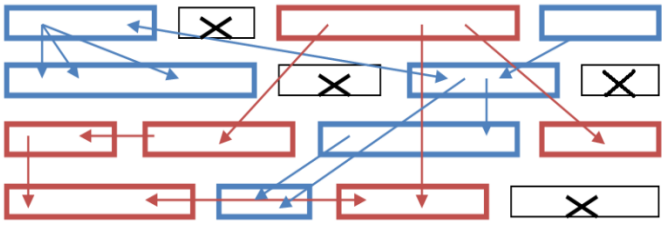


Figure 2. Related sentences and segments, there are two segments with two colors (blow and red)

After extracting the graphs of the input text, the graphs edges were given weight. According to the distance between the words in two sentences, existed in the two sides of the edge, the weighting to the edge is done. To do this we use Average Google normalized distance [19]. NGD takes advantage of the number of hits returned by Google to compute the semantic distance between concepts. The concepts are represented with their labels which are fed to the Google search engine as search terms.

First, using the NGD we define the global and local dissimilarity measure between terms (as shown in [19] the NGD is nonnegative and does not satisfy the triangle inequality, i.e. hence isn't distance and consequently in the further it we shall name dissimilarity measure). According to definition NGD the global dissimilarity measure between terms t_k and t_l also is defined by the formula:

$$NGD^{global}(t_k,t_l)=\frac{max\{\log(f_k^{global}),\log(f_l^{global})\}-\log(f_{kl}^{global})}{\log N_{Google}-min\{\log(f_k^{global}),\log(f_l^{global})\}} \quad (1)$$

Where f_k^{global} is the number of web pages containing the search term t_k, and f_{kl}^{global} denotes the number of web pages containing both terms t_k and t_l, N_{Google} is the number of web pages indexed by Google. The main properties of the NGD [19] are listed as follows:

1) The range of the NGD is in 0 and ∞;

- If $t_k = t_l$ or if $t_k \neq t_l$ but frequency $f_k^{global} = f_l^{global} = f_{kl}^{global} > 0$, Then $NGD^{global}(t_k,t_l) = 0$. That is, the semantics of t_k and t_l, in the Google sense is the same.
- If frequency $f_k^{global} = 0$, then for every term t_k, we have $f_{kl}^{global} = 0$, and the $NGD^{global}(t_k,t_l) = \frac{\infty}{\infty}$, which we take to be 1 by definition.
- If frequency $f_k^{global} \neq 0$ and $f_{kl}^{global} = 0$, we take $NGD^{global}(t_k,t_l) = 1$.

2) NGD (tk,tk) = 0 for every tk. For every pair t_k and t_l, we have $NGD^{global}(t_k,t_l) = NGD^{global}(t_l,t_k)$: It is symmetric.

Formula 3 is the dissimilarity measure between sentences S_i and S_j.

$$\mathrm{diss}_{\mathrm{NGD}}^{\mathrm{global}}(S_i,S_j)=\frac{\sum_{t_k\in S_i}\sum_{t_l\in S_j}\mathrm{NGD}^{\mathrm{global}}(t_k,t_l)}{m_i m_j} \quad (2)$$

That m_i and m_j are the number of words in i-th and j-th sentences. Then, the weighting of the graph, we are selecting the heavier graph (the graph that has heavy nods and light edges). Using the following formula a weight is assigned to each graph.

$$V_g=\frac{1}{L}\times\frac{\sum_{i=1}^{L}F_i\times d_i}{\sum_{j=1}^{E}e_j} \quad (3)$$

That L is number of nodes and E is number of edges in any graph, d_i is the degree of i-th node.

IV. Sentence Extraction

Finally, the graph which is higher than other graphs contains the main topic of the text. In formula 1, sentences can be extracted according to the percent of summarization. If we want to have the summarization of other topics in addition to main topic in the text we extract important sentences from the important graph according to the summarization percent. Using the following formula, each node is evaluated according to its number of incoming and outgoing edges.

$$F_i=\frac{(I_i+O_i)}{L}\times\sum_{t=1}^{m}W_{t_i} \quad (4)$$

Where O_i is number of outputs from i-th sentence and I_i is number of inputs to i-th sentence. We use following formula to calculate the weight of the word W_{ti}.

$$w_{ti}=TF_{ti}\times ISF_{ti} \quad (5)$$

That TF_{ti} is the number of occurrences phrase t in the sentence S_i, and ISF is:

$$ISF=\log(\frac{N}{N_t}) \quad (6)$$

N_t is the number of sentences the word t_i has occurred in it.

V. EXPERIMENTS AND RESULTS

In this section, we conduct experiments to test our summarization method empirically.

A. Datasets

For evaluation the performance of our methods we used two document datasets DUC01 and DUC02 and corresponding 100-word summaries generated for each of documents. The DUC01 and DUC02 are an open benchmark datasets which contain 147 and 567 documents-summary pairs from Document Understanding Conference (http://duc.nist.gov). We use them because they are for generic single-document extraction that we are interested in and they are well preprocessed. These datasets DUC01 and DUC02 are clustered into 30 and 59 topics, respectively. In those document datasets, stop words were removed using the stop list provided in ftp://ftp.cs.cornell.e-du/pub/smart/english.stop and the terms were stemmed using Porter's scheme [9], which is a commonly used algorithm for word stemming in English.

B. Evaluation metrics

There are many measures that can calculate the topical similarities between two summaries. For evaluation the results we use two methods. The first one is by precision (P), recall (R) and F1-measure which are widely used in Information Retrieval. For each document, the manually extracted sentences are considered as the reference summary (denoted by $Summ_{ref}$). This approach compares the candidate summary (denoted by $Summ_{cand}$) with the reference summary and computes the P, R and F1-measure values as shown in formula (9) [12].

$$P = \frac{|summ_{ref} \cap summ_{cand}|}{|summ_{cand}|} \quad (7)$$

$$R = \frac{|summ_{ref} \cap summ_{cand}|}{|summ_{ref}|} \quad (8)$$

$$F_1 = \frac{2PR}{P+R} \quad (9)$$

The second measure we use the ROUGE toolkit [5], [6] for evaluation, which was adopted by DUC for automatically summarization evaluation. It has been shown that ROUGE is very effective for measuring document summarization. It measures summary quality by counting overlapping units such as the N-gram, word sequences and word pairs between the candidate summary and the reference summary. The ROUGE-N measure compares N-grams of two summaries, and counts the number of matches. The measure is defined by formula (10) [5], [6].

$$ROUGE - N = \frac{\sum_{S \in summ_{ref}} \sum_{N-gram \in S} Count_{match}(N-gram)}{\sum_{S \in summ_{ref}} \sum_{N-gram \in S} Count(N-gram)} \quad (10)$$

Where N stands for the length of the N-gram, $Count_{match}$ (N-gram) is the maximum number of N-grams co-occurring in candidate summary and a set of reference–summaries. Count (N_gram) is the number of N-grams in the reference summaries. We use two of the ROUGE metrics in the experimental results, ROUGE-1 (unigram-based) and ROUGE-2 (bigram-based).

C. Simulation strategy and parameters

The parameters of our method are set as follows: depth of tree that is created for any word, *n=3*; extra value for *Lesk* algorithm, λ *=5*; Finally, we would like to point out that algorithm was developed from scratch in C#.net 2008 platform on a Pentium Dual CPU, 1.6 GHz PC, with 512 KB cache, and 1 GB of main memory in Windows XP environment.

D. Performance evaluation and discussion

We compared our method with four methods CRF [12], NetSum [13], Manifold–Ranking [15] and SVM [16]. Tables 1 and 2 show the results of all the methods in terms ROUGE-1, ROUGE-2, and F1-measure metrics on DUC01 and DUC02 datasets, respectively. As shown in Tables 1 and 2, on DUC01 dataset, the average values of ROUGE-1, ROUGE-2 and F1 metrics of all the methods are better than on DUC02 dataset. As seen from Tables 1 and 2 Manifold–Ranking is the worst method, In the Tables 1 and 2 highlighted (bold italic) entries represent the best performing methods in terms of average evaluation metrics. Among the methods NetSum, CRF, SVM and Manifold–Ranking the best result shows NetSum. We use relative improvement $\frac{(our\ method - other\ methods)}{other\ methods} \times 100$ for comparison. Compared with the best method NetSum, on DUC01 (DUC02) dataset our method improves the performance by 3.43% (4.82%), 7.15% (16.30%) and 3.12% (4.28%) in terms ROUGE-1, ROUGE-2 and F1, respectively.

TABLE I. AVERAGE VALUES OF EVALUATION METRICS FOR SUMMARIZATION METHODS (DUC01 DATASET).

Methods	Av.ROUGE-1	Av.ROUGE-2	Av.F1-measure
Our method	0.48021	0.18962	0.48743
NetSum	0.46427	0.17697	0.47267
CRF	0.45512	0.17327	0.46435
SVM	0.44628	0.17018	0.45357
Manifold–Ranking	0.43359	0.16635	0.44368

TABLE II. AVERAGE VALUES OF EVALUATION METRICS FOR SUMMARIZATION METHODS (DUC02 DATASET).

Methods	Av.ROUGE-1	Av.ROUGE-2	Av.F1-measure
Our method	0.47129	0.12986	0.48259
NetSum	0.44963	0.11167	0.46278
CRF	0.44006	0.10924	0.46046
SVM	0.43235	0.10867	0.43095
Manifold–Ranking	0.42325	0.10677	0.41657

VI. CONCLUSION

We have presented the approach to automatic document summarization based on creating graph and text segmentation and extraction of sentences using Wikipedia. Our approach consists of two steps. First creates a two-level tree from the links of the word's abstract, and then creates graph using of previous phase, and finally selects important segments which were created using of previous graph. When comparing our methods with several existing summarization methods on an open DUC01 and DUC01 datasets, we found that our methods can improve the summarization results significantly. The

methods were evaluated using ROUGE-1, ROUGE-2 and F1 metrics. In this paper we also demonstrated that the summarization result depends on the similarity measure. Results of experiment have showed that proposed by us NGD-based dissimilarity measure outperforms the Euclidean distance.

REFERENCES

[1] Alguliev, R. M., & Alyguliev, R. M. (2007). Summarization of text-based documents with a determination of latent topical sections and information-rich sentences. *Automatic Control and Computer Sciences , 41*, 132–140.

[2] Dunlavy, D. M., O'Leary, D. P., Conroy, J. M., & Schlesinger, J. D. (2007). QCS: A system for querying, clustering and summarizing documents. *Information Processing and Management , 43*, 1588–1605.

[3] Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research , 22*, 457–479.

[4] Jones, K. S. (2007). Automatic summarizing: The state of the art. *Information Processing and Management , 43*, 1449–1481.

[5] Lin, C. -Y. (2004). ROUGE: A package for automatic evaluation summaries. *In Proceedings of the workshop on text summarization branches out*, (pp. 74–81). Barcelona, Spain.

[6] Lin, C. -Y., & Hovy, E. H. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. *In Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology (HLT-NAACL 2003)*, (pp. 71–78). Edmonton, Canada.

[7] Mihalcea, R., & Ceylan, H. (2007). Explorations in automatic book summarization. *In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL 2007)*, (pp. 380–389). Prague, Czech Republic.

[8] Navigli, R., & Lapata, M. (2010). An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *IEEE Computer Society , 32*.

[9] Porter, M. (1980). An algorithm for suffix stripping. *Program , 14*, 130–137.

[10] Radev, D., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. *omputational Linguistics , 22*, 399–408.

[11] Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing and Management , 33*, 193–207.

[12] Shen, D., Sun, J. -T., Li, H., Yang, Q., & Chen, Z. (2007). Document summarization using onditional random fields. *In Proceedings of the 20th international joint conference on artificial intelligence (JCAI 2007)*, (pp. 2862–2867). Hyderabad, India.

[13] Svore, K. M., Vanderwende, L., & Burges, C. J. C. Enhancing single-document summarization by combining RankNet and third-party sources. *In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL 2007)*, (pp. 448–457). Prague, Czech Republic.

[14] Wan, X. (2008). Using only cross-document relationships for both generic and topic-focused multi-document summarizations. *Information Retrieval , 11*, 25–49.

[15] Wan, X., Yang, J., & Xiao, J. (2007). Manifold-ranking based topic-focused multidocument summarization. *In Proceedings of the 20th international joint conference on artificial intelligence (IJCAI 2007)*, (pp. 2903–2908). Hyderabad, India.

[16] Yeh, J-Y., Ke, H-R., Yang, W-P., & Meng, I-H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing and Management , 41*, 75–95.

[17] McDonald, D. M., & Chen, H. (2006). Summary in context: Searching versus browsing. ACM Transactions on Information Systems, 24, 111–141.

[18] Fung, P., & Ngai, G. (2006). One story, one flow: Hidden Markov story models for multilingual multi document summarization. ACM Transaction on Speech and Language Processing, 3, 1–16.

[19] Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The Google similarity measure. IEEE Transaction on Knowledge and Data Engineering, 19, 370–383.

[20] Alguliev, Rasim; Aliguliyev, Ramiz. "Evolutionary algorithm for extractive text summarization". http://www.highbeam.com/doc/1G1-214205320.html.

[21] Stergos Afantenos, Vangelis Karkaletsis, Panagiotis Stamatopoulos. "Summarization from medical documents: a survey".

[22] Xiaojun Wan. "Towards a Unified Approach to Simultaneous Single-Document and Multi-document Summarizations".

Automated Periodontal Diseases Classification System

Aliaa A. A. Youssif
Department of Computer Science,
Faculty of Computers and Information
Helwan University
Cairo, Egypt

Abeer Saad Gawish
Department of Oral Medicine and Periodontology, Faculty of Dental Medicine (Girls' Branch)
Al-Azhar University
Cairo, Egypt

Mohammed Elsaid Moussa
Department of Computer Science,
Faculty of Computers and Information
Helwan University
Cairo, Egypt

***Abstract*— This paper presents an efficient and innovative system for automated classification of periodontal diseases, The strength of our technique lies in the fact that it incorporates knowledge from the patients' clinical data, along with the features automatically extracted from the Haematoxylin and Eosin (H&E) stained microscopic images. Our system uses image processing techniques based on color deconvolution, morphological operations, and watershed transforms for epithelium & connective tissue segmentation, nuclear segmentation, and extraction of the microscopic immunohistochemical features for the nuclei, dilated blood vessels & collagen fibers. Also, Feedforward Backpropagation Artificial Neural Networks are used for the classification process. We report 100% classification accuracy in correctly identifying the different periodontal diseases observed in our 30 samples dataset.**

Keywords-Biomedical image processing; epithelium segmentation; feature extraction; nuclear segmentation; periodontal diseases classification.

I. INTRODUCTION

Periodontitis is a chronic inflammatory disease of vascularized supporting tissues of the teeth [1], Periodontal disease occurs when inflammation or infection affect the gingiva and extend to the periodontal apparatus [2], The 1999 classification system for periodontal diseases and conditions listed seven major categories of periodontal diseases [3,4]: Gingivitis, Chronic periodontitis, Aggressive periodontitis, Periodontitis as a manifestation of systemic disease, Necrotizing ulcerative gingivitis/periodontitis, Abscesses of the periodontium, Combined periodontic-endodontic lesions; The latter 4 are associated with systemic diseases [5,6,7,8,9]; Hence, this work preliminary focuses on identifying the different types of periodontal diseases by using computer-assisted microscopy system for automated classification of periodontal diseases to increase the accuracy and reduce the workload in classifying and diagnosis of the different categories of periodontal diseases, which help in designing the treatment plan used.

The paper is organized as follows: Section II explains the materials and methods that we followed to get our dataset, Section III describes the implementation details of our proposed system, Section IV shows experimental results and discussions about these results, and Section V concludes the paper and introduces the future work that can be done.

II. MATERIALS AND METHODS

A. Study Cases

The study was conducted on 32 patients attending between February 2009 and March 2011 to the Oral Medicine, Periodontology, Oral Diagnosis and Radiology Department, Faculty of Dental Medicine Girls' Al-Azhar University.

Patients were suffering from gingival inflammation, which may extend to include the periodontium (different types of periodontitis) or gingival overgrowth (due to different etiological factors).

Patients were excluded from the present study if they were smokers, pregnant or post-menopausal women. All selected patients had not undergone any periodontal therapy for at least six months.

B. Collected Clinical Data

The following clinical parameters were collected and recorded on six sites at each tooth; all linear measurements were recorded to the nearest 0.5 mm using William graduated periodontal probe: Plaque index (PI) [10] Pocket Depth (PD) and Clinical attachment level (CAL) was measured from the CEJ to the apical part of the sulcus, All included patients were indicated for surgical flaps as a line of their treatment.

C. Histopathological Sample Preparation

After the surgical procedures; the excised tissue samples were immersed in 10% formalin and decalcified in multiple baths of 10% trichloroacetic acid. The blocks were immersed in paraffin, and semi-serial 4 µm histologic sections were stained with Haematoxylin and Eosin (H&E).

D. Image Capture

Representative sections were photographed using a Leitz DMRD Microscope (Leica, Wetzlar, Germany) with 20 X objective UPLanFl (resolution 0.67 µm) at a size of 1600 X 1200 pixels (interpixel distance 0.62 µm) using a JVC KY-55B 3-CCD colour camera attached to a 24 bit RGB frame grabber (Imaging Technologies IT4PCI, Bedford,

MA, U.S.A.) controlled by Optimas v. 6.51 (Media Cybernetics, Silver Spring, MD, U.S.A.)

E. Periodontal Diseases Groups

The 32 patients were classified into 4 diagnostic groups according to their periodontal status; 16 patients were classified into Gingival Enlargement group, 7 patients into Chronic Gingivitis group, another 7 patients into Chronic Periodontitis group, and the last 2 patients into Aggressive Periodontitis group.

We excluded the patients which were classified into the Aggressive Periodontitis group from our classification experiments; because of the very small number of study cases that we have for this group; which was not sufficient for our neural networks based classification system to work with. Although we included these 2 cases in our preprocessing and feature extraction experiments, to make sure that our proposed system is generic enough to handle all of the different types of diseases.

We also included another study case for a healthy person to our preprocessing and feature extraction experiments.

III. PROPOSED SYSTEM

Our proposed system can be divided into several well-defined stages as illustrated in Figure 1. We first get the H&E stained slide, do some preprocessing steps to it to remove background and undesired objects, segment it to the tissue's main components epithelium and connective tissue, then extract features from both parts, and then use these features along with the clinical data of the patient to start the classification process.

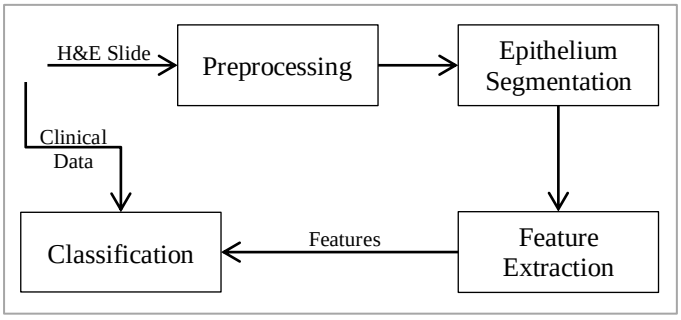


Figure 1 Schematic structure of the proposed system

A. Pre-processing

Mainly images consist of 2 parts: Tissue (Epithelium + Connective Tissue) and Background. In this preprocessing phase, we need to remove the background areas from the image, allowing the further processing to be done only on the tissue part.

We implemented 2 different algorithms to achieve this task; the first one is fully-automated, while the second one is semi-automated because it requires some clicks from the user to specify the seed points in the background area. The pseudo codes for fully-automated and semi-automated algorithms are shown below in Pseudocode I & Pseudocode II respectively. 0shows one sample result of the 2 algorithms. Also, sensitivity and specificity results of these algorithms are displayed in the Results and Discussion section.

Pseudocode I — FULLY-AUTOMATED PREPROCESSING ALGORITHM

1. Get the "a" channel from L*a*b* color space of the original image, $Image_1$.
2. Smoothing of $Image_1$ with a $5 \cdot 5$ average filter to preserve only large detail [11], $Image_2$.
3. Convert $Image_2$ to black & white with 0.2 threshold value, and then invert it, $Mask_1$.
4. Morphologically open $Mask_1$ to remove all black objects that have fewer than 1000 pixels, $Mask_2$.
5. Morphologically open $Mask_2$ to remove all white objects that have fewer than 1000 pixels, $Mask_3$.
6. Perform a morphological dilatation with a $10 \cdot 10$ window size on the $Mask_3$, $Mask_4$.
7. Fill Holes of $Mask_4$, $Mask_5$.
8. Perform a morphological erosion with a $10 \cdot 10$ window size on the $Mask_5$, $Mask_6$.
9. Remove areas from $Mask_6$ that are not on border [12], $Mask_7$.
10. Morphologically open $Mask_7$ to remove all objects that have more than 10000 pixels, $Mask_8$.
11. Apply final smoothing for $Mask_8$, $Mask_9$.

Pseudocode II — SEMI-AUTOMATED PREPROCESSING ALGORITHM

1. Get the "Green" channel of the original image, $Image_1$.
2. Get Seed Point from the pathologist, inside the background region, then set the initial threshold to the intensity value of this seed point.
3. Based on the initial threshold, get all the pixels that its intensity in a range of 50 values, $Mask_1$, i.e.
   abs(newPixelValue - initalThreshold) <= 50
4. Perform image reconstruction on the $Mask_1$ to get only the pixels that are connected to the seed point, $Mask_2$.
5. Perform a morphological erosion with a $1 \cdot 1$ window size on the $Mask_2$, $Mask_3$.
6. Morphologically open $Mask_3$ to remove all white objects that have fewer than 1000 pixels, $Mask_4$.
7. Perform a morphological dilation with a $2 \cdot 2$ window size on the $Mask_4$, $Mask_5$.
8. Morphologically open $Mask_5$ to remove all black objects that have fewer than 1000 pixels, $Mask_6$.
9. Perform a morphological erosion with a $1 \cdot 1$ window size on the $Mask_6$, $Mask_7$.
10. Repeat steps 2 to 10 for each separate background region.
11. Edge shifting for the $Mask_7$ by performing morphological erosion with a disk-shaped structuring element [13], [14] with radius 2 the $Mask_7$, $Mask_8$.

For very small background regions, we skipped steps from 6 to 10 for these regions only. This happened only 3 times over our 33 images.

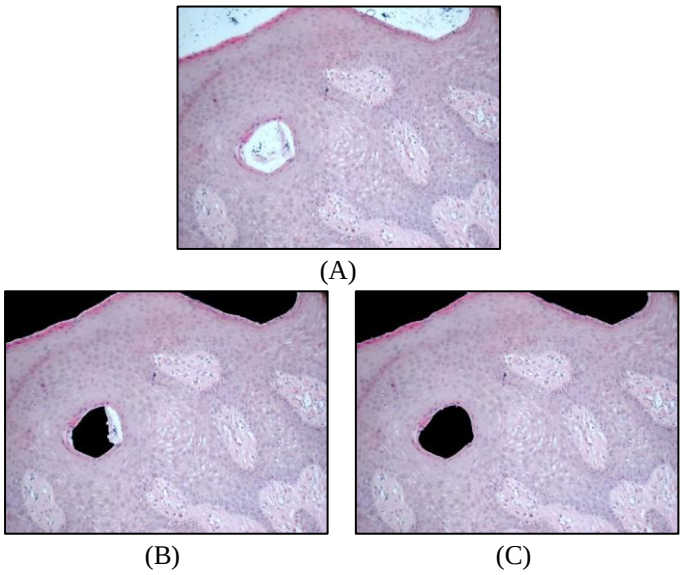


Figure 2 Pre-processing sample (A) The original image. (B) After applying fully-automated algorithm. (C) After applying semi-automated algorithm.

B. Epithelium Segmentation

In this phase of our system, we need to segment the tissue into Epithelium and Connective Tissue; Epithelium segmentation was targeted in some studies [15,16,17]. As stated in [15] there are a lot of challenges doing this task, and they are mainly caused by staining artifacts, lighting acquisition conditions, and undesired touching objects. Some studies [15] used the saturation channel in the HSV color space, others [16] used multiple gray level automated thresholding over the image's green channel, while others [17] used region growing algorithm [18] over the gray channel of the image.

We found that 73% of our samples give better results when using the HSV Saturation channel, while the remaining 27% give better results when using the RGB green channel. We implemented an automated way for this segmentation process; Pseudocode III describes our epithelium segmentation algorithm.

We also implemented a post-processing step, to move some parts from epithelium to connective tissue or vice versa manually, because our proposed algorithm didn't achieve highly accurate results for all study cases, because the brightness value of the infiltrate was similar to that of the epithelium in images with high inflammatory infiltrate [16]. Information about sensitivity and specificity of our algorithm is found in the Results and Discussion section. Figure 3 shows one sample result of this algorithm.

C. Feature Extraction

We extracted a set of features from our Haematoxylin and Eosin (H&E) stained slides, that we thought it will be helpful in the classification process later. Below are the extracted features along with the algorithms we used to extract it:

Pseudocode III EPITHELIUM SEGMENTATION ALGORITHM

1. Get the "Saturation" channel from the HSV color space of the tissue image [a], $Image_1$, which retrieved from the preprocessing phase, $Channel_1$.
2. Adjust [b] intensity values in the grayscale $Channel_1$, to map it to new values in such that 1% of data is saturated at low and high intensities of the $Channel_1$. This increases the contrast of the output channel, $Channel_2$. [15].
3. Apply $10 \cdot 10$ median filter [c] for removing the noise in the $Channel_2$ without harming edges, $Channel_3$. [19]
4. Convert $Channel_3$ to black & white by applying global thresholding through Otsu's method [20], $Mask_1$. [d]
5. Fill Holes of $Mask_1$, $Mask_2$.
6. Perform a morphological erosion with a $10 \cdot 10$ window size on the $Mask_2$, $Mask_3$.
7. Morphologically open $Mask_3$ to remove all white objects that have fewer than 1000 pixels, $Mask_4$.
8. Perform a morphological dilation with a $20 \cdot 20$ window size on the $Mask_4$, $Mask_5$.
9. Morphologically open $Mask_5$ to remove all black objects that have fewer than 1000 pixels, $Mask_6$.
10. Perform a morphological erosion with a $10 \cdot 10$ window size on the $Mask_6$, $Mask_7$.
11. Edge shifting for the mask by performing a morphological erosion with a strel $2 \cdot 2$ window size on the $Mask_7$, $Mask_8$. [13], [14].

[a] For 9 samples of our 33 samples, we used RGB Green channel instead of HSV Saturation channel, because we found it giving better results.

[b] For 1 sample, we replaced image adjustment step with histogram equalization.

[c] For 6 samples, we used a $20 \cdot 20$ window size for median filtering instead of the $10 \cdot 10$ window size, and for another 1 sample we used a $9 \cdot 9$ window size.

[d] For 12 samples, we applied histogram equalization after step 4.

1) Epithelium Percentage

After segmentation of the tissue image into its components (Epithelium and Connective Tissue) in the previous phase, we calculated the percentage of the Epithelium part.

(Note that the background regions are already excluded)

2) Nuclei in Epithelium

We segmented the nuclei that found in the epithelium part, and used its count and percentage from the whole epithelium in our features list, Nuclei segmentation was targeted in some studies such as [11], [21], [22] which were using nuclear localization achieved using the colour deconvolution algorithm developed by [23] to obtain the optical density of the Haematoxylin stain alone plus spatial partition of the epithelial compartment representing the exclusive area of influence of each nucleus profile, [24] which is based on the multiscale Laplacian-of-Gaussian (LoG) filter, [25] which depends on ellipse fitting and the watershed transform, [26] which relies on some morphology-based techniques.

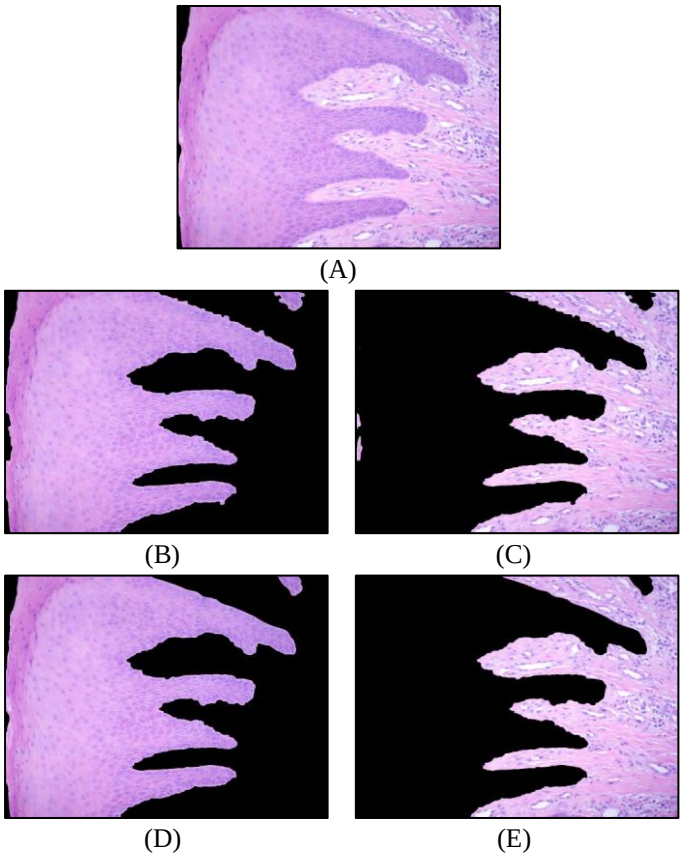


Figure 3 Epithelium Segmentation (A) Original tissue image after pre-processing phase. (B) Automatically segmented epithelium. (C) Automatically segmented connective tissue. (D) Epithelium after post-processing. (E) Connective-Tissue after post-processing.

In our proposed system, we used steps in Pseudocode IV to segment the nuclei in the epithelium part of the image; Figure 4 shows two sample results of this algorithm.

3) Nuclei in Connective Tissue

We segmented the nuclei that found in the connective tissue component, and used its count and percentage from the whole connective tissue in our features list.

We do this by getting 2 masks; the first one represents the light areas in the connective tissue (blood vessels) while the second one represents the dark areas (nuclei). Then we need to filter the nuclei mask to exclude the ones that are related to the dilated blood vessels. Our steps for doing this segmentation part are described in Pseudocode V. Figure 5 shows one sample result of this algorithm.

4) Dilated Blood Vessels

We segmented the dilated blood vessels that found in the connective tissue part, and used its count and percentage from the whole connective tissue in our features list.

We do this with almost the same steps described in the Pseudocode V that were used in the connective tissue nuclei segmentation. We get the first mask with exactly the same steps, while excluding steps e, f and j from the ones required to get the second mask. And to get the final mask, we filter parts in $Mask_1$ (dilated blood vessels candidates) to only the ones that intersect at least 3 times with parts of $Mask_2$ (nuclei); finally, we count and get the percentage of all objects in $Mask_3$. 0Figure 6 shows one sample result of this algorithm.

Pseudocode IV EPITHELIUM NUCLEI SEGMENTATION ALGORITHM

1. Get the optical density of the Haematoxylin as an 8 bit grayscale image, by colour deconvolution [23] of the epithelium part, retrieved from the epithelium segmentation phase, $Channel_1$. [a]
2. Normalize intensity values in the grayscale $Channel_1$, to change the range of pixel intensity values to be between 0 and 255, $Channel_2$. [b]
3. Convert $Channel_2$ to black & white by computing the extended-minima transform, which is the regional minima of the H-minima transform [27], with a value of 30, $Mask_1$.
4. Fill Holes of $Mask_1$, $Mask_2$.
5. Remove areas from $Mask_2$ that are not on border, $Mask_3$.
6. Morphologically open $Mask_3$ to remove all objects that have less than 15 pixels, $Mask_4$.
7. Morphologically open $Mask_4$ to remove all objects that have more than 800 pixels, $Mask_5$.
8. Convert each connected object to the ellipse that fit it, $Mask_6$.
9. Finally we count and get the percentage of all objects in $Mask_6$.

[a] For 9 samples of our 33 samples, we used RGB Red channel instead of Haematoxylin channel, because we found it giving better results.

[b] For 3 samples, we also apply image adjustment after step 2.

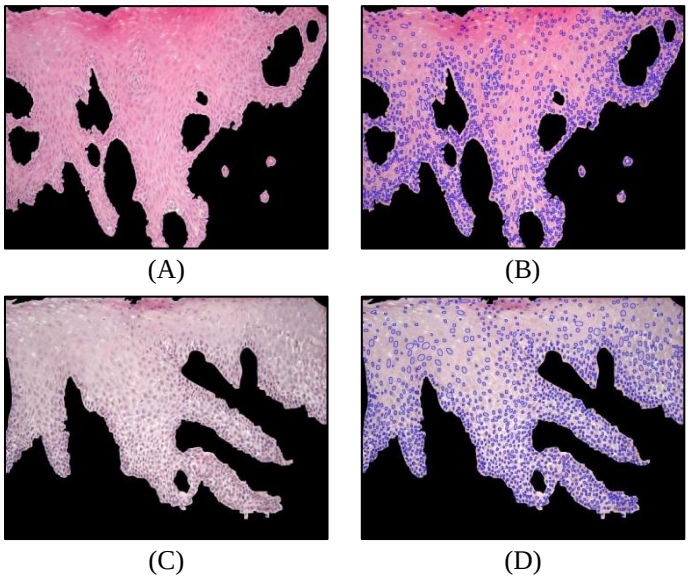


Figure 4 Nuclei in Epithelium. (A), (C) Two samples of epithelium images. (B), (D) After applying epithelium nuclei segmentation algorithm over (A) and (C) respectively. [Nuclei are surrounded by blue ellipses].

Pseudocode V — CONNECTIVE-TISSUE NUCLEI SEGMENTATION ALGORITHM

1. Get $Mask_1$
   a. Get the "Green" channel of the connective tissue part, retrieved from the epithelium segmentation phase, $Channel_1$.
   b. Enhance the contrast of the grayscale channel $Channel_{1\_1}$ by transforming the values using contrast-limited adaptive histogram equalization (CLAHE) [28], $Channel_{1\_1}$.
   c. Apply 3 • 3 median filter for removing the noise [19] in the $Channel_{1\_1}$ without harming edges, $Channel_{1\_2}$.
   d. Convert $Channel_{1\_2}$ to black & white with 0.8 threshold value, $Mask_{1\_1}$.
   e. Morphologically open $Mask_{1\_1}$ to remove all objects that have less than 80 pixels, $Mask_1$.
2. Get $Mask_2$
   a. Get the optical density of the Haematoxylin as an 8 bit grayscale image, by colour deconvolution [23] of the connective tissue part, retrieved from the epithelium segmentation phase, $Channel_2$.
   b. Adjust intensity values in the grayscale $Channel_2$, to map it to new values in such that 1% of data is saturated at low and high intensities of the $Channel_2$. This increases the contrast of the output channel, $Channel_{2\_1}$. [15].
   c. Apply 2-D adaptive filtering with a 5 · 5 window size over $Channel_{2\_1}$, $Channel_{2\_2}$.
   d. Convert $Channel_{2\_2}$ to black & white by computing the extended-minima transform, which is the regional minima of the H-minima transform [27], with a value of 30, $Mask_{2\_1}$.
   e. Morphologically open $Mask_{2\_1}$ to remove all objects that have less than 15 pixels, $Mask_{2\_2}$.
   f. Morphologically open $Mask_{2\_2}$ to remove all objects that have more than 800 pixels, $Mask_{2\_3}$.
   g. Apply watershed algorithm for $Mask_{2\_3}$, $Mask_{2\_w}$.
   h. Perform a morphological dilation with a 10 • 10 window size on the $Mask_{2\_3}$, $Mask_{2\_4}$.
   i. Apply AND operator between $Mask_{2\_4}$ and $Mask_{2\_w}$, $Mask_{2\_5}$.
   j. Remove areas from $Mask_{2\_5}$ that are not on border, $Mask_2$.
3. Get Final Mask
   a. Filter parts in $Mask_2$ (Nuclei) to only ones that not intersect with parts of $Mask_1$ (Light Parts) that intersects with at least 3 times with parts of $Mask_2$, $Mask_{3\_1}$.
   b. Apply image reconstruction algorithm between $Mask_{3\_1}$ and $Mask_{2\_3}$, $Mask_{3\_2}$.
   c. Convert each connected object in $Mask_{3\_2}$ to the ellipse that fit it, $Mask_3$.
4. Finally we count and get the percentage of all objects in $Mask_3$.

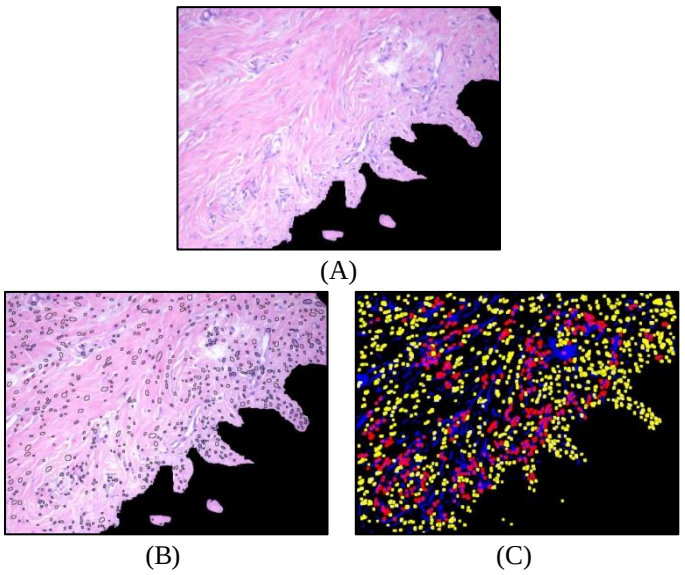


Figure 5 Nuclei in Connective Tissue. (A) Sample of connective tissue image. (B) Nuclei in connective tissue surrounded by black ellipses. (C) $Mask_1$is in blue, $Mask_2$ is in red and yellow, and Final Mask is in yellow.

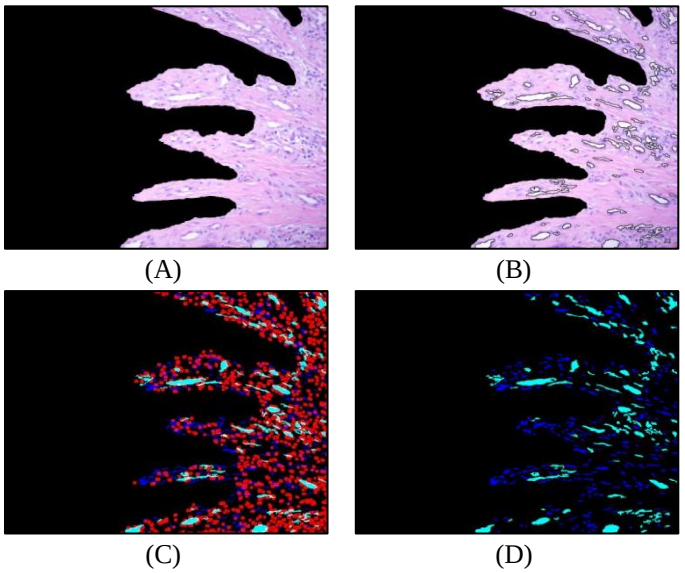


Figure 6 Dilated Blood Vessels. (A) Sample of connective tissue image. (B) Dilated Blood Vessels in connective tissue surrounded by black ellipses. (C) Candidate blood vessels are in blue and light blue. Red areas are nuclei. (D) Candidate blood vessels are in blue and light blue. Final Mask (passed blood vessels) is in light blue.

5) Collagen Fiber

We segmented the collagen fiber that found in the connective tissue part, and used its percentage from the whole connective tissue in our features list.

We do this by getting 2 masks, the first one represents the light areas in the connective tissue, and the second one

represents the dark areas, then we need to filter the nuclei mask to exclude the ones that are related to the dilated blood vessels. Our steps for doing this segmentation part are described in Pseudocode VI. 0Figure 7 shows one sample result of this algorithm.

Pseudocode VI COLLAGEN FIBER SEGMENTATION ALGORITHM

1. Get $Mask_1$
   a. Get the "Green" channel of the connective tissue part, retrieved from the epithelium segmentation phase, $Channel_1$.
   b. Enhance the contrast of the grayscale channel $Channel_{1\_1}$ by transforming the values using contrast-limited adaptive histogram equalization (CLAHE) [28], $Channel_{1\_1}$.
   c. Apply 3 • 3 median filter for removing the noise [19] in the $Channel_{1\_1}$ without harming edges, $Channel_{1\_2}$.
   d. Convert $Channel_{1\_2}$ to black & white with 0.75 threshold value, $Mask_1$.
2. Get $Mask_2$
   a. Get the optical density of the Haematoxylin as an 8 bit grayscale image, by colour deconvolution [23] of the connective tissue part, retrieved from the epithelium segmentation phase, $Channel_2$.
   b. Adjust intensity values in the grayscale $Channel_2$, to map it to new values in such that 1% of data is saturated at low and high intensities of the $Channel_2$. This increases the contrast of the output channel, $Channel_{2\_1}$. [15].
   c. Apply 2-D adaptive filtering with a 5 · 5 window size over $Channel_{2\_1}$, $Channel_{2\_2}$.
   d. Convert $Channel_{2\_2}$ to black & white by applying global thresholding algorithm using the value got from Otsu's method [20] multiplied by 1.1, and then invert it, $Mask_2$.
3. Get Final Mask by excluding $Mask_1$ and $Mask_2$ from the original connective tissue area, $Mask_3$.
4. Finally we get the percentage of all objects in $Mask_3$.

D. Classification

After extracting all the features, we used Feedforward Backpropagation Artificial Neural Networks in the classification process.

We used 11 inputs; 3 clinical data (Plaque Index, Pocket Depth, and Attachment Level) and 8 extracted features (Epithelium Percentage, Epithelium Nuclei Count & Percentage, Connective Tissue Nuclei Count & Percentage, Dilated Blood Vessels Count & Percentage, and Collagen Fiber Percentage), 10 hidden neural networks, and 3 outputs representing our 3 diseases groups (Gingival Enlargement, Chronic Gingivitis, and Chronic Periodontitis); the used neural network model is shown in Figure 8.

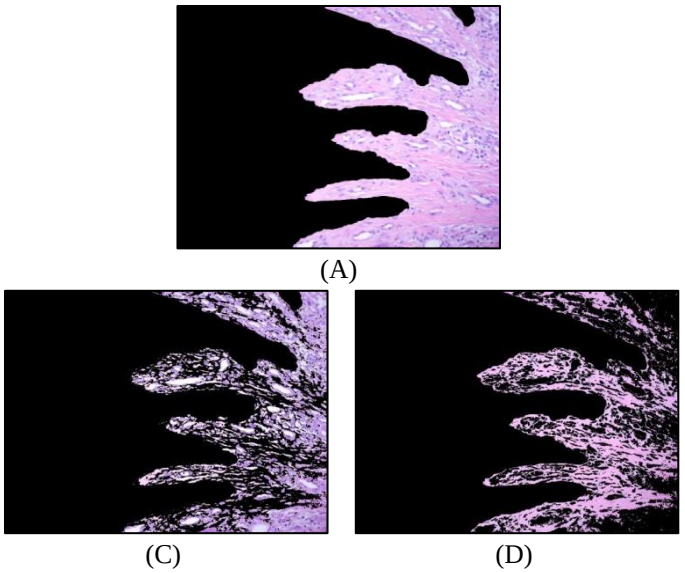


Figure 7 Collagen Fiber (A) Sample of connective tissue image. (B) Mask1 + Mask2. (C) Final Mask (Collagen Fiber).

We divided our samples randomly into 3 groups, 20 samples for the training set which were presented to the neural network during training, and the network was adjusted according to its error, 5 samples for the validation set which were used to measure network generalization, and to halt training when generalization stops improving, and the last 5 samples for the testing set which had no effect on training so providing an independent measure of network performance during and after training.

We wrote some algorithm to make sure that each data set contains a relative number of study cases for each disease group. We used scaled conjugate gradient backpropagation function for network training. Also, we used mean squared error as our performance function.

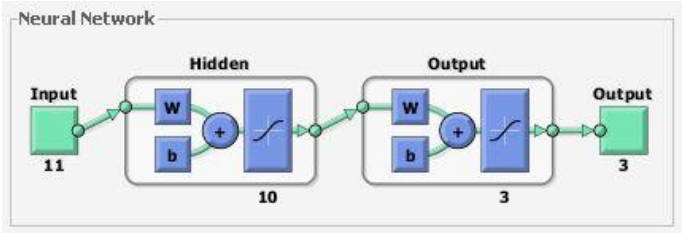


Figure 8 Neural Network Model

IV. RESULTS AND DISCUSSION

For background removal preprocessing phase, we applied both fully-automated and semi-automated preprocessing algorithms described in Pseudocode I and Pseudocode II respectively. We considered the semi-automated results as the ground truth, and compared the results of the fully-automated algorithm to it; over our 33 samples, we achieved a sensitivity of 68.42% and specificity of 98.56%. 6 samples of our samples actually contain no background at all, and when removing their results from our statistics, we achieved a sensitivity of 83.33% and specificity of 98.66% over our remaining 27 samples.

For epithelium segmentation phase, we applied the algorithm described in Pseudocode III over our date, and since the results of this step has to be accurate since it is used in all later phases, we also implemented a manual way for moving parts from the epithelium part to the connective tissue part or vice versa. We considered the results after manual processing the ground truth, and found that our algorithm achieved a sensitivity of 84.99% and specificity of 88.40% over our 33 samples dataset, below are some statistics about most notable results:

- 16 Samples (48.5%) give Sensitivity >= 95%, 15 Samples (45.5%) give Specificity >= 95%, and 11 Samples (33.3%) give Sensitivity & Specificity >= 95%.
- 16 Samples (48.5%) give Sensitivity >= 90%, 20 Samples (60.6%) give Specificity >= 90%, and 13 Samples (39.4%) give Sensitivity & Specificity >= 90%.
- 5 Samples (15.2%) give Sensitivity <= 60%, 2 Samples (6.1%) give Specificity <= 60%, and 2 Samples (6.1%) give Sensitivity & Specificity <= 60%.

For nuclei in epithelium segmentation phase, we applied the algorithm described in Pseudocode IV over our data; we found that optical density of the Haematoxylin stain retrieved through colour deconvolution [23] is the best channel for nuclei segmentation in 24 samples (73%) while the RGB Red channel is the best for the remaining 9 samples (27%). Information regarding average (count, area, and percentage) of nuclei found in the epithelium for each disease group can be found in Table I.

TABLE I. EPITHELIUM NUCLEI SEGMENTATION STATISTICS.

Group	No. of Samples	Avg. Count	Avg. Area	Avg. Percentage
Normal	1	826	117887	10.31
Gingival Enlargement	16	916	108013	09.46
Chronic Gingivitis	7	1359	131323	12.41
Chronic Periodontitis	7	1020	125925	11.96
Aggressive Periodontitis	2	1284	125068	10.44

For nuclei in connective tissue segmentation phase, we applied the algorithm described in Pseudocode IV over our data; Information regarding average (count, area, and percentage) of nuclei found in connective tissue for each disease group can be found in Table II.

TABLE II. CONNECTIVE TISSUE NUCLEI SEGMENTATION STATISTICS.

Group	No. of Samples	Avg. Count	Avg. Area	Avg. Percentage
Normal	1	413	27722	03.65
Gingival Enlargement	16	409	22835	03.07
Chronic Gingivitis	7	438	26042	03.06
Chronic Periodontitis	7	510	30459	03.45
Aggressive Periodontitis	2	444	19670	02.92

For dilated blood vessels segmentation phase, we found the following statistics that are described in Table III.

TABLE III. DILATED BLOOD VESSELS SEGMENTATION STATISTICS.

Group	No. of Samples	Avg. Count	Avg. Area	Avg. Percentage
Normal	1	92	64005	08.29
Gingival Enlargement	16	94	75001	10.57
Chronic Gingivitis	7	77	68666	08.90
Chronic Periodontitis	7	82	65443	08.53
Aggressive Periodontitis	2	117	84809	12.61

For collagen fiber segmentation phase, we applied the algorithm described in Pseudocode VI over our data; Information regarding average (area, and percentage) of collagen fibers for each disease group can be found in Table IV.

TABLE IV. COLLAGEN FIBER SEGMENTATION STATISTICS.

Group	No. of Samples	Avg. Area	Avg. Percentage
Normal	1	347325	45.72
Gingival Enlargement	16	423448	58.54
Chronic Gingivitis	7	383753	51.01
Chronic Periodontitis	7	404528	51.03
Aggressive Periodontitis	2	373231	56.24

For our classification results; Figure 9 shows our neural network performance, Figure 10 0shows the confusion matrix for the training dataset, validation dataset, test dataset, & the whole dataset, while 0Figure 11 shows Receiver Operating Characteristic ROC for them.

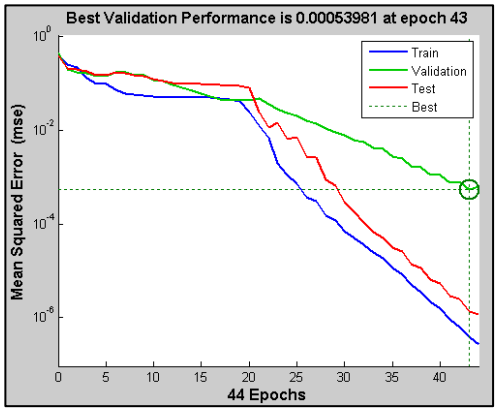


Figure 9 Neural Network Training Performance

I. CONCLUSION AND FUTURE SCOPE

An automated system has been developed for classification of periodontal diseases using H&E stained microscopic images of the tissues around the affected teeth along with clinical data. The epithelium percentage of the whole tissue, count & percentage of nuclei in epithelium & connective tissue, dilated blood vessels count & percentage, and collagen fiber percentage are used as features during the classification process which is done using Feedforward Backpropagation Artificial Neural Networks. It was found that using these mixed features together achieve a more accurate classification results than using only clinical data or H&E stained images' extracted features.

Figure 10 Neural Network Training Confusion Matrix

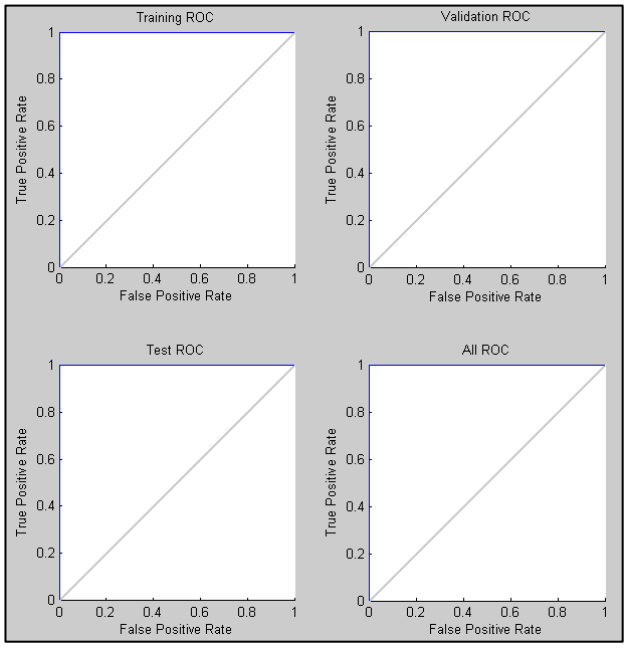


Figure 11 Neural Network Training Receiver Operating Characteristic ROC

We suggest the following as a future work for this study:

1) Get a larger dataset of diseases, to provide better training and testing of the proposed system, also get test cases in the Aggressive Periodontitis group to include them in the training and testing phases of the program to make it as generic as possible for all periodontal diseases.
2) Enhance the fully-automated background removal algorithm found in Pseudocode I, to achieve similar results to the semi-automated one found in Pseudocode II. Also, enhance the epithelium segmentation algorithm found in Pseudocode III, to achieve higher Specificity and sensitivity results.
3) Try to find more generic algorithms' parameters, to decrease the number of study cases that we had to alter its parameters to get better results; as shown in the footnotes of Pseudocode III and Pseudocode IV.

REFERENCES

[1] D. V, Prapulla, P. B. Sujatha, and A. R. Pradeep, "Gingival Crevicular Fluid VEGF Levels in Periodontal Health and Disease," *Journal of Periodontology*, vol. 78, no. 9, pp. 1783-1787, September 2007.

[2] C. B. Wiebe and E. E. Putnins, "The periodontal disease classification system of the American Academy of Periodontology--an update," *J Can Dent Assoc.*, vol. 66, no. 11, pp. 594-597, December 2000.

[3] G. C. Armitage, "Development of a classification system for periodontal diseases and conditions," *Annals of Periodontology*, vol. 4, pp. 1-6, December 1999.

[4] (2012, January) Wikipedia, the free encyclopedia. [Online]. http://en.wikipedia.org/wiki/Periodontitis

[5] L. J. Brown and H. Löe, "Prevalence, extent, severity and progression of periodontal disease," *Periodontol 2000*, vol. 2, pp. 57-71, June 1993.

[6] G. C. Armitage, "Periodontal diseases: diagnosis," *Annals of Periodontology*, vol. 1, no. 1, pp. 37-215, 1996.

[7] P. N. Papapanou, "Periodontal Diseases: Epidemiology," *Annals of Periodontology*, vol. 1, no. 1, pp. 1-36, November 1996.

[8] Löe H., Anerud A., Boysen H., and Morrison E., "Natural history of periodontal disease in man. Rapid, moderate and no loss of attachment in Sri Lankan laborers 14 to 46 years of age," *J Clin Periodontol*, vol. 13, no. 5, pp. 431-445, May 1986.

[9] R. Attström and U. van der Velden, "Consensus report (epidemiology)," in *Proceedings of the 1st European Workshop on Periodontics, 1993*, London, 1994, pp. 120-126.

[10] H. LOE and J. SILNESS, "Periodontal disease in pregnancy. I. Prevalence and severity," *Acta Odontol Scadinavia*, vol. 21, pp. 533-551, December 1963.

[11] G. Landini and I. E. Othman, "Estimation of tissue layer level by sequential morphological," *Journal of Microscopy*, vol. 209, pt 2, pp. 118– 125, February 2003.

[12] P. Soille, *Morphological Image Analysis: Principles and Applications*, 1st ed.: Springer-Verlag, 1999, pp. 164-165.

[13] R. Adams, "Radial Decomposition of Discs and Spheres," *Computer Vision, Graphics, and Image Processing: Graphical Models and Image Processing*, vol. 55, no. 5, pp. 325–332, September 1993.

[14] R. Jones and P. Soille, "Periodic lines: Definition, cascades, and application to granulometrie," *Pattern*

Recognition Letters, vol. 17, pp. 1057–1063, 1996.

[15] M.M. Sami, M. Saito, H. Kikuchi, and T. Saku, "A computer-aided distinction of borderline grades of oral cancer," in *16th IEEE International Conference on Image Processing (ICIP)*, Cairo, 2009, pp. 4205 - 4208.

[16] R. Abu Eida and G. Landini, "Quantification of the Global and Local Complexity of the Epithelial-Connective Tissue Interface of Normal, Dysplastic, and Neoplastic Oral Mucosae Using Digital Imaging," *Pathology - Research and Practice*, vol. 199, no. 7, pp. 475-482, 2003.

[17] M. Pal et al., "Quantitative dimensions of histopathological attributes and status of GSTM1–GSTT1 in oral submucous fibrosis," *Tissue and Cell*, vol. 40, no. 6, pp. 425-435, December 2008.

[18] J. Maeda, S. Novianto, S. Saga, Y. Suzuki, and V. V. Anh, "Rough and accurate segmentation of natural images using fuzzy region-growing algorithm," in *IEEE Proceedings Image Processing. ICIP 99*, vol. 3, Kobe, 1999, pp. 227-231.

[19] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed.: Prentice Hall, 2008.

[20] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62-66, 1979.

[21] G. Landini and I. E. Othman, "Quantification of Local Architecture Changes Associated with Neoplastic Progression in Oral Epithelium using Graph Theory," in *Fractals in Biology and Medicine*.: Birkhäuser Basel, 2005, vol. 4, pp. 193-201.

[22] G. Landini, "Quantitative analysis of the epithelial lining architecture in radicular cysts and odontogenic keratocysts," *Head & Face Medicine*, vol. 2, 2006.

[23] A. C. Ruifrok and D. A. Johnston, "Quantification of histological staining by color deconvolution," *Anal Quant Cytol Histol*, vol. 23, pp. 291-299, 2001.

[24] Y. Al-Kofahi, W. Lassoued, and W. Lee, "Improved Automatic Detection and Segmentation of Cell Nuclei in Histopathology Images," *IEEE transactions on biomedical engineering*, vol. 57, no. 4, pp. 841-852, April 2010.

[25] M. Park et al., "Automatic cell segmentation in microscopic color images using ellipse fitting and watershed," in *The 201 0 IEEEIICME International Conference on Complex Medical Engineering*, Gold Coast, Australia, 2010, pp. 69-74.

[26] S. DiCataldo, E. Ficarra, A. Acquaviva, and E. Macii, "Achieving the way for automated segmentation of nuclei in cancer tissue images through morphology-based approach - A quantitative evaluation," *Computerized Medical Imaging and Graphics*, vol. 34, pp. 453–461, 2010.

[27] P. Soille, *Morphological Image Analysis: Principles and Applications*, 1st ed.: Springer-Verlag, 1999, pp. 170-171.

[28] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in *Graphic Gems IV*. San Diego: Academic Press Professional, 1994, pp. 474–485.

AUTHORS PROFILE

Aliaa A. A. Youssif, professor of computer science, Faculty of Computers and Information, Helwan University. Cairo, Egypt. She received her B.Sc and MSc. degree in telecommunications and electronics engineering from Helwan University. Prof. A. Youssif received the PhD degree in computer science from Helwan University in 2000. She was a visiting professor at George Washington University (Washington DC, USA) in 2005. She was also a visiting professor at Cardiff University in UK (2008). She is currently vice dean for the faculty of Computers and Information, Helwan University. Her fields of interest include pattern recognition, AI researches, and medical imaging. She published more than 40 papers in different fields. She can be reached at aliaay@helwan.edu.eg

Abeer Saad Gawish, professor of Oral Medicine, Periodontology, Oral Diagnosis and Radiology, Faculty of Dental Medicine (Girls'), Al-Azhar University. Cairo, Egypt. She received her B.Sc degree in Oral and Dental Medicine from Alexandria University. Prof. A. Gawish received the MSc. And the PhD degree Oral and Dental Medicine from Cairo University in 2000. She was a visiting professor at Pan Arab Association of Oral &Maxillofacial Surgeon KSA 2010. She is currently vice dean for the faculty of Oral and Dental Medicine, Sinai University. Her fields of interest include periodontal medicine, genetics in periodontal medicine, implantology. She published more than 37 papers in different fields. She can be reached at abeergawish@yahoo.com

Mohammed Elsaid Moussa received his BSc. degree in computer science from Faculty of Computers and Information, Helwan University, Cairo, Egypt in 2005. He is currently working as a teacher assistant at the same faculty, and a master's degree student under the supervision of prof. Aliaa A. A.Youssif and prof. Abeer Saad Gawish. His areas of interests include image processing, medical imaging, cloud computing, and programming languages. He can be reached at engmohammedelsaid@hotmail.com

Communication and migration of an embeddable mobile agent platform supporting runtime code mobility

Mohamed BAHAJ
Department of Mathematics and Computer Science,
Université Hassan 1er, FSTS, LABO LITEN
Settat, Morocco

Khaoula ADDAKIRI
Department of Mathematics and Computer Science,
Université Hassan 1er, FSTS, LABO LITEN
Settat, Morocco

Noreddine GHERABI
Department of Mathematics and Computer Science,
Université Hassan 1er, FSTS, LABO LITEN
Settat, Morocco

***Abstract*—In this paper we present the design and the implementation of Mobile-C, an IEEE Foundation for Intelligent Physical Agents (FIPA) compliant agent platform for mobile C/C++ agents. Such compliance ensures the interoperability between a Mobile-C agent and other agents from heterogeneous FIPA compliant mobile agent platforms. Also, the Mobile-C library was designed to support synchronization in order to protect shared resources and provide a way of deterministically timing the execution of mobile agents and threads. The new contribution of this work is to combine the mechanisms of agent migration and their synchronization.**

Keywords- Mobile agent; Mobile agent platform;Agent communication.

I. INTRODUCTION

Mobile agent is a design program with a persistent identity which migrates in the network and communicates with its environment and other agent [1]. It has been applied to a variety of distributed applications, such as manufacturing [2-4], electronic commerce [5-7], network management [8, 9], health care [10], and entertainment [11]. During the execution, mobile agents can be dynamically created and sent to the destination systems to perform tasks with the up-to-date code. The mobility allows mobile agent to migrate from one host to another in the network and provides a several applications with flexibility and adaptability that are both able to satisfy the requirement and condition in a distributed environment.

The importance of Mobile agent technology in the design of distributed applications on the web has led the OMG (Object Management Group) to define the specifications MASIF (Mobile Agent System Interoperability Facility) for interoperability between different systems to mobile agents. Another effort is made by FIPA (Foundation for Intelligent Physical Agents) to specify the architecture and also the semantics of communication between mobile agents.

The majority of mobile agent platforms in use are Java-oriented. Multiple mobile agent platforms supporting Java mobile agent code include Mole [12], Aglets [13], Concordia [14], JADE [15], and Agents [16]. Using a standard language like the mobile agent code language that provides both high-level and low-level functionalities is a good choice to treat with the large number of distributed applications. The choice of C/C++ is a proper for a mobile agent code language because it's provides powerful functions in terms of memory access. Furthermore, C is a language which can easily interface with a variety of low-level hardware devices. Ara [17, 18] and TACOMA [19] are two mobile agent platforms supporting C mobile agent code, while Ara also supports C++. Mobile agent code is compiled as byte code [20] and machine code [21] for execution in both Ara and TACOMA, respectively.

Mobile-C [22-25] was originally developed as a stand-alone, IEEE FIPA compliant mobile agent platform to accommodate applications where low-level hardware is involved and embedded systems [26]. Most of the systems are written in C/C++; Mobile-C chose C/C++ as the mobile agent language because C has an advantage for easy interfacing with control programs and underlying hardware. Additionally, Mobile-C integrated an embeddable C/C++ interpreter, Ch [27-29], as the Agent Execution Engine (AEE) in order to run the mobile agent code. The migration of mobile agent in Mobile-C is achieved through FIPA agent communication language (ACL) messages. Using FIPA ACL messages for agent migration in FIPA compliant agent systems simplifies agent platform, reduces development effort and easily achieves inter-platform migration through well-designed communication mechanisms provided in the agent platform. Messages for agent communication and migration are presented in FIPA ACL and encoded in XML. Also, the Mobile-C library was designed to support synchronization [26] in order to protect shared resources and provide a way of deterministically timing the execution of mobile agents and threads.

In this paper we present the Mobile-C library that can embed Mobile-C into any C/C++ programs to facilitate the design of mobile agent-based applications, also the possibility

to combine the migration of the mobile agent over the network and the synchronization mechanism existing in Mobile-C. Mobile agents are an application that can control the agent platform, its modules and other mobile agents, as well as smoothly interface with a variety of low-level hardware devices. Using FIPA ACL messages for agent migration in FIPA compliant agent systems simplifies agent platform since both agent communication and migration can be achieved through the same communication mechanism provided in the agent platform. Flexible synchronization mechanisms have been added for execution and interaction of several mobile agents. This paper proposes a new approach by combining two concept migration and synchronization supports in Mobile-C.

The remainder of the article is structured as follows. In section 2, we present the concept and the properties of mobile agent. Section 3 introduces the architecture of Mobile-C. Section 3 presents the migration of mobile agent over the network from multiple hosts. Section 4 describes the program structure and implementation of the component of agency. Section 5 gives an example of a mobile agent that migrates from hosts via mobile agent messages and illustrates the synchronization support in Mobile-C.

II. MOBILE AGENTS

An agent is defined as "person who's acting on behalf of other people" [30]. In the context of computer science, mobile agent is considered as an entity that moves from one machine to another in the network to perform certain tasks on behalf of the user [31].

Mobile agents have the following properties which distinguish them from other programs [32]:

- **Adaptability** - Mobility of agent required to learn about user's behavior and adapt it to suit the user. Indeed, to evolve adequately the differences between heterogeneous systems, the agent must be able to adopt the changes during the execution.
- **Autonomy**- Mobile agent must be able to make his own decision to be performed to achieve the user's tasks, also he must be able to migrate from one machine to another in the network and execute the user's tasks.
- **Communication** - Mobile agent must have the ability to communicate with others agents of the system in order to exchange information and benefit from the knowledge and expertise of other agents.
- **Mobility**- Mobile agent has the ability to move from one host to another, either by moving the agent's code or by serializing both code and state to allow the agent to continue the execution in a new context.
- **Persistence** - A persistent agent it will be able to retain the knowledge and state over extended period of time to be accessed later on. Once the mobile agent is set up, it is not dependent on system that has been initiated and it is automatically recovered when the agent is terminated or when it is flushed from memory to the database.

III. THE ARCHITECTURE OF MOBILE-C

The system of mobile-C is shown in figure1. Agencies are the major building blocks of the system and abode in each node of a network system in order to support Stationary Agents (SA) and Mobile Agents (MA) at runtime. They serve for locating and messaging agents, moving mobile agents, collecting knowledge about other agents and providing several places where the agent can be run. The core of an agency provides local service for agents and proxies remote agencies. The principle of an agency and their functionalities can be summarized as follows [33]:

- Agent Management system (AMS): The AMS manages the life cycle of agents in the system. It relates the creation, authentication, registration, deletion,execution, migration and persistence of agents. AMS is also responsible for receiving and dispatching mobile agent's .Each agent must register with an AMS in order to get a valid AID.
- Agent Communication Channel (ACC): The ACC forwardsmessages between local and remote entities. The interaction and coordination of mobile agents and host systems can be performed through agent communication language (ACL).
- Agent Security Manager (ASM): The ASM is responsible for protection of access for platform and infrastructure.
- Directory Facilitator: DF serves yellow page services. Agents in the system can register their services with DF for providing to the community. They can also look up required services with DF.
- Agent Execution Engine (AEE): AEE serves as the workhorsefor mobile agents. Mobile agents must reside inside an engine to execute. AEE has to be platform independent in order to support a mobile agent executingin a heterogeneous network.

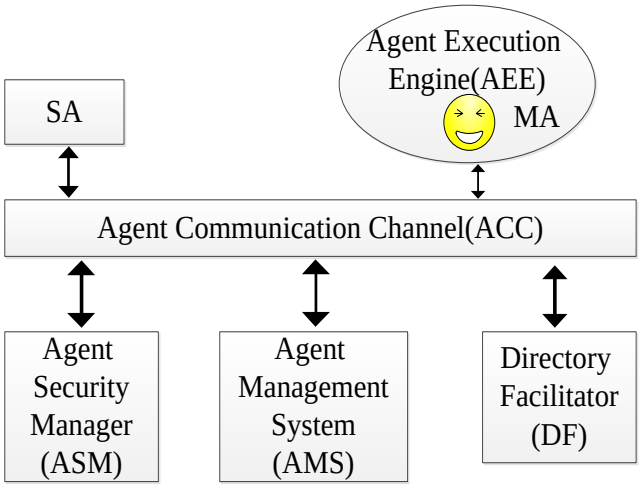


Figure 1. The system architecture of agencies in Mobile-C.

IV. MOBILE AGENT MIGRATION

Mobile agent is a software agent who is able to migrate from one host to another over the network and resume the

execution in the new host. The migration and the execution of mobile agents are supported by a mobile agent system. In previous studies, Chen et al. have developed a mobile agent system called Mobile-C.The Mobile-C supports weak migration. The task of a mobile agent can be divided into several subtasks whichcan be executed in different hosts and listed in a list of tasks as shown in figure 2. The task list can be modified by adding new subtasks and new conditions. Changing dynamically the task list improves the flexibility of a mobile agent. Thus, once we start the execution of a subtask in a host, the mobile agent cannot move until the end of execution.

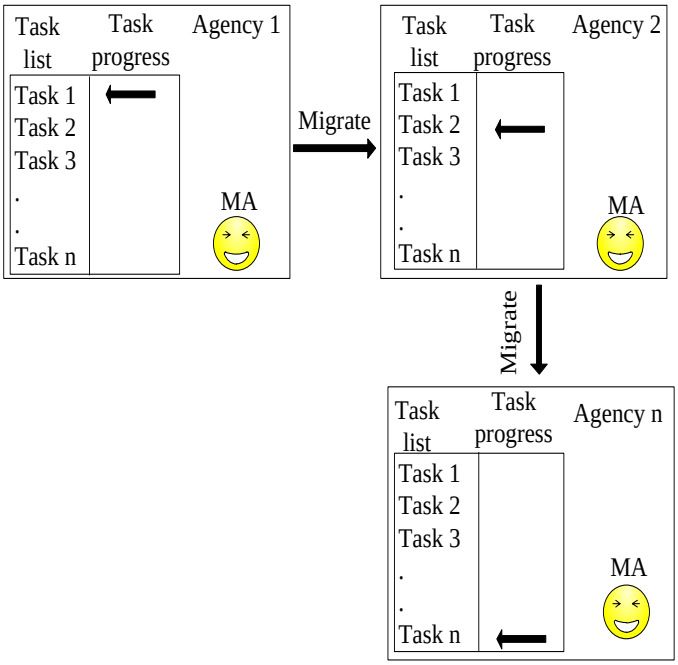


Figure 2. Agent migration based on a task list and a task progress pointer.

Mobile agent migration is achieved through ACL mobile agent messages encoded to XML, which convey mobile agents as the content of a message. Mobile agent message containsthe data state and the code of an agent. The data state of mobile agent include general information about mobile agent as agent name, agent owner and agent home , also the tasks that mobile agent will performed on destination hosts. The data state and code will be wrapping up into an ACL message and transmitted to a remote host trough Agent Communication Channel. Mobile agent migration based on ACL messages is simple and effective for agent migration in FIPA compliant systems because these systems have mandatory mechanisms for message communication, transmission and procession.

V. THE PROGRAM STRUCTURE AND IMPLEMENTATION OF COMPONENT OF AGENCY

An agency is a principle program running in each node of the network [23]. When the execution of an agency is started, the system is initialized and threads are created for all of the components in the agent platform.

After the initialization of the system, the agency waits for defined events. When the agency receives a request to run a mobile agent, it creates a new thread and embeds an Embeddable C/C++ Interpreter – Ch into the thread for executing mobile agent code. After the mobile agent migrates to the other hosts, this thread is terminated automatically (figure 3). If the agency receives a system termination request, the execution of agent platform and the system itself will be finished. In the current implementation, each mobile agent runs in an embeddable Ch inside its own thread.

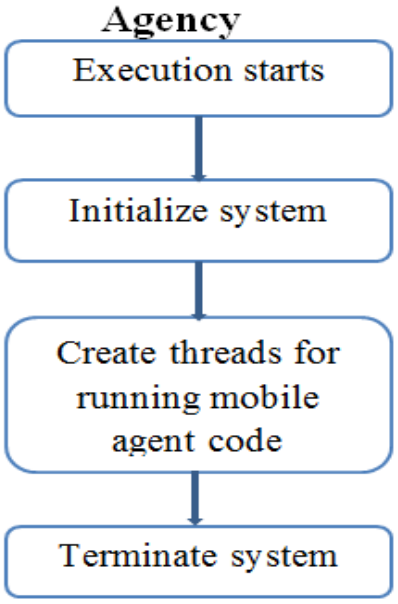


Figure 3. The program structure of an agency

According to the FIPA specifications, each agency should provide mechanisms to receive and send messages. This requirement is satisfied by three components: listening thread, connecting thread and ACC processing thread as shown in Figure 4. The listening thread serves to listen for client connections. When a new connection client is accepted, it will be added to the connection list. Also, the connecting thread is responsible for making connections with other hosts. The ACC processing thread processes the lists of client connections and requests for connecting remote hosts. The ACC facilitates remote agent to agent communication and remote agent platform to agent platform communication via ACL messages. Remote horizontal communication in Mobile-C is implemented on top of TCP/IP and the transport protocol uses HTTP (HyperText Transfer Protocol).

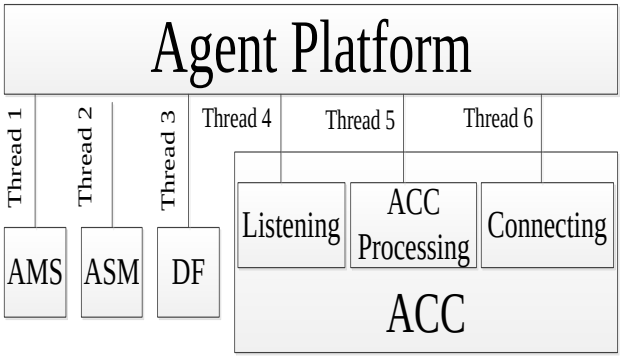


Figure 4. The multi-thread implementation of an agent platform

VI. SYNCHRONIZATION SUPPORT IN MOBILE-C

Amongproperties of mobile agents is the ability to immigrate to perform tasks that exist in the remote host. The purpose of this example is to use an embeddable mobile agent system to protect shared resources by used the synchronization and combines it with the migration of mobile agents from hosts.

The Mobile-C library allows synchronization via mutex. The mutex is a program that allows multiple threads to share the same resource, but not simultaneously.

The example below demonstrates the capabilityof a Mobile-C mutex to protect a resource that may be shared between two or more agents in several hosts. As shown in program 1, a mobile agent is transferred by an agency in the host fsts1 visits remote host fsts2 then host fsts3 .The mobile agent message is represented in Extensible Markup Language (XML), it contains information of the mobile agent and tasks that will be performed on destination hosts. The general information of a mobile agent contains:agent name, agent owner, and the home of the agent. The task information for example the statement <TASK task="2" num="0"> shows that this mobile agent has two tasks to perform and no task has been done yet.The DATA element overall information about the number of element, the name of the task's return variable, the completeness of the task and the host to perform the task.The sub-element DATA ELEMENT contains the return data from the execution task and the sub-element AGENT_CODE contains a C program which will be executed in remote host.

```
<NAME>mobileagent</NAME>
<OWNER>fsts</OWNER>
<HOME>fsts1.fsts.ac.ma:5125 </HOME>
<TASK task= "2" num= "0">
<DATA number_of_elements ="0" name = "results_fsts2"
complete = "0" server = "fsts2.fsts.ac.ma:5138">
<DATA_ELEMENT></ DATA_ELEMENT >
<AGENT_CODE>
Mobile agent code on fsts2
</AGENT_CODE>
</DATA>
<DATA number_of_elements ="0" name = "results_fsts3"
complete = "0" server = "fsts3.fsts.ac.ma:5135">
<DATA_ELEMENT></ DATA_ELEMENT >
<AGENT_CODE>
Mobile agent code on fsts3
</AGENT_CODE>
</DATA>
</TASK>
```

Program 1: The content of the mobile agent message from the host fsts1 to fsts2 and to host fsts3

As shown in Program2, the mobile agent 1" MA1" initialize a mutex with an ID 55 via the function mc_SyncInit() and defines two functions, SetN1() and GetN1() in the host fsts2 . After visiting this host, the mobile agent 1 "MA1" visits the host fsts3 and defines also two functions, SetN2() and GetN2().The result obtained from the host fsts2 is sent to the host fsts3 and the return data will be included in the sub-element DATA_ELEMENT.

```
<NAME>MA1</NAME>
<OWNER>fsts</OWNER>
<HOME> fsts1.fsts.ac.ma:5125 </HOME>
<TASK task= "2" num= "0">
<DATA number_of_elements ="0" name = "results_fsts2"
complete = "0" server = "fsts2.fsts.ac.ma:5138">
<DATA_ELEMENT></ DATA_ELEMENT >
<AGENT_CODE>
<![CDATA[
int N1;
int main () {
intmutex_id=55;
mc_SyncInit(mutex_id);
return 0;
}
void SetN1(inti){
N1 +=i;
if(N1 > 1000){
N1=0;
}
}
intGetN(){
return N
}
]]>
</AGENT_CODE>
</DATA>
<DATA number_of_elements ="0" name = "results_fsts3"
complete = "0" server = "fsts3.fsts.ac.ma:5135">
<DATA_ELEMENT></ DATA_ELEMENT >
<AGENT_CODE>
<![CDATA[
int N2;
int main () {
intmutex_id=55;
mc_SyncInit(mutex_id);
return 0;
}
void SetN2(inti){
N2*=i;
if(N2 > 1000){
N2=0;
}
}
intGetN(){
return N2
}
]]>
</AGENT_CODE>
</DATA>
</TASK>
```

Program2 .A mobile agent that contains a global variable and defines function to access the global variables

As shown in Program 3, the task of the mobile agent 2"MA2" is to perform the operation setting the variable. The operation includes locking the mutexthrough the function mc_MutexLock(), setting the global variable by calling

SetN1() from the host fsts2 and SetN2()from the host fsts3via the function mc_CallAgentFunc(), and unlocking the mutex via the function mc_MutexUnlock().

```
<NAME>MA2</NAME>
<OWNER>fsts</OWNER>
<HOME> fsts1.fsts.ac.ma:5125 </HOME>
<TASK task= "2" num= "0">
<DATA number_of_elements ="0" name = "results_fsts2"
complete = "0" server = "fsts2.fsts.ac.ma:5138">
<DATA_ELEMENT></ DATA_ELEMENT >
<AGENT_CODE>
<![CDATA[
#include <stdio.h>
int main (){
MCAgent_t agent;
inti= 0,mutex_id =55 ,retval;
agent= mc_FindAgentByName("MA1");
wile(1){
mc_MutexLock(mutex_id);
mc_CallAgentFunc(agent,"SetN1",NULL,i);
printf("N1:%d\n",retval);
mc_MutexUnlock(mutex_id)
i++;
if(i==20) {
i=0;
      }
}
return 0;
]]>
</AGENT_CODE>
</DATA>
<DATA number_of_elements ="0" name = "results_fsts3"
complete = "0" server = "fsts3.fsts.ac.ma:5135">
<DATA_ELEMENT></ DATA_ELEMENT >
<AGENT_CODE>
<![CDATA[
#include <stdio.h>
int main (){
MCAgent_t agent;
inti= 0,mutex_id =55 ,retval;
agent= mc_FindAgentByName("MA1");
wile(1){
mc_MutexLock(mutex_id);
mc_CallAgentFunc(agent,"SetN2",NULL,i);
printf("N2:%d\n",retval);
mc_MutexUnlock(mutex_id)
i++;
if(i==20) {
i=0;
      }
}
return 0;
]]>
</AGENT_CODE>
</DATA>
</TASK>
```

Program3. A mobile agent that sets a variable

Likewise, as shown in Program 4, the task of the mobile agent 3“MA3” is locks the mutex, get the global variable, and unlocks the mutex from both the host fsts2 and fsts3

```
<NAME>MA3</NAME>
<OWNER>fsts</OWNER>
<HOME> fsts1.fsts.ac.ma:5125 </HOME>
<TASK task= "2" num= "0">
<DATA number_of_elements ="0" name = "results_fsts2"
complete = "0" server = "fsts2.fsts.ac.ma:5138">
<DATA_ELEMENT></ DATA_ELEMENT >
<AGENT_CODE>
<![CDATA[
#include <stdio.h>
int main (){
MCAgent_t agent;
inti ,mutex_id =55 ,retval;
agent= mc_FindAgentByName("MA1");
mc_MutexLock(mutex_id);
mc_CallAgentFunc(agent,"GetN",&retval,NULL);
printf("N1:%d\n",retval);
mc_MutexUnlock(mutex_id)
return 0;
]]>
</AGENT_CODE>
</AGENT_CODE></DATA>
<DATA number_of_elements ="0" name = "results_fsts3"
complete = "0" server = "fsts3.fsts.ac.ma:5135">
<DATA_ELEMENT></ DATA_ELEMENT >
<AGENT_CODE>
<![CDATA[
#include <stdio.h>
int main (){
MCAgent_t agent;
inti ,mutex_id =55 ,retval;
agent= mc_FindAgentByName("MA1");
mc_MutexLock(mutex_id);
mc_CallAgentFunc(agent,"GetN",&retval,NULL);
printf("N2:%d\n",retval);
mc_MutexUnlock(mutex_id)
return 0;
]]>
</AGENT_CODE>
</DATA>
</TASK>
```

Program4. A mobile agent that gets a variable from another agent.

The results of the mobile agent 1”MA1”, mobile agent 2 “MA2” and mobile agent3 “Ma3” obtained from both the host fsts2 and fts3 are send back to the home agency fsts.

VII. CONCLUSION

In this work we present the design and implementation of an IEEE FIPA compliant agent platform, Mobile-C. Mobile-C integrates an embeddable C/C++ interpreter—Ch—into the platform as a mobile agent execution engine in order to support mobile agent. The migration of mobile agent is achieved through ACL messages. Mobile agents, including both its data state and code, are transported to a remote agent platform via

ACL messages which is encoded in XML, and the execution of mobile agents is resumed by the task progress pointer. The Mobile-C library supports the synchronization among mobile agents and threads because the synchronization functions protect shared resources and provide a way of deterministically timing the execution of mobile agents and the migration to a remote host.

In our future work, this framework will be tested and extended in various types of industrial applications like e-commerce and network management.

ACKNOWLEDGMENT

The authors thank the referees for valuable constructive comments and suggestions which lead to a significant improvement of this paper.

REFERENCES

[1] D.Chess,C.Harrison, A.Kershenbaum,"Mobile Agentes: Are They a Good Idea?", IBM ResearchReport, 1998 [Online] Available : http://www.research.ibm.com/iagentes/paps/mobile -idea.ps.

[2] W. Shen, D. Xue, D.H. Norrie, An agent-based manufacturing enterprise infrastructure for distributed integrated intelligent manufacturing systems, in: Proceedings of the 3rd International Conference on the Practical Applications of Agents and Multi-Agent Systems (PAAM-98), London, UK, 1998, pp. 533–548.

[3] H. Wada, S. Okada, An autonomous agent approach for manufacturing execution control systems, Integrated Computer-Aided Engineering 9 (3) (2002) 251–262.

[4] H.V.D. Parunak, A.D. Baker, S.J. Clark, The AARIA agent architecture: from manufacturing requirements to agent-based system design, Integrated Computer-Aided Engineering 8 (1) (2001) 45–58.

[5] M. Yokoo, S. Fujita, Trends of internet auctions and agent-mediated web commerce, New Generation Computing 19 (4) (2001) 369–388.

[6] T. Sandholm, eMediator: a next generation electronic commerce server, Computational Intelligence 18 (4) (2002) 656–676.

[7] S.P.M. Choi, J. Liu, S. Chan, A genetic agent-based negotiation system, Computer Networks: The International Journal of Computer and Telecommunications Networking 37 (2) (2001) 195–204.

[8] W.E. Chen, C. Hu, A mobile agent-based active network architecture for intelligent network control, Information Sciences 141 (1–2) (2002) 3–35.

[9] L. Chou, K. Shen, K. Tang, C. Kao, Implementation of mobile-agent-based network management systems for national broadband experimental networks in Taiwan, Holonic and Multi-Agent Systems for Manufacturing (Lecture Notes in Computer Science) 2744 (2003) 280–289.

[10] J. Huang, N.R. Jennings, J. Fox, Agent-based approach to health care management, Applied Artificial Intelligence 9 (4) (1995) 401–420.

[11] I. Noda, P. Stone, The RoboCup soccer server and CMUnited clients: implemented infrastructure for MAS research, Autonomous Agents and Multi-Agent Systems 7 (1–2) (2003) 101–120.

[12] K.Straber, J.Baumann and F.Hohl.. Mole - A Java Based Mobile Agent System. Institute for Parallel and Distributed Computer Systems, University of Stuttgart,1997

[13] D. Lange, M.Oshima. Programming and Deploying Java Mobile Agents with Aglets. Addison-Wesley: MA, 1998.

[14] D.Wong, N.Paciorek, T.Walsh, J.DiCelie, M.Young, B.Peet. Concordia: An infrastructure for collaborating mobile agents. Proceedings of the First International Workshop on Mobile Agents (MA'97) (Lecture Notes in Computer Science, vol. 1219). Springer: Berlin, 1997; 86–97.

[15] F.Bellifemine, G.Caire, A.Poggi, G.Rimassa.JADE: A software framework for developing multi-agent applications.Lessons learned. Information and Software Technology 2008; 50(1–2):10–21.

[16] R.Gray, G.Cybenko, D.Kotz,R.Peterson, D.Rus. D'Agents: Applications and performance of a mobile-agent system. Software—Practice and Experience 2002; 32(6):543–573.

[17] H.Peine. Run-time support for mobile code. PhD Dissertation, Department of Computer Science, University of Kaiserslautern, Germany, 2002.

[18] H.Peine .Application and programming experience with the Ara mobile agent system. Software—Practice and Experience 2002; 32(6):515–541.

[19] D.ohnansen, K.Lauvset, R.V.Renesse, F.B. Schneider, N.P. Sudmann, K. Jacobsen. A TACOMA retrospective.Software—Practice and Experience 2002; 32(6):605–619.

[20] MACE—Mobile agent code environment. Available at: http://wwwagss.informatik.uni-kl.de/Projekte/Ara/mace.html [last modified 10 August 2004].

[21] N.P.Sudmann,D.Johansen. Adding mobility to non-mobile web robots. Proceedings of the IEEE ICDCS00 Workshop on Knowledge Discovery and Data Mining in the World-wide Web, Taipei, Taiwan, 2000; 73–79.

[22] B.Chen, H.H.Cheng. A run-time support environment for mobile agents. Proceedings of ASME/IEEE International Conference on Mechatronic and Embedded Systems and Applications, No. DETC2005-85389, Long Beach, CA, September 2005.

[23] B.Chen,H.H.Cheng,J.Palen. Mobile-C: A mobile agent platform for mobile C/C++ agents. Software—Practice and Experience 2006; 36(15):1711–1733.

[24] B.Chen, D.Linz, H.H.Cheng. XML-based agent communication, migration and computation in mobile agent systems. Journal of Systems and Software 2008; 81(8):1364–1376.

[25] Mobile-C: A multi-agent platform for mobile C/C++ code. Available at: http://www.mobilec.org.

[26] Y.C. Chou, D. Ko, H. H. Cheng, An embeddable mobile agent platform supporting runtime code mobility,interaction and coordination of mobile agents and host systems, Information and Software Technology 52 (2010) 185–196

[27] H.H.Cheng. Scientific computing in the Ch programming language. Scientific Programming 1993; 2(3):49–75

[28] H.H.Cheng.Ch: A C/C++ interpreter for script computing. C/C++ User's Journal 2006; 24(1):6–12.

[29] Ch—An embeddable C/C++ interpreter. Available at: http://www.softintegration.com .

[30] P.M .Reddy, Mobile Agents Intelligent Assistants on the Internet, July 2002

[31] A. Kaur , S.Kaur, Role of Mobile Agents In Mobile Computing, Proceedings of National Conference on Challenges & Opportunities in Information Technology (COIT-2007) RIMT-IET

[32] M. L. Griss, Software Agents as Next Generation Software Components, Chapter 36 in Component-Based Software Engineering: Putting the Pieces Together, Edited by George T. Heineman, Ph.D. & William Councill, M.S., J.D., May 2001,Addison-Wesley

[33] B. Chen, H. H. Cheng, J.Palen, Integrating mobile agent technology with multi-agent systems for distributed traffic detection and management systems, Transportation Research Part C 17 (2009) 1–10

An Adaptive parameter free data mining approach for healthcare application

Prof. Dipti Patil[#1]
[#] Asst. Professor, Computer Engg. Dept, MITCOE Pune, India

Bhagyashree Agrawal[#2]
[#]Computer Engg. Dept, MITCOE Pune, India

Snehal Andhalkar[*3]
[#]Computer Engg. Dept, MITCOE Pune, India

Richa Biyani[#4]
[#]Computer Engg. Dept, MITCOE Pune, India

Mayuri Gund[#5]
[#]Computer Engg. Dept, MITCOE Pune, India

Dr. V.M.Wadhai[#6]
[#]Professor, Computer Engg. Dept, MITCOE Pune, India

***Abstract*—In today's world, healthcare is the most important factor affecting human life. Due to heavy work load it is not possible for personal healthcare. The proposed system acts as a preventive measure for determining whether a person is fit or unfit based on his/her historical and real time data by applying clustering algorithms viz. K-means and D-stream. Both clustering algorithms are applied on patient's biomedical historical database. To check the correctness of both the algorithms, we apply them on patient's current biomedical data. The Density-based clustering algorithm i.e. the D-stream algorithm overcomes drawbacks of K-means algorithm. By calculating their performance measures we finally find out effectiveness and efficiency of both the algorithms.**

Keywords- Data stream mining; clustering; healthcare applications; medical signal analysis.

I. INTRODUCTION

Today the health care industry is one of the largest industries throughout the world. It includes thousands of hospitals, clinics and other types of facilities which provide primary, secondary & tertiary levels of care. The delivery of health care services is the most visible part of any health care system, both to users and the general public [2]. A health care provider is an institution or person that provides preventive, curative, promotional or rehabilitative health care services in a systematic way to individuals, families or community. The physiological signals such as SpO2, ABPsys, ABPdias, HR affects person's health. In health care the data mining is more popular and essential for all the healthcare applications. In healthcare industry having the more amounts of data, but this data have not been used properly for the application. In this health care data is converted in to the useful purpose by using the data mining techniques [1].

The data mining is the process of extracting or mining the knowledge from the large amounts of data, database or any other data base repositories. The main purpose of the data mining is to find the hidden knowledge from the data base. In health care industry, the data having some unwanted data, missing values and noisy data. This unwanted data will be removed by using preprocessing techniques in data mining. Preprocessing is the process of removing noise, redundant data and irrelevant data. After the preprocessing the data will be used for some useful purpose. In recent years different approaches are proposed to overcome the challenges of storing and processing of fast and continuous streams of data.

Data stream can be conceived as a continuous and changing sequence of data that continuously arrive at a system to store or process. Traditional OLAP and data mining methods typically require multiple scans of the data and are therefore infeasible for stream data applications. Whereby data streams can be produced in many fields, it is crucial to modify mining techniques to fit data streams. Data stream mining has many applications and is a hot research area[3]. Data stream mining is the extraction of structures of knowledge that are represented in the case of models and patterns of infinite streams of information. These data stream mining can be used to form the clusters of medical health data. This paper proposed two main clustering algorithms namely, K-means algorithm and density based clustering.

The K-means clustering algorithm is incompetent to find clusters of arbitrary shapes and cannot handle outliers. Further, they require the knowledge of k and user-specified time window. To address these issues, D Stream, a framework for clustering stream data using a density-based approach. The algorithm uses an online component which maps each input data record into a grid and an offline component which computes the grid density and clusters the grids based on the density. The algorithm adopts a density decaying technique to capture the dynamic changes of a data stream. Exploiting the intricate relationships between the decay factor, data density and cluster structure, our algorithm can efficiently and effectively generate and adjust the clusters in real time. Further, a theoretically sound technique is developed to detect and remove sporadic grids mapped to by outliers in order to dramatically improve the space and time efficiency of the system. The technique makes high-speed data stream clustering

feasible without degrading the clustering quality. The experimental results show that our algorithm has superior quality and efficiency, can find clusters of arbitrary shapes, and can accurately recognize the evolving behaviors of real-time data streams [4].

II. Related Work

All Several health care projects are in full swing in different universities and institutions, with the objective of providing more and more assistance to the elderly. The application of data clustering technique for fast retrieval of relevant information from the medical databases lends itself into many different perspectives.

Health Gear: a real-time wearable system for monitoring and analyzing physiological signals [5] is a real-time wearable system for monitoring, visualizing and analyzing physiological signals. This system focused on an implementation of Health Gear using a blood oximeter to monitor the user's blood oxygen level and pulse while sleeping. The system also describes two different algorithms for automatically detecting sleep apnea events, and illustrates the performance of the overall system in a sleep study with 20participants. A Guided clustering Technique for Knowledge Discovery – A Case Study of Liver Disorder Dataset, [6] presents an experiment based on clustering data mining technique to discover hidden patterns in the dataset of liver disorder patients. The system uses the SOM network's internal parameters and k-means algorithm for finding out patterns in the dataset. The research has shown that meaningful results can be discovered from clustering techniques by letting a domain expert specify the input constraints to the algorithm.

Intelligent Mobile Health Monitoring System (IMHMS), [7] Author proposed the system which can provide medical feedback to the patients through mobile devices based on the biomedical and environmental data collected by deployed sensors. The system uses the Wearable Wireless Body/Personal Area Network for collecting data from patients, mining the data, intelligently predicts patient's health status and provides feedback to patients through their mobile devices.

The patients will participate in the health care process by their mobile devices and thus can access their health information from anywhere any time. But actual implementation of data mining framework for decision support system is not done.

Real-Time analysis of physiological data to support medical applications [8], proposed a flexible framework to perform real-time analysis of physiological data and to evaluate people's health conditions. Patient or disease-specific models are built by means of data mining techniques. Models are exploited to perform real time classification of physiological signals and continuously assess a person's health conditions. The proposed framework allows both instantaneous evaluation and stream analysis over a sliding time window for physiological data. But dynamic behavior of the physiological signals is not analyzed also the framework is not suitable for ECG type of signals.

Performance of Clustering Algorithms in Healthcare Database [1], proposed a framework where they used the heart attack prediction data for finding the performance of clustering algorithm. In final result shows the performance of classifier algorithm using prediction accuracy and the visualization of cluster assignments shows the relation between the error and the attributes. The comparison result shows that, the make density based clusters having the highest prediction Accuracy.

III. Method

We present a framework that will perform clustering of dataset available from medical database effective manner. The flow of the system is depicted in Fig.1.

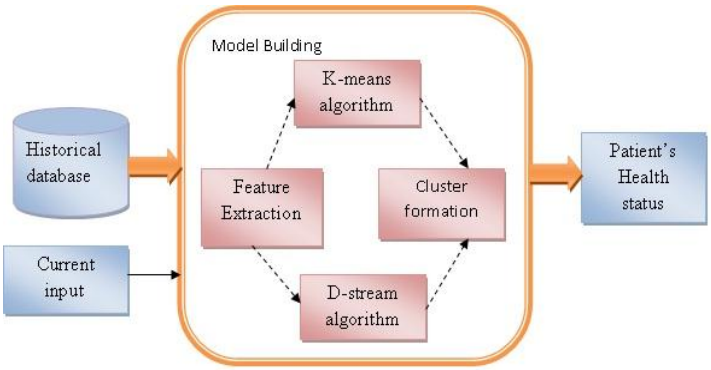


Figure 1.Flow of the system

The target is to cluster the patient's records into different groups with respect to the test report attributes which may help the clinicians to diagnose the patient's disease in efficient and The evaluation steps are the following-

1) Data set collection

The data set contains 7 attributes, SpO2, ABPsys, ABPdias, HR, heredity, obesity, cigarette smoking. These attributes are the risk factors that can help in predicting the patient's health status. Attributes such as SpO2, ABPsys, ABPdias, HR can be collected form MIMIC database [9] and the other attributes are influenced by the person's behavior. These all attributes values are discrete in nature .The dataset will be in preprocessed format.

2) Model Building

In model building phase features of the available data will be extracted and then clustering algorithm will be applied on extracted features.

A. Feature Extraction

For each physiological signal x among the X monitored vital signs, we extract the following features [8].

1) Offset

The offset feature measures the difference between the current value x(t) and the moving average (i.e., mean value over the time window). It aims at evaluating the difference between the current value and the average conditions in the recent past.

2) Slope

The slope function evaluates the rate of the signal change. Hence, it assesses short-term trends, where abrupt variations may affect the patient's health.

3) Dist

The dist feature measures the drift of the current signal measurement from a given normality range. It is zero when the measurement is inside the normality range.

B. Risk Components

The signal features contribute to the computation of the following risk components.

1) Sharp changes

The z1 component aims at measuring the health risk deriving from sharp changes in the signal (e.g., quick changes in the blood pressure may cause fainting)

2) Long-term trends

The z2 component measures the risk deriving from the h weighted offset over the time window. While z1 focuses on quick changes, z2 evaluates long-term trends, as it is offset-based.

3) Distance from normal behavior

The z3 component assesses the risk level given by the distance of the signal from the normality range. A patient with an instantaneous measurement outside the range may not be critical, but her/his persistence in such conditions contributes to the risk level

From above risk components, risk functions and global risk components will be calculated. These values will be further used in clustering algorithms as an input for cluster formation.

C. Cluster formation

The proposed flow of the system uses two algorithms K-means and D-stream. The comparison between two clustering algorithms will be performed using the above described attributes.

K-MEANS ALGORITHM

1) The algorithm is composed of the following steps: [10] Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids. Assign each object to the group that has the closest centroid. When all objects have been assigned, recalculate the positions of the K centroids.

Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Figure 2. Algorithm K-means algorithm

D-STREAM ALGORITHM

The D-stream algorithm is explained as follows [4]

1.procedure D-Stream
2. Tc = 0;
3. Initialize an empty hash table grid list;
4. while data stream is active do
5. read record x = (x1, x2, · · · , xd);
6. determine the density grid g that contains x;
7. if(g not in grid list) insert g to grid list;
8. update the characteristic vector of g;
9. if tc == gap then
10. call initial clustering(grid list);
11. end if
12. if tc mod gap == 0 then
13. detect and remove sporadic grids from grid list;
14. call adjust clustering(grid list);
15. end if
16. tc = tc + 1;
17. end while
18. end procedure

Figure 3: The overall process of D-Stream.

The overall architecture of D-Stream, which assumes a discrete time step model, where the time stamp is labeled by integers 0, 1, 2, · · · , n, · · · . D-Stream has an online component and an offline component. The overall algorithm is outlined in Figure 1.

For a data stream, at each time step, the online component of D-Stream continuously reads a new data record, place the multi-dimensional data into a corresponding discretized density grid in the multi-dimensional space, and update the characteristic vector of the density grid (Lines 5-8 of Figure 3). The density grid and characteristic vector are to be described in detail later. The offline component dynamically adjusts the clusters every gap time steps, where gap is an integer parameter. After the first gap, the algorithm generates the initial cluster (Lines 9-11). Then, the algorithm periodically removes sporadic grids and regulates the clusters (Lines 12-15).

D-Stream partitions the multi-dimensional data space into many density grids and forms clusters of these grids. This concept is schematically illustrated in Figure 4.

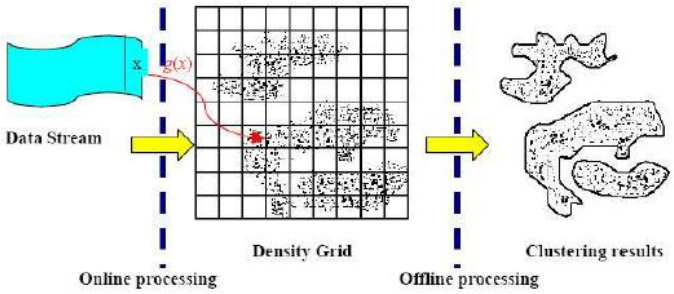


Figure 4.Illustration of the use of density grid.

The input data has d dimensions, and each input data record is defined within the space

$$S = S_1 \times S_2 \times \cdots \times S_d, \quad \text{..... (1)}$$

where Si is the definition space for the ith dimension.

In D-Stream, we partition the d−dimensional space S into density grids. Suppose for each dimension, its space Si, I = 1, · · · , d is divided into pi partitions as

$$S_i = S_{i,1} \bigcup S_{i,2} \bigcup \cdots \bigcup S_{i,p_i},$$ -----(2)

then the data space S is partitioned into N density grids. For a density grid g that is composed of $S_{1,j_1} \times S_{2,j_2} \cdots \times S_{d,j_d}, j_i = 1, \ldots, p_i,$ we denote it as $g = (j_1, j_2, \cdots, j_d).$ ----------(3)

A data record x = (x1, x2, · · · , xd) can be mapped to a density grid g(x) as follows:

$g(x) = (j_1, j_2, \cdots, j_d)$ where xi 2 Si,ji .

For each data record x, we assign it a density coefficient which decreases with as x ages. In fact, if x arrives at time tc, we define its time stamp T(x) = tc, and its density coefficient D(x, t) at time t is

$$D(x,t) = \lambda^{t-T(x)} = \lambda^{t-t_c},$$ ----------(4)

where λϵ (0, 1) is a constant called the decay factor.

Definition (Grid Density) For a grid g, at a given time t, let E(g, t) be the set of data records that are map to g at or before time t, its density D(g, t) is defined as the sum of the density coefficients of all data records that mapped to g. Namely, the density of g at t is:

$$D(g,t) = \sum_{x \in E(g,t)} D(x,t).$$

1) procedure initial clustering (grid list)
2) update the density of all grids in grid list;
3) assign each dense grid to a distinct cluster;
4) label all other grids as NO CLASS;
5) repeat
6) for each cluster c
7) for each outside grid g of c
8) for each neighboring grid h of g
9) if (h belongs to cluster c')
10) if (|c| > |c'|) label all grids in c' as in c;
11) else label all grids in c as in c';
12) else if (h is transitional) label h as in c;
13) until no change in the cluster labels can be made
14) end procedure

Figure 3: The procedure for initial clustering.

1) procedure adjust clustering (grid list)
2) update the density of all grids in grid list;
3) for each grid g whose attribute (dense/sparse/transitional) is changed since last call to adjust clustering()
4) if (g is a sparse grid)
5) delete g from its cluster c, label g as NO CLASS;
6) if (c becomes unconnected) split c into two clusters;
7) else if (g is a dense grid)
8) among all neighboring grids of g, find out the grid h whose cluster ch has the largest size;
9) if (h is a dense grid)
10) if (g is labeled as NO CLASS) label g as in ch;
11) else if (g is in cluster c and |c| > |ch|)
12) label all grids in ch as in c;
13) else if (g is in cluster c and |c| <= |ch|)
14) label all grids in c as in ch;
15) else if (h is a transitional grid)
16) if ((g is NO CLASS) and (h is an outside grid if g is added to ch)) label g as in ch;
17) else if (g is in cluster c and |c| >=|ch|)
18) move h from cluster ch to c;
19) else if (g is a transitional grid)
20) among neighboring clusters of g, find the largest one c' satisfying that g is an outside grid if added to it;
21) label g as in c';
22) end for
23) end procedure

Figure 4: The procedure for dynamically adjusting clusters.

The calculated values of z1, z2, z3 components will be applied as an input for both the clustering algorithms to form the clusters based on their risk level.

3) Patient's Health status

Using clustering algorithm we form the clusters for attributes stated above. And then for patient's current input we predict patient's health status i.e. patient is fit or unfit.

IV. Experimental Result

The above described algorithms used for formation of clusters on medical database. The data will be collected from the Switzerland data set. The data set contains the 107 instances and the 14 attributes. The attributes are age, sex, Blood Pressure, Cholesterol, Chest Pain and etc. The performance of these algorithms will be computed by using correctly predicted instance. [1]

Performance Accuracy= correctly predicted Instance/
Total Number of Instance

Table I. Performance of Clustering Algorithm

Cluster Category	Cluster Algorithm	Measures		
Clusters		Correctly Classified Instance	In correctly Classified Instance	Prediction Accuracy
	Simple K-means	89	18	83.18
	D-stream	94	13	87.85

From above table we observed that the performance of density based algorithm is better than simple K-means. Accuracy of D-stream algorithm is more than K-means.

V. Future Scope and Conclusion

K-means is unable to handle arbitrary cluster formation because prediction of the number of classes to be formed is not fixed. The D-stream algorithm has superior quality and efficiency, can find clusters of arbitrary shapes, and can accurately recognize the evolving behaviors of real-time data streams. Therefore, D-stream will perform better in biomedical applications. This system can be further developed for real time analysis of biomedical data to predict patient's current health status.

The proposed system can be used for monitoring elderly people, Intensive Care Unit (ICU) Patient. Also the system gives the health status of patient, it can be used be used by clinicians to keep the records of patients.

The proposed system is adaptive since it can handles more than one physiological signal. The proposed system uses historical biomedical data which is very useful for prediction of current health status of a patient by using clustering algorithms like K-means, D-stream, etc. Prediction of health status is very sensitive job, D-stream will perform better here, as it supports arbitrary cluster formation which is not supported by K-means. Also D-stream is particularly suitable for users with little domain knowledge on the application data that means it won't require the K-values. Hence D-stream is parameter free and proves to give more accurate results than K-means when used for cluster formation of historical biomedical data.

References

[1] P.Santhi, V.Murali Bhaskaran Computer Science & Engineering Department Paavai Engineering College, "Performance of Clustering Algorithms in Healthcare Database", International Journal for Advances in Computer Science, Volume 2, Issue 1 March 2010

[2] Sellappan Pandian, Rafigh Awang,"Heart Disease Prediction System using Data Mining Techniques",IEEE Computer, Vol 7, PP.295-304,August2008.

[3] Mahnoosh Kholghi, Mohammadreza Keyvanpour, "An Analytical Framework For Data Stream Mining Techniques Based On Challenges And Requirements", Mahnoosh Kholghi et al. / International Journal of Engineering Science and Technology (IJEST), Vol. 3 No. 3 Mar 2011

[4] Yixin Chen, Li Tu, "Density-Based Clustering for Real-Time Stream Data"in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007

[5] Nuria Oliver , Fernando Flores-Mangas, "HealthGear: A Real-time Wearable System for Monitoring and Analyzing Physiological signals" in proceeding BSN'06 Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks, 2006

[6] Vikram Singh, Sapna Nagpal "A Guided clustering Technique for Knowledge Discovery – A Case Study of Liver Disorder Dataset", International Journal of Computing and Business Research, Vol.1, no. 1, Dec 2010

[7] RifatShahriyar, Md. Faizul Bari, GourabKundu, Sheikh IqbalAhamed and Md. Mustofa Akbar 5,"Intelligent Mobile Health Monitoring System(IMHMS)", International Journal of Control and Automation, vol 2,no.3, Sept 2009, pp 13-27.

[8] Daniele Apiletti, Elena Baralis, Member, IEEE, Giulia Bruno, and Tania Cerquitelli, "Real-Time Analysis of Physiological Data to Support Medical Applications", IEEE Transactions On Information Technology In Biomedicine, Vol. 13, No. 3, May 2009.

[9] The MIMIC database on PhysioBank (2007, Oct.) [Online]. Available: http://www.physionet.org/physiobank/database/mimicdb

[10] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques" , Second Edition

Question Answering System for an Effective Collaborative Learning

Prof. Kohei Arai[1]
Information Science,
Saga University - Japan

Anik Nur Handayani [1,2]
Electrical and Information Technology
State University of Malang - Indonesia

***Abstract*— The increasing advances of Internet Technologies in all application domains have changed life styles and interactions. With the rapid development of E-Learning, collaborative learning is an important for teaching, learning methods and strategies. Interaction between the students also student with the teacher is important for student to gain knowledge. Based on the four basic teaching styles formal authority, demonstrator or personal model, facilitator and delegator, today combined between facilitator and delegator style is responsible for student learning. It is student centered and the teacher as facilitates the material and activities, but learning becomes part of valuable and effective when they collaborate with each other, and as the teacher who will delegates and facilitates the responsibility of learning to the students. In this paper, we introduce an effective question answering Q&A system for collaborative learning, which can act not just like a virtual teacher, but also virtual discussion for student. With the proposed system, brings a new Q&A system, student can attach their question when they want collaborate using collaborative learning capitalize on one another's resources and skills. Students can ask their questions to the group when they want to collaborate with others, asking one another for information, evaluating one another's ideas, then each of the answer will compare with encyclopedia data base. In this research, the Q&A system for the Senior High School in Indonesia, in this subject of Information Communication Technology implemented. From the the 40 question and 120 answer, the result is 90,48% precision 50% recall.**

Keywords – component; E-Learning; Collaborative Learning; Q&A; Knowledge Base.

I. INTRODUCTION

The concept of Collaborative Learning is two or more people learn or attempt to learn something together than independent. Different with individual learning, in collaborative learning people can exploit and share their resources and skills by asking for information, evaluating, monitoring one another's information and idea, etc [1]. Collaborative Learning is a model that knowledge can be created by sharing experiences within a population where members actively interact [2] [3]. In the Collaborative Robotic Instruction (A Graph Teaching Experience Computers and Education), the goal of collaborative learning is methodologies and environments which learners or students where each depends on and is responsible to each other [4]. Including both directly with face-to-face conversations [5] or using computer discussions (online forums, chat rooms, etc.) [6].

In [3] authors indicate that when they found some problem, students learn better when they learn together more frequently than working individually as members in a group. Indeed, the effectiveness of collaborative learning on the internet has been identified by various studies. Interaction among learners is fostered as communication over the internet is unpretentious and convenient when addressing to a single user or multiple users. However, interaction students and a teacher address a problem, the teacher cannot constantly online every time, and it is not possible for the teacher to deal with lots of question from students all the time and in a timely manner. In the collaborative learning, students are encouraged to ask question [7]. Therefore, there is a need to describe an automated Q and A system to support learning efficiency of collaborative learning.

In this paper, we proposed question answering system for an effective collaborative learning. With the proposed system, brings a new Q&A system, student can attach their question when they want collaborative using collaborative learning capitalize on one another's resources and skills. Students can ask their questions to the group when they want to collaborate with others, asking one another for information, evaluating one another's ideas. The method also considers that students can communication through Q&A interaction such a discussion forum to support information sharing. The paper is organized as follows. First is the motivation for the question and answering system. Section 2 presents the related works on collaborative learning. Section 3 explains about the question answering mechanism proposed. In section 4, present the architecture of system that we proposed an interactive effective Q&A system for collaborative learning. The user interactions between user and the system are explained. Finally, section 6, are summary and conclusion of this paper.

II. RELATED WORK

Collaborative learning is one of the study groups. Some studies show that students get the most current learning through group rather than independently [8] [9]. Studies by the OTTER Group [10] have shown that the ideal class is organized around 50/50 rule. At least 50% of the time students spend is spent interacting with and learning about the other student in the virtual classroom. The social aspect of the classroom is an important factor. If social aspect missing, than student dissatisfaction rises dramatically, as does the attrition rate. In this learning mode, which is collaborative learning, students who are interested in sharing their knowledge from a learning group to communicate and discuss all kinds of questions, asking one another for information, evaluating one another's

idea for help and teach each other. Therefore, learning is both a group activity and a social process and thus learning performance is strongly affected peers [11].

In the development of networks, comprise all forms of electronically supported learning and teaching that can eliminate the obstacles of time and space. In the collaborative learning, students can take part by computer at anyplace, at the same time or different time (synchronous and asynchronous). Researchers have used activity theory to analyze Computer Supported Collaborative Work (CSCW) system [12]. Group communication relationship [13] refers to the intra-group relationships determined by the interactions among members. However, how to form a learning group after the group is a problem in collaborative learning.

Several study about Q&A for collaborative learning had been done. An application of Question Answering System for Collaborative Learning has been designed. In this application learners can attach their question to the group when they want to collaborate with others, and the teacher providing answers to them. In this case, the collaborative between the students and the teacher to gain knowledge, and becomes question answering system like a virtual teacher [7] [14]. In this paper, we proposed question answering system for an effective collaborative learning. The originality of the system is bringing a new Q&A system, in which students can attach their questions to the group when they want to gaining and sharing knowledge with others, by collaborative learning capitalize on one another's resources and skills (asking and evaluating one another's idea), then each of the answer will compare with encyclopedia data base.

III. Proposed E-Learning System With Knowledge-Base System

E-learning system, which is proposed here is based on knowledge base system which allows acquires knowledge about students and uses the acquired knowledge. Once students learn with the proposed system through questions and answers, then all students can use the previously acquired Q&A information which relates to the current question. Figure 1 shows the complete overview proposed of Question Answering Mechanism.

A. Proposed Question Answering System

An online Wikipedia-encyclopedia was chosen as the corpus for the task. Wikipedia is a free, web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation. Indonesian Wikipedia is the edition of Wikipedia in the Indonesian language; there were over 100.000 articles in the Indonesian Wikipedia project. In the research proposed, we using Indonesian Q&A system specifically in the Information Communication Technology Subject, for the Senior High School.

An automated Q&A system in collaborative learning operates proposed work based on the question answering knowledge base. When a student needs some information, he or she can ask a question through a designed interface. When a new asked question enters the system, query is created. Then other students will response the question with answering and evaluating one another's ideas by vote. This representation is then compared with the representations of Wikipedia data base.

A similarity percentage is given between the student answer and any existing Wikipedia data base. Based on the biggest vote some time is not describing the best answer, after time for answering and voting finish, similarity percentage will be show for every answer. It aims to estimate whether which one of the answered considered with the Wikipedia data base. Also, student can access the topic question more clearly through the link provided. Such as [7], if no any information from Wikipedia comparing with neither student answer, nor the student is satisfied with any the answers (no proper match from the knowledge base), and then the teacher will answer or might be sent to the student. After the teacher manually answer the question, the new Q&A set is formed and entered into the question answering knowledge base. When student meeting some difficulties or having no difficulties, a student can see what problems other students have encountered in learning now and in the past and see the answers or solutions the teacher offered by browsing the knowledge base. The whole process of question answering in the effective of collaborative learning is shown in figure 1.

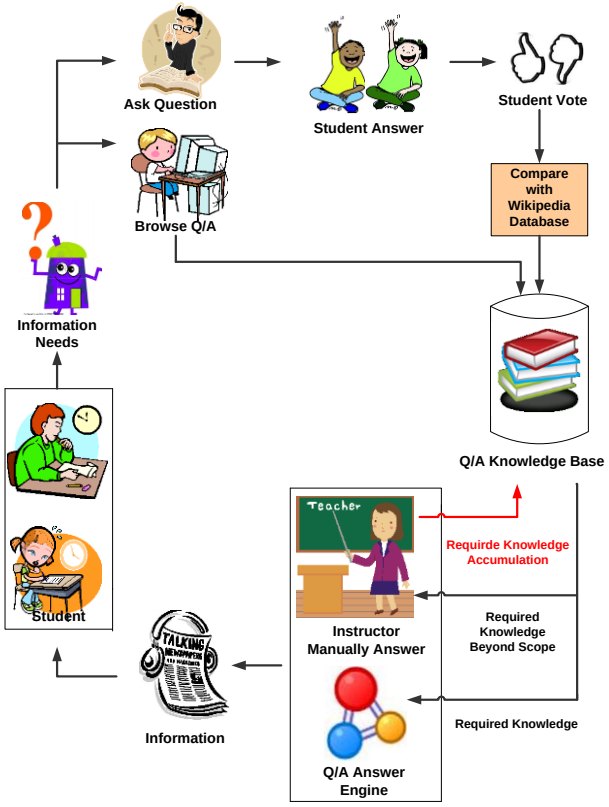


Figure 1. Question Answering Mechanism

B. Architecture of Question and Answering System

The architecture of the question answering system is shown in figure 2. There are eight main components in the system, including the student agent Q, the student agent A, question analysis and query generation, Q&A acquirer, question answering knowledge base, similarity machine, Q&A browsing component, answer generator.

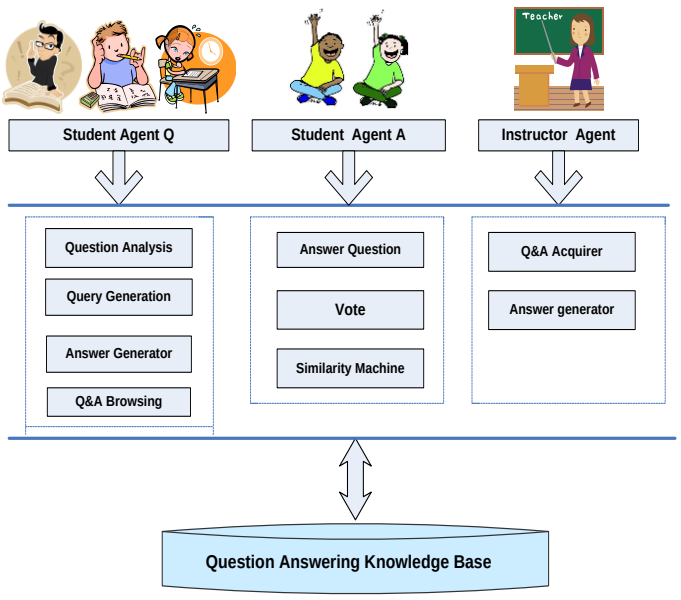


Figure 2. Architecture of Question Answering System

The functions of the components are briefly described as follows,

- Student Agent Q

Student agent Q is to be the interface between student and the system. Student can send his or her question and receive answers (feedback from other student) from this interface. Beside from student also provide feedback from the teacher for manually answering and question if the system are not satisfied.

- Student Agent A

Student agent A is the component interface between student and the system provides an answer and voting question. In this component also provide similarity machine to match student answer with Wikipedia database, to ensure the answer considering biggest vote not necessary correct answer.

- Question analysis, query generation

An analyzed question represents an asyntactic and semantic analysis of a question. It serves as an interface between the question analysis and query generation stages. A query is a search engine query generated at the query generation stage and executed at the search stage.

- Q&A Acquirer

If there is an unanswered question, the questions will temporarily stores at this component. When the teacher gets online, and then will manually answer the question with an interface (Instructor Agent). This component are very important for a reason that student's question might be not answer by other student or student's answer cannot compare with Wikipedia database, cause the information not relevant. A new Q&A will be formed and saved to the Q&A knowledge base, after a question replied by the teacher.

- Question Answering Knowledge Base

A knowledge base provides a means for collected, organized, shared, searched and utilized information. Q&A knowledge base is a knowledge base where questions corresponding with answer. The knowledge will be accumulated and rich, through the process of student question answer and manually teacher answer.

- Q&A Browsing Component

Beside asking a question, when student meeting some difficulties or having no difficulties, a student can see what problems other students have encountered in learning now and in the past and see the answers or solutions the teacher offered by browsing the knowledge base.

- Answer Generator

Several interfaces have been designed, to let students and teacher is able to interact with the system. The interfaces for the student to ask question and get answers (Student Agent A) are developed as shown in Figure 3. Under the interface in figure 3, student can key in an Indonesian question and get an answer immediately without waiting for the teacher to get online. Then other student can answer the question and can discuss whether which one of their answer as same as their opinion. After that, Q&A system makes a point for each of student answer that closely with Q&A system answer. If no satisfactory answer is found, then it will present the teacher demanding a manual answer whenever he or she is online.

- Similarity Machine

This component uses Levenstein and Word similarity method to calculate the similarity between students answer and Wikipedia database. In the question analysis, extracting question query is necessary. Specific interrogation all has question query focus, usually question word (Qw) is the question feature, such Qw as what, when, where, who, why and how. In this system we using what, when, where, who type of question. How and why, are notable mission because it's procedural answer. Using Levenstein for what question (Definition expectations), to calculate percentage Levenstein distance between two length string (students answer and the knowledge base answer). And using Word similarity for when, where, and who question. In the word similarity, the distance of words is a real number in [0, ∞] a word and its own distance is zero. The following equation is word similarity formula.

$$Sim\ (W1, W2) = \frac{\propto}{Dis\ (W1, Ws) + \propto} \qquad (1)$$

Where α is an adjustable parameter.

C. User Interactions with Question and Answering System

Several interfaces have been designed, to let students and teacher be able to interact with the system. The interfaces for the student to ask questions and get answers (Student Agent A) are developed as shown in figure 3. Under the interface in figure 3, student can key in an Indonesian question and get answer immediately without waiting for the teacher to get online. Then other student can answer the question and can discuss whether which one of their answer as same as their opinion. After that, Q&A system makes point for each of student answer that closely with Q&A system answer. If no satisfactory answer is found, then student can press a button to

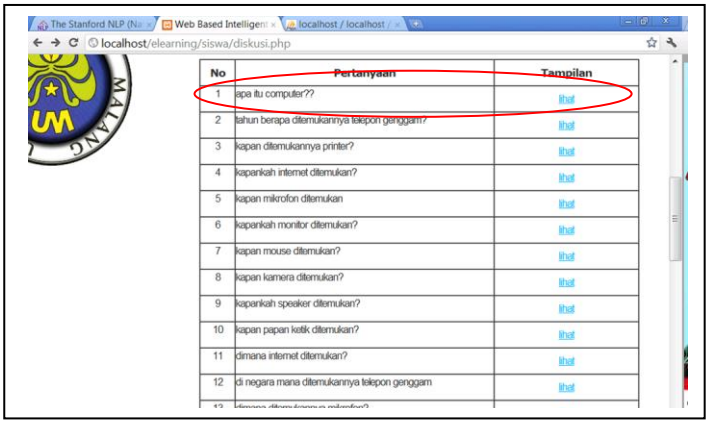
send the question to the Q&A acquirer, which will then present the answered questions to the teacher demanding for a manual answer whenever he is online.

(a)

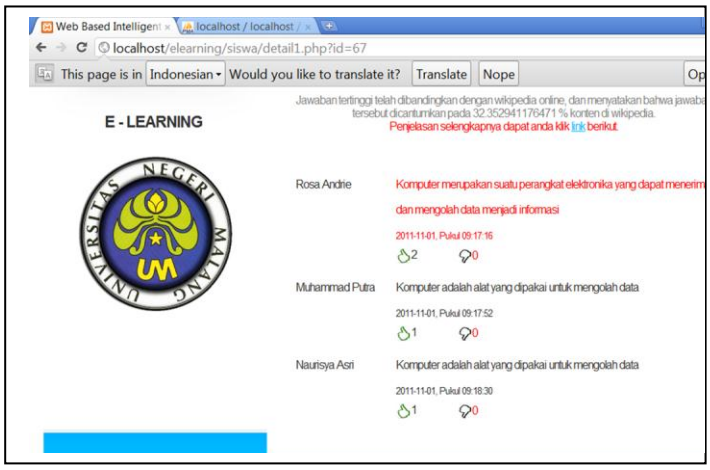
(b)

Figure 3. Student's interfaces question and get the answer

Beside asking a question, when student meeting some difficulties or having no difficulties, through the interfaces in figure 4, a student can see what problems other students have encountered in learning now and in the past and see the answers or solutions the teacher offered by browsing the knowledge base.



(a)



(b)

Figure 4. Browsing activities

IV. EXPERIMENT

According to the method described above and the structure of question and answering system, we build an experimental system in Indonesian Q&A system specifically in the Information Communication Technology Subject, for the Senior High School. We choose 40 questions, and there are two indexes in this experimental, precision and recall. The calculation formula is as follow:

$$precision = \frac{a}{a+c} \times 100\% \quad (2)$$

$$recall = \frac{a}{a+b} \times 100\% \quad (3)$$

In the formula, ^{a} is the number of right matching of questions; b is the number of without matching of question; c is the numbers of wrong matching questions. Through the experiment, we can get the data of precision and recall, the result is in Table 1.

TABLE 1. RESULT OF EXPERIMENT

Question	40
Answer	120
Right Matching	57
Without Matching	57
Wrong Matching	6
Precision	90,48 %
Recall	50 %

From the table, from the 40 question and 120 answer there are 57 answer are without matching because where the machine similarity get the keyword from the student question, there is not information about keyword question in the Wikipedia data based. In this case, the teacher will provide answers to students. After a question is replied by the teacher, not only the student will be notified, but also a new Q&A set will be formed and saved to the question answering knowledge base. We have this component for a reason that student's questions might be unanswered by the system due to no suitable or desirable answers for them in the knowledge base.

V. CONCLUSION

In this research, a system for online automatically answering students' questions in the collaborative learning environment has been designed. The system operated upon the question answering knowledge base. In the knowledge base, pairs of question with its corresponding answer (Q&A sets) were collected through the process of students asking questions and other students will response the question with answering and evaluating one another's ideas by vote. This representation is then compared with the representations of Wikipedia data base. A similarity percentage is given between the student answer and any existing Wikipedia data based. Based on the biggest vote some time is not describe the best answer, after time for answering and voting finish, similarity percentage will show for every answer. It aims to the students who asked the question and information for students who answer and vote can estimate approximately correct answer. Also, students can access the topic question more clearly through the link provided. It was very important to have such a system in collaborative learning environment. It benefited both the teacher and the students; students can look for the answers to their questions without the constraint of time and space.

REFERENCES

[1] Dillenbourg, P. (1999). Collaborative Learning: Cognitive and Computational Approaches. Advances in Learning and Instruction Series. New York, NY: Elsevier Science, Inc.

[2] Chiu, M. M. (2000). Group problem solving processes: Social interactions and individual actions. Journal for the Theory of Social Behavior, 30, 1, 27-50.600-631.

[3] Chiu, M. M. (2008).Flowing toward correct contributions during groups' mathematics problem solving: A statistical discourse analysis. Journal of the Learning Sciences, 17 (3), 415 – 463

[4] Mitnik, R., Recabarren, M., Nussbaum, M., & Soto, A. (2009). Collaborative Robotic Instruction: A Graph Teaching Experience. Computers & Education, 53(2), 330-342.

[5] Chiu, M. M. (2008). Effects of argumentation on group micro-creativity. Contemporary Educational Psychology, 33, 383 – 402.

[6] Chen, G., & Chiu, M. M. (2008). Online discussion processes. Computers and Education, 50, 678 – 692

[7] Wang, C.C., Hung J.C., Yang C.Y., Shih T.K. (2006). An Apllication of Question Answering System for Collaborative Learning. IEEE Conference on ICDCSW'06

[8] Johnson, D.W. and Johnson R.T. (1999) Cooperation and competition: Theory and research. Edina. MN: Interaction Book Company

[9] Slavin, R. (1996) Research on cooperative learning and achievement: what we know, what we need to know. Contemporary Educational Pshchology, 21, 1, pp. 43-69.

[10] Gilrory, K.(2001). Collaborative e-learning: the right approach ([Online].Available at : http://www.ottergroup.com/otter-with-comments/right_approach.html).

[11] Lave, J., and Wenger, E. (1991) Situated Learning: Legitimate Peripheral Participation. Cambridge University Press, Cambridge

[12] Kuutti, K. (1991) The concept of activity as a basic unit of analysis for CSCW research. Proceedings of the Second European Conference on Computer-Supported Co-operative Work: EC-CSCW'91 (eds. L.J. Bannon, M. Robinson &K. Schmidt) pp. 249-264, Kluwer, Dordrecht.

[13] Watzlawick, P. (1967) Pragmatics of Human Communications: A Study of Interactional Patterns. Pathologies and Paradoxes. W.W. Norton, New York.

[14] Bahreininejad A, Alinaghi Tanaz, 2011, A Multi Agent Question Answering System for E-Learning nad Collaborative Learning, nternational Journal of Distance Education Technologies, 9(2), 23-39, April-June 2011.

[15] Xu Jinzhong, Jia Keliang, Fu Jibin. 2008. Research of Automatic Answering System in Network Teaching. The 9th International Conference for Young Computer Scientist.

AUTHORS PROFILE

Kohei Arai received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 and also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada.

He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission A of ICSU/COSPAR since 2008. He wrote 26 books and published 227 journal papers.

Anik Nur Handayani received the B.E. degree in electronics engineering from Brawijaya University, and the M.S. degree in Electrical Engineering, from Institute of Technology Sepuluh Nopember, Surabaya, Indonesia, in 2004 and 2008, respectively. She is currently a PhD Student at Information Science in Saga University, Japan.

An Efficient Method For Multichannel Wireless Mesh Networks With Pulse Coupled Neural Network

S.Sobana
Assistant professor
Department of ECE
PSNA College of Engg & Tech-Dindigul 624622

S.Krishna Prabha
Associate professor
Department of Mathematics
PSNA College of Engg & Tech-Dindigul 624622

***Abstract*—Multi cast communication is a key technology for wireless mesh networks. Multicast provides efficient data distribution among a group of nodes, Generally sensor networks and MANETs uses multicast algorithms which are designed to be energy efficient and to achieve optimal route discovery among mobile nodes whereas wireless mesh networks needs to maximize throughput. Here we propose two multicast algorithms: The Level Channel Assignment (LCA) algorithm and the Multi-Channel Multicast (MCM) algorithm to improve the throughput for multichannel sand multi interface mesh networks. The algorithm builds efficient multicast trees by minimizing the number of relay nodes and total hop count distance of the trees. Shortest path computation is a classical combinatorial optimization problem. Neural networks have been used for processing path optimization problem. Pulse Coupled Neural Networks (PCNNS) suffer from high computational cast for very long paths we propose a new PCNN modal called dual source PCNN (DSPCNN) which can improve the computational efficiency two auto waves are produced by DSPCNN one comes from source neuron and other from goal neuron when the auto waves from these two sources meet the DSPCNN stops and then the shortest path is found by backtracking the two auto waves.**

Keywords-Wireless Mesh Networks; Multicast; Multichannel; Multiinterface; Shortest path; DSPCNN; Auto wave; Search space.

I. INTRODUCTION

Unlike mobile adhoc networks or wireless sensor networks route recovery are energy efficiency is not the major concern for mesh network due to limited mobility and the rechargeable characteristics of mesh nodes. Moreover supporting major applications such as video on demand poses a significant challenge for the limited bandwidth of WMNs it is necessary to design an effective multicast algorithm for mesh networks. It improves the system throughput by allowing simultaneous close-by transmissions with multichannel and multi – interfaces. It assigns all the available channels to the interfaces instead of just the non-overlapping channels.

We propose level channel assignment algorithm multichannel multicast algorithm to improve throughput for multichannel and multi interface mesh networks. Our design builds a new multicast backbone - tree mesh which partitions mesh network into different levels based on the Breadth First Search (BFS), and then heuristically assigns channel to different interfaces. The Pulse Coupled Neural Network is a very active neural network .The PCNN is modified so that the output pulses decay in times. These modified PCNN models need fewer neurons than other approach. This paper proposes a faster PCNN model, which can improve the computational efficiency significantly.

II. LEVEL CHANNEL ASSIGNMENT ALGORITHM

The nodes obtain their level information. The BFS is used to traverse the whole network. All the nodes are portioned into different levels according to the hop count distances between the source and the nodes.

If node a (in level i) and b (in level i+1) are within each other's' communication range, then 'a' is called the parent of 'b', and 'b' is called the child of 'a'.

We build a multicast tree based on the node level information. Initially, the source and all the receivers are included in the tree. Then, for each multireceiver v, if one of its parents is a tree node then connect it with that parent, and stop. Otherwise randomly choose one of its parents, say fv, as relay node on the tree, and connect v and fv. Afterwards, we try to find out the relay node for fv recursively. The process repeats until all the multireceivers are included in the multicast tree.

The tree nodes decide their channel assignment with the level information.

- The source node (level 0) only uses one interference, which is assigned channel 0. This interference is responsible for sending packets to the tree nodes in level 1.
- The internal tree node in level i ($i \geq 1$) uses two interfaces: one is assigned channel i-1, which is used to receive packets from the upper level; the other is assigned channel 1, which is used to forward the packets to the tree nodes at level i+1.
- The leaf in the level I ($i \geq 1$) uses two interfaces: one uses Channel i-1 to receive the packets from level i-1,the other uses channel I to forward the packets to the mesh clients within the communication range that desire to receive the packets.

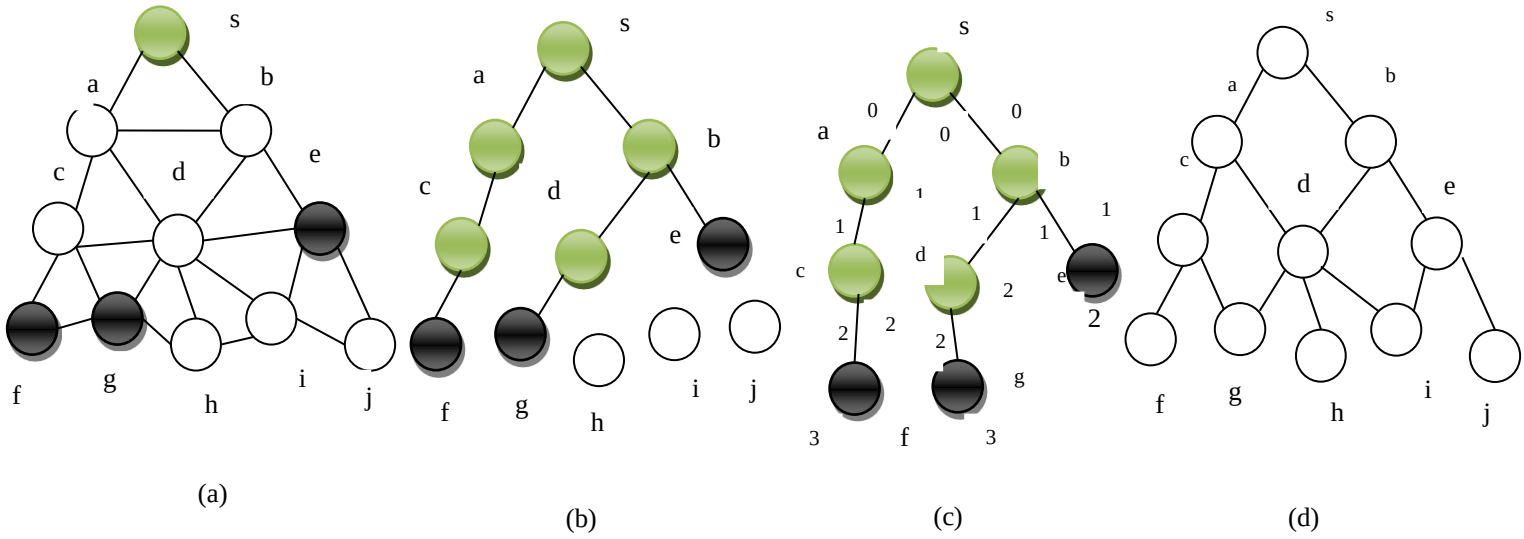


Fig.1.An example for LCA and tree mesh,(a)Network topology ,(b) multicast tree, (c) channel assignment, and (d) tree mesh

For example in fig.1, the node s is the source and nodes f, g, e are the multireceivers. In fig1.a {s, f, g, and e} are included in the multicast tree. Since nodes of g's parents are tree nodes, it randomly selects d as a parent node and connects node g with d. Then choose d's parent b as a tree node and connect d with b. Since b's parent s is a tree node connect b with s. Next, we start from multireceiver e. Connect e with its parent node b and stop because b is already connected with tree node s. Similarly the third multireceiver f, connect f with c, c with a and then a with s. Thus the tree construction is completed by connecting all the receivers with the tree.

The LCA algorithm has two advantages: simple implementation and throughput improvement. At the same time the use of multiple channels reduces the close-by interference and allows more simultaneous transmissions.

To improve the system throughput by the following ways, first LCA cannot diminish the reference among the same level s since it uses the same channel at the same level. Second, when the number of available channels is more than that of the levels, some channels will not be utilized, which is a waste of channel diversity. Third, the channel assignment does not take the overlap property of the two adjacent channels into account. For all I, channel and channel i+1 are adjacent in frequency, so they partially interfere with each other. Thus, the channel I for level i sti ll has some interference effect with the channel i+1 for level i +1.

III. MULTICHANNEL MULTICAST

A. Algorithm

To improve the system throughput, the MCM algorithm is proposed to minimize the number of relay nodes and the hop count distances between the source and the destinations, and further reduce the interference by exploiting all the partially overlapping channels instead of just the orthogonal channels.

B. Construction of multicast protocol

When all the Nodes are multireceivers, the multicast problem becomes the broadcast problem. We can say that the broadcast is a special case of multicast. The broadcast structure in the mesh network is built by the following steps.

After the BFS traversal, all the nodes are divided into different levels. Delete the edges between any two nodes of the same level, with which we get the elementary communication structure "tree mesh".Fig.1a and 1d given an example of the original network topology and its corresponding tree mesh.

Identify the minimal number of relay nodes that form the broadcast tree. Using more relay node means more transmissions in the network. Because the number of available channel is limited by current technical conditions, more transmissions would result in more interference and result in more bandwidth cost. Hence, minimizing the multicast tree size helps to improve the throughput. The purpose of this step is to identify the relay node for a node that has more than one parent nodes so that the number of relay node is minimal.

C. Structure of Multicast protocol

In broadcast structure unnecessary branches are present if the destinations do not involve all the nodes. Hence, we propose to construct a slim structure y using the MCM Tree Construction algorithm.

The goal of the algorithm is to discover the minimal number of relay nodes needed to construct a multicast tree. The search process starts from the bottom to the top. We use a simple example to explain the process in a tree mesh in Fig.3a, where nodes 6, 7, and 8 are the multireceivers. First select node 4 at level 2 because it covers all the multireceivers at level 3.Next select node 2 at level 1, which covers all the multi receivers and the relay node at level 2.By doing these steps finally we get the multicast tree in Fig.3b

D. Channel Assignment

Multi receivers can be connected with the gateway through mini al hop count distance as discussed earlier. Now we discuss about how to assign channels to the interference of tree nodes, for that we propose two allocation algorithms: ascending channel allocation and heuristic channel assignment.

a) Ascending channel Allocation

The interference that a node uses to receive packets from its relay node at the upper layer, called as Receiver Interference (RI), is disjoint from the interference the node uses to forward packets to children, called Send-Interference (SI).To guarantee that the relay node can communicate with its children, each node's RI is associated with the SI of its relay node, i.e., they should be assigned the same channel.

The algorithm is explained as follows: From the top to bottom in the tree, the channels are assigned to the interfaces in the ascending order until the maximum channel number is reached, then start from the channel 0 again. Although simple, this approach avoids the situation that the same channel is assigned to two nearby links that interfere with each other. In Fig .4, the numbers of orthogonal channels are three, the number above the node represents the channel number used for its RI, and the number below the node represents the channel number for its SI.

b) Heuristic Channel Assignment

We noticed that the interference range decreases with the increase of the channel separation for two wireless links which have short physical distance. To minimize the sum of the interference area of all the transmissions this algorithm is proposed.

We use IR (uv) to indicate the interference range of sender u of one link with respect to sender v of another link According to our consideration all the have the same transmission range R, IR (u v) =R* δ [iu-iv].iu and iv are the channels of u and v for their SIs, and δ t is the interference factor. When allocating a channel for relay node u, the channel assignment should take a channel that minimize the sum of the square of the IRs between u and u's neighboring relay nodes, i.e., minimize $\sum_{v \in N(u)} IR^2(u_v)$,where N(u) is the set of neighboring relay nodes of u. This is because the bigger the interference area means bigger chance two transmissions may interfere. The interference area is approximated as a circle whose area is determined by IR2 (uv).Since

$$\sum_{v \in N(u)} IR^2(u_v) = \sum_{v \in N(u)} \left(R * \delta_{|i_u - i_v|}\right)^2 ,$$

The heuristic Channel assignment is used to minimize

$$\sum_{v \in N(u)} \left(R * \delta_{|i_u - i_v|}\right)^2$$

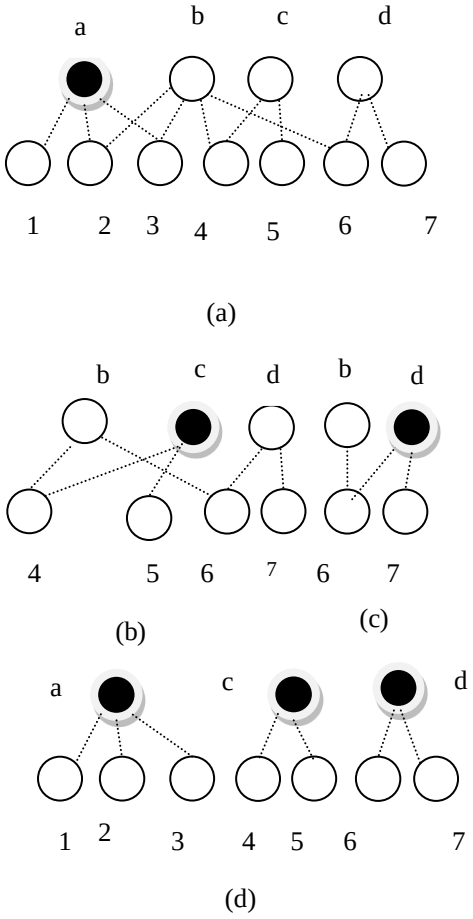


Figure 2. Relay node search example

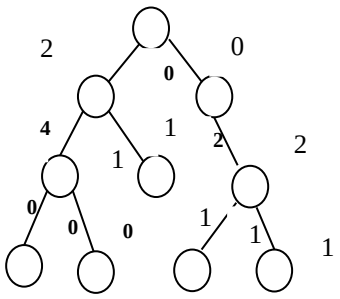


Fig.4. Ascending channel allocation example

IV. DSPCNN MODEL

A. Preliminaries

The input to the preprocessing stage is an undirected graph G = (V, E) with n vertices and m edges, and non-negative lengths l(e) for each edge e. Another two inputs are a source node s and a goal node g. The goal of this algorithm is to find a shortest path from s to g. Let dist(s, g) denote the shortest path length from s to g with respect to l. i.e. dist(s, g) =dist (g, s).

B. Model Of DSPCNN

To compute the point-point shortest path more efficiently a Dual Source Pulse Coupled Neural Network (DSPCNN) model is proposed. This model can produce two Auto waves from two different firing sources. At t=t0, the source neuron and goal neuron fire and emit pulses simultaneously. Then, the two Auto waves propagate in parallel by their neighboring neurons at next instant till they meet together. In order to differentiate two auto waves, the auto wave propagating from source neuron is denoted as Ps, and the auto wave propagating from goal is denoted as Pg. If neuron fires on the simulating of Ps auto wave it outputs Ps pulses. If a neuron fires on the simulating of Pg auto wave, it output Pg pulses. If a neuron fires on the simulating of both Ps and Pg pulses, it indicates that the two auto waves meet and the model should stop.

C. Shortest Path Computation Using DSPCNN

To compute the shortest path for networks, first they are all transformed to a graph for further processing. The next step is to map the graph into DSPCNN model. Each node in the graph corresponds to a DSPCNN neuron, and each edge associated with a link between neurons. The cost of an edge can be viewed as an external input for the two neurons connected by the edge.

During time t=0, Source neuron and Goal neuron fire simultaneously. Then the auto waves Ps and Pg from the firing sources propagate to their neighbors. A variant meeting is used to determine whether the two auto waves meet together, and the meeting neutron is denoted by Nm. If Nj fires on the simulation of Ni, we call Ni is the precursor of neuron Nj.

V. CONCLUSION

In our paper we investigate the multicast algorithm wireless networks. In order to achieve efficient multicast in WMNs, two multicast algorithms are proposed by using multichannel and multi interfaces.

These algorithms are focused on increasing the throughput and decreasing delay. With neural networks the proposed DSPCNN is used to achieve higher efficiency and involve lower search space, which can save the run time significantly.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their helpful comments. We would like to thank our mother for her kind support. We would like to thank our institution and management for their cooperation.

REFERENCES

[1] J.So and N.Vaidya, 2004.Multichannel Mac for Ad Hoc Networks: Handling Multi-channel Hidden Terminals Using a single transceiver.

[2] K.Ramachandran, E.M.Belding, K.Almeroth,and M.Buddhiko, interference- Aware channel Assignment in multiradio wireless messnetworks"Proc.IEEE INFOCOM,2006.

[3] http://www.seattlewireless,net,2009.

[4] A.Mishra,V.Shrivatsava,and S.Banarjee,"Partialy Overlapped channels Not Considred Harmful,"

[5] E.W.Dijkstra .A note on two prolems in connection with aphs.Numeische Mathematick.1959,(1):269-271

[6] R.CM.Folyd.Algorthms 97;shortest path Communications of the ACM.1962,6(5):345

[7] H.Qu.Z.Yi.A new algorithm for finding the shortest paths using PCNN s.Chaos, solutions and fractals.2007,4(33):1220-1229

[8] T.H.Cormen,C.E.Leiserson,R.L.Rivest "Introduction to Algorithms.The MIT press,2001.

[9] B.Yu,L.Zhang.pulse coupled Neural Networks for contour and motion matchings.IEEE Transanctions on Newral networks. 2004, 5(15):1186-1201

[10] Guokai Zeng, Bo Wang, Yong Ding, Li Xiao, Matt Mutka, Multicast Algorithms for Multi-Channel Wireless Mesh Networks

AUTHORS PROFILE

S.SOBANA.-received the B.E (with distinction) in Electronics andCommunication Engineering and M.E (with distinction) in Applied Electronics from RV.S College of Engineering and Technology, Anna University Chennai, Tamil Nadu, India .in 2005 and 2007 respectively. Currently working as an assistant professor in Electronics and Communication Engineering Department, P.S.N.A College of Engineering and Technology, Tamil Nadu, India. Her research mainly focuses on ad hoc networks, congestion control, power management techniques in wireless networks.

S.Krishna Prabha-receiver her B.Sc Mathematics degree in 2000 from G.T.N Arts College, MK University, Tamil Nadu, India. MSc and MPhil degree in Mathematics from M.K. University, Tamil Nadu, India in 2002 and 2004 respectively. Currently she is pursuing the M.E degree from System Engineering and Operation Research department, Anna University, Trichy, Tamil Nadu, India. Currently working as an associate professor in Department of Mathematics, P.S.N.A College of Engineering and Technology, Tamil Nadu, India. Her research interests include graph theory, operation research, boundary value problems, ad hoc networks.

A Congestion Avoidance Approach in Jumbo Frame-enabled IP Network

Aos Anas Mulahuwaish, Kamalrulnizam Abu Bakar, Kayhan Zrar Ghafoor
Faculty of Computer Science and Information System
Universiti Teknologi Malaysia
Johor, Malaysia

Abstract— Jumbo frame is as an approach that allows for higher utilization of larger packet sizes on a domain-wise basis, decreasing the number of packets processed by core routers while not having any adverse impact on the link utilization of fairness. The major problem faced by jumbo frame networks is packet loss during queue congestion inside routers is as the RED mechanism that is recommended to combine with jumbo frame treats jumbo frame encapsulation as one packet by drop the whole jumbo frame with packets encapsulate during the congestion time. RED dropping the whole jumbo frame encapsulation randomly from head, middle and tail inside queue of router during periods of router congestion, leading to affect the scalability and performance of the networks by decreasing throughputs and increasing queue delay. This work proposes the use of two AQM techniques with jumbo frame, modified Random Early Detection MRED and developed drop Front technique DDF, which are used with the jumbo frame network to reduce packet drop and increase throughput by decreasing overhead in the network. For the purpose of evaluation, network simulator NS-2.28 was set up together with jumbo frame and AQM scenarios. Moreover, for justification objectives, the proposed algorithm and technique for AQM with jumbo frame were compared against the existing AQM algorithm and techniques that are found in the literature using metrics such as packet drop and throughput.

Keywords- Jumbo Frame; Queue Congestion; AQM; RED.

I. INTRODUCTION

Computer networks have experienced rapid growth over the years, from transferring simple email messages to now being a full media resource where full length movies are commonly transmitted. Many users have begun to use the internet for many things; as a result, the connections of internet have started to become strained where before the common solution of the internet service provider (ISP) was capable of providing sufficient bandwidth to users in the network. However, recent research has found that the users' access speed has increased and thus affects the efficiency of the network. Therefore new techniques need to improve the efficiency of the network traffic. Many techniques from multicasting to packet caching have been used to improve the efficiency of the network, but with limited success as these techniques suffer from one or more drawbacks including global network support, application support, asymmetric and computation overhead. The current assumption with networking research is also that it affects an individual network flow's quality of service (QoS) including the packet loss, end to end delay and jitter; however these researches presented techniques that investigate the possibility of trading a minimal amount of an individual flow's QoS typical delay so as to obtain better overall network performance [1].

One of the issues facing networks is the number of packets required to be processed per second, whereby the gigabit link core router may have to route anywhere from 90,000 to 2,000,000 packets per second. As line speed increases to greater rates, so does the number of packets that need to be processed; one way to reduce the load on the router is to increase the maximum transmission unit (MTU) of the network. Unfortunately while the MTU of Ethernet is 1500 bytes, up to 50% of the packets transferred across the network are 64 bytes or less.

Jumbo frame is a technique that aims to reduce the number of packets processed by the core routers, by reducing the number of packets. This is accomplished by transmitting many packets in the domain into a single large jumbo frame for transmission across the core network. In ordering the aggregate packets together in a jumbo frame, incoming packets are queued briefly by egress point. Once the jumbo frame reaches the egress of the domain, the original packets are rebuilt and transmitted on toward their final destination.

II. RELATED WORK

A jumbo frame has a common size of 9000 bytes, which is exactly six times the size of a standard Ethernet frame [5]. A 9k byte jumbo frame would be 9014-9022 bytes together with the Ethernet headers. This makes it large enough to encapsulate a standard network file system (NFS) data block of 8192 bytes, yet not large enough to exceed the 12,000 byte limit of Ethernet's error checking in cyclic redundancy check algorithm (CRC) [5].Undoubtedly, smaller frames usually mean more CPU interruptions and more processing overhead for a given data transfer size [9]. When a sender sends a data, every data unit plus its headers have to be processed and read by the network components between the sender and the receiver. The receiver then reads the frame and TCP/IP headers before processing the data. This whole process, plus that of adding the header to frames and packets from the sender to the receiver consumes CPU cycles and bandwidth [13]. For these reasons, increasing the frame size by sending data in jumbo frames means fewer frames are sent across the network when considering the fact of high processing cost of network packets [3]. These generate improvements in CPU utilization and bandwidth by allowing the system to concentrate on the data in the frames, instead of the frames around the data. The justification behind increasing the frame size is clear; larger frames reduce the number of packets to be processed per

second. A single 9k Jumbo Frame replaces six 1.5k standard frames, producing a net reduction of five frames as only one TCP/IP header and Ethernet header is required instead of six, resulting in 290 (5*(40+18)) fewer bytes transmitted over the network [14].

In terms of improving bandwidth, it takes over 80,000 standard Ethernet frames per second to fill a gigabit Ethernet pipe, which in turn consumes a lot of CPU cycles and overhead. By sending the same data with 9,000 bytes jumbo frames, only 14,000 frames need to be generated and the reduction in header bytes frees up 4 Mbps of bandwidth. The resources used by the server to handle network traffic are proportional to the number of frames it receives. Therefore, using fewer large frames dramatically improves server and application performance, compared to a larger number of smaller frames [14]. Jumbo frame improves core router scalability, by encapsulating packets with the same next autonomous systems (AS) and egress point into larger packets for transmission across the domain. Critically, the design of jumbo frame functions on a domain-wise scale, instead of end-to-end; the external entities (other domains and end hosts) are unaware that any conversion took place. The overall jumbo frame shown in Figure 1:

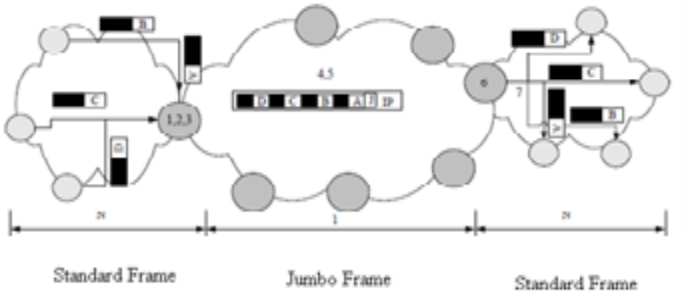


Figure 1. Jumbo Frame Structure

The description of the structure as shown in Figure 1 is that when packets arrive at an ingress node to the domain, the ingress node and the packets are sorted into queues based upon their egress point of the network in their path that is obtained from the border gateway protocol (BGP) routing table [12]. A jumbo frame Encapsulation Timer (JFET) is started for the queue. Packets that are being sent through the same egress point are combined into the same jumbo frame, subject to MTU; once the JFET for the queue has expired, the Jumbo Frame is released towards the next AS. The jumbo frame is routed through the core of the network, with the routing provided by using the standard routing mechanism of the network. The jumbo frame arrives at the egress node and the original packets are separated out, after which the original packets are forwarded onto their final destinations. There are two main benefits of using jumbo frame [1]. The first benefit is that jumbo frame lowers the number of packets that the core routers are responsible for processing, thus allowing better scaling for the network as line speeds increase. The second beneficial aspect of jumbo frame is that data is more efficiently transferred by reducing the number of physical layer headers used (due to a lower number of packets).

A. *Fast Packet Encapsulation*

The jumbo frame is structured to allow for efficient encapsulation, inspection, and de-capsulation [5]; packet overhead is minimal and is offset by the reduction in physical layer headers. The structure of the jumbo frame is shown in Figure 2 containing the following fields:

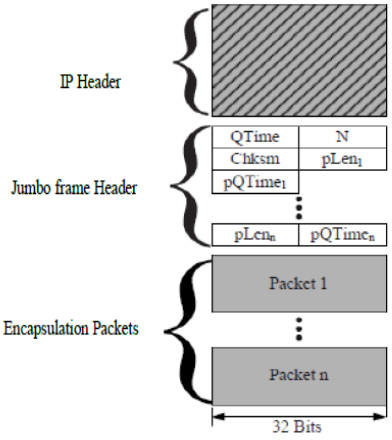


Figure 2. Jumbo Frame Structure

The destination address of the jumbo frame is the same as the first packet stored in the jumbo frame. For a multiprotocol label switching (MPLS) network, the destination address is the MPLS address of the first packet stored in the group. This ensures proper routing for all packets as all encapsulated packets contained in the jumbo frame would arrive at the next correct AS in their path. The design of the jumbo frame allows the original packets to be de-capsulated with minimal effort while also keeping the overhead of the jumbo frame to a minimum. As shown, the overhead of the jumbo frame is 6 + 4N bytes. However, the overhead is offset by the reduction in physical layer headers. The net cost (or benefit) of jumbo frame can be stated as:

Equation 1: Cost = HIP + HJG + (N _ 4) − (Hp _ (N − 1)) (1)

The cost of jumbo frame in the above equation comes from the size of the IP header (HIP), the jumbo frame header (HJG), and the number of encapsulated packets (N). The reduction in bandwidth comes from the reduction in physical headers 5 (Hp). For example, if the network is an Ethernet network and two packets were encapsulated into a jumbo frame, then HIP = 20, HJG = 6, N = 2, Hp = 38, and the total cost would be −4 bytes. In other words, 4 bytes of bandwidth would be conserved.

B. *Egress Shaping*

When the jumbo frame reaches its destination, the packets need to be de-capsulated and released to the next node on their path to the destination. If all the packets are released as soon as they are removed from the jumbo frame, this can lead to dropped packets at the client due to the receive buffer overflowing [2]. Hence packets are shaped at the egress

according to the differences in their arrival time (pQTime). In other words, if two packets arrive at the ingress node 4 ms apart, they are released from the egress node 4 ms apart.

C. Active Queue Management (AQM) with Jumbo Frame

The structure of the jumbo frame allows active queue management AQM scheme techniques and methods are an important type of technology with aims to improve the utilization of the network [4] and [8]. While jumbo frame can be combined with AQM techniques and methods, this allows the combination between the jumbo frame and AQM techniques and methods to solve many problems in jumbo frame networks, and also enhance the efficiency and scalability of jumbo frame network by decreasing the packet loss and end to end delay, reducing the overhead and increasing throughput for jumbo frame networks to perform optimally, RED is one of the AQM methods that work with jumbo frame [4] and [8], for preventing the gateway router from becoming full and ensuring that jumbo frame can transmit to the destinations.

In [6] and [11], two different methods that RED queues can use to determine the queue utilization are presented. The first is through the number of packets in the queue and the second is to determine queue utilization by number of bytes in the queue. RED detect the congestion in jumbo frame networks, and decrease the congestion of overflow by randomly drop whole jumbo frame, RED treat jumbo frame as a one big packet, so when the drop occur RED will used the same drop operation with standard packet.

D. Random Early Detection (RED) and Drop from Front

Random early detection (RED) Algorithm was first proposed by [6]. This discipline maintains a moving average of the queue length to manage congestion. If this moving average of the queue length lies between a minimum threshold value and a maximum threshold value, then the packet is either marked or dropped with a probability. If the moving average of the queue length is greater than or equal to the maximum threshold then the packet is dropped. Even though it tries to avoid global synchronization and has the ability to accommodate transient bursts, in order to be efficient RED must have sufficient buffer spaces and must be correctly parameterized. In contrast, RED algorithm uses packet loss and link utilization to manage congestion. RED gateways can be useful in gateways with a range of packet-scheduling and packet-dropping algorithms. For example, RED congestion control mechanisms could be implemented in gateways with drop preference, where packets are marked as either essential or optional, and optional packets are dropped first when the queue exceeds a certain size. Similarly, for the example of a gateway with separate queues for real time and non-real time traffic, RED congestion control mechanisms could be applied to the queue for one of these traffic classes.

The RED congestion control mechanisms monitor the average queue size for each output queue, and by using randomization chooses connections to notify of that congestion. Transient congestion is accommodated by a temporary increase in the queue. Longer-lived congestion is reflected by an increase in the computed average queue size, and results in randomized feedback to some of the connections to decrease their windows. The probability that a connection is notified of congestion is proportional to that connection's share of the throughput through the gateway. In addition, gateways detecting congestion before the queue overflows are not limited to packet drops as the method for notifying connections of congestion. RED gateways can mark a packet by dropping it at the gateway or by setting a bit in the packet header, depending on the transport protocol. When the average queue size exceeds a maximum threshold, the RED gateway marks every packet that arrives at the gateway. If RED gateways mark packets by dropping them, rather than by setting a bit in the packet header, then the RED gateway controls the average queue size even in the absence of a cooperating transport protocol when the average queue size exceeds the maximum threshold. One advantage of a gateway congestion control mechanism is that it works with current transport protocols and does not require that all gateways in the internet use the same gateway congestion control mechanism; instead it could be deployed gradually in the current Internet. RED gateways are a simple mechanism for congestion avoidance that could be implemented gradually in current TCP/IP networks with no changes to transport protocols.

Drop from front technique drops the head of the queue if the incoming packet sees the queue as full. With the drop from front policy that governs when a packet arrives to a full buffer, the arriving packet is allowed in, with space being created by discarding the packet at the front of the buffer. This shows that for networks using TCP, the Internet transport protocol, a drop from front policy results in better performance than is the case under tail dropping and its variations [10]. Drop from Front a partial solution to the problem of throughput collapse in networks where TCP represents a sizeable part of the load. Drop from Front can be used in conjunction with other strategies such as Partial Packet Discard. In [10], showed that moving to a drop from front strategy considerably improves performance and allows use of smaller buffers than is possible with tail drop. Drop from Front is also applicable to both the switch and routers. During congestion episodes when buffers are full, Drop from Front causes the destination to see missing packets in its received stream approximately one buffer drain time earlier than would be the case under tail drop. The sources correspondingly receive earlier duplicate acknowledgements, causing earlier reduction in window sizes.

However, drop from front has the advantage that the switch and router does not need to maintain a table of drop probabilities and does not have to know the traffic type being carried. This is because drop from front also reduces latencies for successfully transmitted packets and hence is a sensible policy to use for delaying sensitive non-feedback controlled traffic as well. This reduction in latency has been described by [15], who considered a "drop from front" scheme for a very different problem where none of the sources were feedback controlled. They found that drop from Front resulted in shorter average delay in the buffer for eventually transmitted packets and recommended its use for time-constrained traffic.

III. Methodology and Results

Modified random early detection (MRED): a RED queue is an important technique that aims to improve the utilisation of the network and remove the synchronisation that tends to occur

with TCP flows when the network becomes congested. There are two different methods that RED queues can use to determine the queue utilisation. The first method is to simply use the number of packets, while the second is to use the number of bytes consumed in order to determine queue utilisation. The second method has more overhead, however, it allows for smaller packets to be favoured over larger packets. This effectively gives priority (less chance to be dropped) to smaller packets (eg.TCP acknowledgments). In jumbo frame networks if RED is not modified in any way, jumbo frame will be treated the same as any other packet. This behaviour is not advantageous as a jumbo frame has the same percent chance to be dropped as does any other packet. However, any time a jumbo frame is dropped, all encapsulated packets are lost. Because multiple packets are lost, this can result in poor TCP performance, as packets from the flow can be dropped, thus resulting in a greater than desired reduction of traffic.

MRED will start to calculate the new average queue size and the time for the new flows that coming to the queue and MRED will do compression between the number of jumbo frames and the capacity of the queue and check if the capacity of queue are enough to receive new flows or not. If the queue has enough space for all flows then MRED will allow all jumbo frames to queue up for forwarding out to the different destinations. However if the capacity is not enough, there is a congestion over flow problem that will happen in this queue, all that will be calculated based on the MRED detection mechanism. In this case MRED will do the calculation for each jumbo frame for probabilities drop. From here MRED will check the header of jumbo frame and will exactly check and compare two of fields inside the header. It will check the average length of each jumbo frame to calculate out the percentage of jumbo frame packets to distinguish that jumbo frame is not like any normal packet (this is because the average length size is high), MRED will also check the number of packets which encapsulated within the field header to verify there are encapsulation packets inside. Here MRED will only work with the average length and the number of packets that encapsulated within jumbo frame and will not work with the header of capsule frame.

After that, MRED will register out all of the information from the header for each jumbo frame encapsulation; then based on the percentage of packets that have been encapsulated, MRED will mark jumbo frame for drop sub packets inside, the percentage of packets that will drop are different from jumbo frame to another that are in the same flows. This calculation is based on the percentage of upper and lower bounds for each jumbo frame with the packets encapsulate, in which this calculation based on specific mathematical formulas. MRED will compare the percentage of packets inside each of jumbo frame with average queue size for the queue of router. Then MRED will decide the percentage of packets dropped from each of jumbo frame, to make the average queue size stable between them and during the congestion overflow time, and to reduce losing the whole encapsulation of jumbo frame but for only subs of packets. The marking operation of MRED for jumbo frame and the packets inside are related with time that sets for each jumbo frame, after that the probability marking drop will be set.

In this work MRED are combined with DDF, so based on this mechanism it will only mark the jumbo frame at the head of queue and the packets at the head of those jumbo frames. MRED will distribute the drop marking operation with the different jumbo frames to reduce the congestion and to let some of the packets inside those encapsulations left within without dropping it whole. This mechanism will reduce losing the whole packets inside each jumbo frame at the same flow; Figure 3 shows the diagram of MRED operation structure.

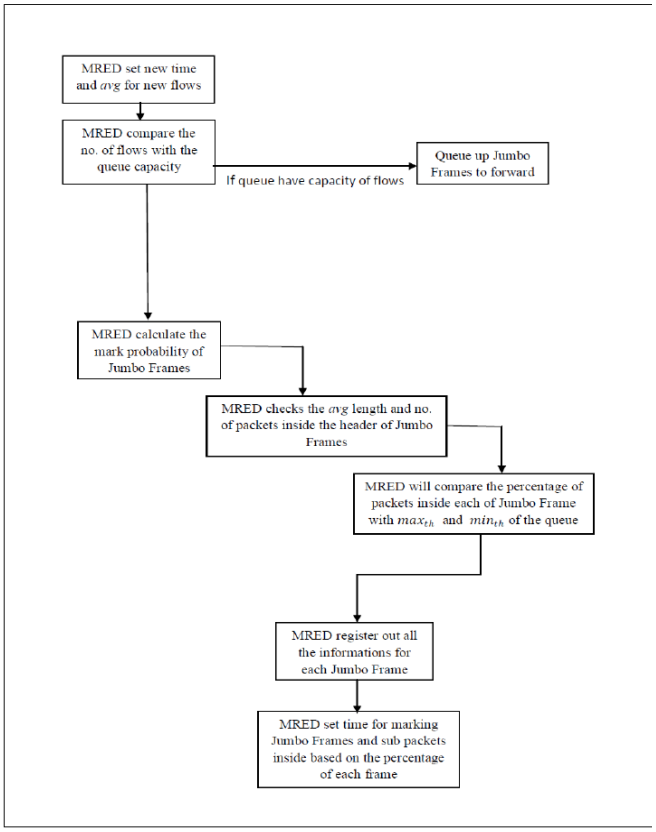


Figure 3. MRED operation structure

Developed drop front (DDF): development drop front mechanism is combined with modified RED for the steps of the packets drop in jumbo frame networks. After MRED has marked the jumbo frame that needs to be drop inside it, by calculating the upper and lower bounds for the encapsulations based on the percentage of jumbo frame encapsulations. When the MRED marked jumbo frames for dropping process, only the sub packets inside the jumbo frame will be dropped; the marked sub packets inside jumbo frame encapsulation will be done in the head of encapsulation frame, based on the mechanism of DDF which combines with MRED, so there are no random packets dropped inside jumbo frame. DDF will wait until the processing time finishes for the MRED with all the flow packets, then the time for DDF operation will start; DDF checks how many encapsulated for jumbo frame that marked by MRED, based on the percentage for each jumbo frame inside the queue. After checking the numbers of marked jumbo frames, DDF calculates the sub packets that are marked for drop by MRED inside each marked encapsulation.

DDF will set new time differently with time that was set before by MRED for each encapsulate frame that has been marked by MRED to do drops operation. Each marked jumbo frame has its own time drop packets. This time is set based on how many packets that are marked to drop each sub packets' time that have been marked to drop for this operation. There is a delay time for packet drop from one packet to another and this time will be calculated and set for the total drop operation time for the whole jumbo frame and each jumbo frame have different time with other. The drop operation starts with the first jumbo frame in the head of queue that was marked for drop operation. Inside this marked encapsulation sub packets drop operation will start with the first packet in the jumbo frame encapsulation that has been marked to drop. The DDF operation will start to drop packets by packets inside each jumbo frame encapsulation, and the packets drop will set in sequence number of router queue for each jumbo frames. Then it will send notification to the source for retransmitting the loss packets. In this operation, not all the encapsulation of jumbo frame is lost and the drop operation for the sub packets did not happen randomly but only from the front of jumbo frame, Figure 4 shows the DDF operation.

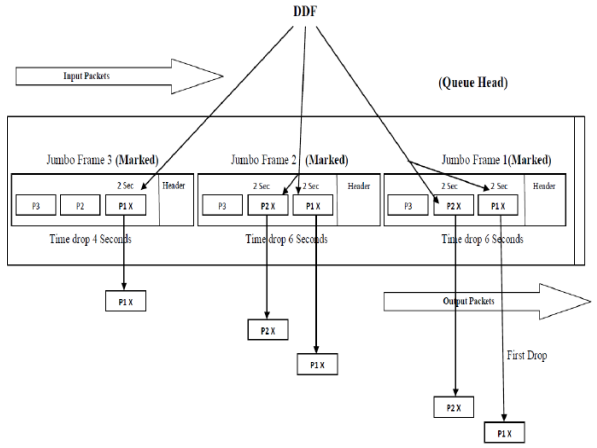


Figure 4. DDF operation

DDF allows the possibility of dropping partial packets without significant overhead. Firstly DDF looks at the number of packets stored in the jumbo frame encapsulations. Once the number of packets to be dropped is decided, the packets will removed from the head of the jumbo frame. The length of jumbo frame is shortened by the lengths of the packets that are to be removed and their lengths in the jumbo frame header are set to zero. The number of packets field for each jumbo frame got marked to drop sub packets will not be modified, and also the average length field in header will not be modified. This is due to the need for correct parsing at the egress router and the need for simplicity in modifying the packets in flight. Removing the lengths that are zeroed out is not a desirable option because multiple memory copies would have occur before the packet could be forwarded. So here jumbo frame will forward out without restructuring the sequencing of packets that were encapsulated, only the number of packets and average length fields in the jumbo frame header are not modified, DDF will set zero at jumbo frame header instead each packet has been dropped directly and one by one based on the time has been set for each jumbo frame marked and for each packets inside need to be dropped to remove the restructure operation, Figure 5 shows the average length of packets after drop operation inside jumbo frame header.. After that jumbo frame will de-encapsulate the rest packets to the destination address by the egress operation. DDF eliminates the random marked jumbo frame and dropped the packets inside encapsulation. However, if the packets are able to be removed randomly by MRED in jumbo frame, the complexity of the partial drop would substantially increase. The increase in complexity is from performing an MRED calculation on each encapsulated and from memory move operations needed to close the gaps in the jumbo frame after drop sub packets in different places in the encapsulation. DDF eliminates the restructure operation for each jumbo frame; all that will decrease the overhead in jumbo frame networks.

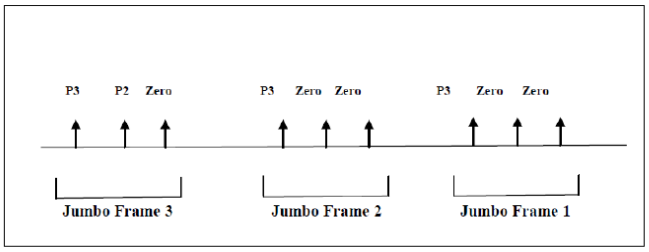


Figure 5. The average length of packets inside jumbo frame header after packets drop

A. *Simulation Setup*

The simulations presented here illustrate MRED with DDF well-understood dynamic of the average queue size varying with the congestion level, resulting from MRED with DDF and normal RED with tail drop fixed mapping from the average queue size to the packet dropping probability and the percentage of throughput. These simulations focus on the transition period from one level of congestion to another.

These simulations used a simple dumbbell topology with 6 nodes, the congested link of 1.5Mbps. The buffer accommodates 20 packets, which, for 3000 byte packet size and MTU 3000 byte, corresponds to a queuing delay of 0.28 seconds. In all of the simulations, weight of queue is set as a default in NS-2 to 0.0027, the choice of Wq determines the queue weight of the averaging for the average queue size, if Wq is too law, then the estimated average queue size is probably responding too slowly to transient congestion, if □□□□ it too high, then the estimated average queue size is too closely tracks the instantaneous queue size, MINth is set to 5 packets, the setting for MINth depends on exactly what the desired tradeoffs is at that router between low average delay and high link utilization. In the NS-2 MINth is set to a default of 5 packets because if MINth is set as small as one or two packets would only denied burstiness in the arrival process, and MAXth is set to 15 packets; there times more than MINth. Maximum value for the current marking of packet probability MAXp is constrained to remain within the range [0.01, 0.5] (or equivalently, [1%, 50%]), and the percentage of Jumbo Frame packets is 0.025, the average size of encapsulated packet is read from the Jumbo Frame header, not calculated at the router.

B. Simulation Scenario

The first scenario is for the increased average queue size in congestion, which used for testing the proposed MRED with DDF and also for test normal RED with tail drop in jumbo frame networks, this scenario is focus for the increase the average queue size in router queue during the congestion over flows at the transition period. The new flows are more than the buffer size capacity, the over flows burst in the specific simulation time, the average queue size has been increased because this over flows and been near or over the MAXth, so the congestion and packet drop happened, with decrease in throughput. This simulation is test the efficiency and scalability for the proposed MRED with DDF algorithm and compare the results with normal RED with tail drop results, for reduce the packet loss in and increase throughput with jumbo frames.

C. Results for MRED with DDF an Increased in Congestion Scenario

For this simulation scenario, the forward traffic consists of two long-lived TCP flows, and the reverse traffic consists of one long-lived TCP flow. At 25 seconds time, there are 20 new flows started, one every 0.1 seconds, each with a maximum window of 20 jumbo frames. This is not intended to model a realistic load, but simply to illustrate the effect of a sharp with the average queue size changing as a function of the packet drop rate. However after roughly 10 seconds, and because of the new 20 flows of jumbo frames the congestion happened, the algorithm of MRED detected the congestion and started to calculate the average queue size in the overflow time, MRED marked packets inside jumbo frames by put the drop probability first, and then mark sub packets inside jumbo frames at the head of queue and at the head of jumbo frames to decrease the congestion and then the drop will be done at the head of those jumbo frames by DDF without changing the length of information inside the header for each jumbo frames marked for drop. Here MRED with DDF has brought the average queue size back down to the range, between (6 – 7 packets). That means the proposed algorithm makes the average queue size away from the MAXth by making the probability of the packet drop less (MINth ≤ avg < MAXth).

The simulations with MRED with DDF have a higher throughput with smaller packet loss (drop), at the first half part of simulation, the throughput percentage is 42.45% and the packet drop is 0.69%. In the end of simulation scenario, the throughput becomes 91.7% and packet drop 8.24%. Figure 6 shows the MRED with DDF an increase in average queue size in congestion, the green trend represents the instantaneous change of queue length and the red trend shows the average queue size.

D. Result for Normal RED with Tail Drop an Increased in Congestion Scenrio

For this scenario simulation, it used the same simulation with MRED and DDF but with normal RED and tail drop instead. There are also at 25 seconds time where there are 20 new flows start, one for every 0.1 seconds, and each with a maximum window of 20 jumbo frames. In Figure 7 the graph illustrates normal RED with tail drop, with the average queue size changing as a function of the packet drop rate. With 20 new jumbo frames flow, congestion happened and packet dropped, because RED detected congestion and the RED algorithm dropped marked jumbo frames totally by tail drop at the tail of queue only. The packet drop rate changes from 0.90% with throughput 41.06% over the first half of the simulation, to 8.50% with the throughput 90.20% over the second half of simulation That means the average queue size here is become near to MAXth because of the algorithm for normal RED with drop tail did not reduce the number of packets that dropped during the congestion time. Due for that reason, the average queue size has been increased and the throughput has been decreased. Figure 7 shows the normal RED with tail drop with an increase in congestion, here can be noticed that at 25 second during the congestion the trend of average queue size increases and almost near with MAXth which means more packet drop happened.

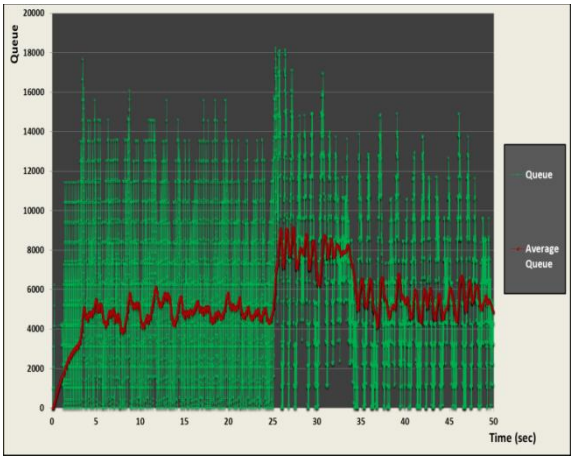


Figure 6. MRED with DDF with an increase avg in congestion

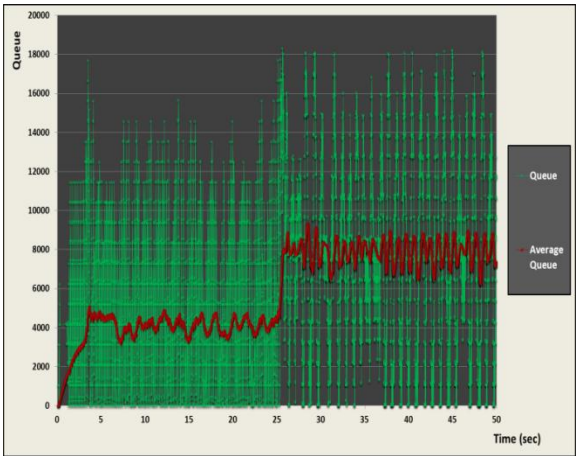


Figure 7. Normal RED with tail drop with an increase avg in congestion

E. Results Comparision

Four scenarios were compared in this study, starting from results for MRED with DDF with an increase of average queue size in congestion compared with results for normal RED with tail drop with an increase in congestion too; results for MRED with DDF with a decrease of average queue size in congestion compared with results for normal RED with tail drop with a decrease in congestion also. All those comparisons are based on the simulation metrics packet drop and throughput.

It can be observe in Figure 8 and 9 the comparison between the results for MRED with DDF and normal RED with tail drop in the same scenario with an increased of average queue size during the congestion. It has shown at the end of simulation lower percentage packet drop decrement 26% when used RED with DDF than in normal RED with tail drop, and throughput increment 1.56% when used with MRED with DDF than in normal RED with tail drop; it can be observed that when there are over flow in the queue the MRED with DDF makes the average queue size lower than MAXth by decreasing drop of jumbo frame encapsulation and just drop packets inside jumbo frame encapsulation during over flow in queue and increases the throughput. This means the proposed MRED with DDF technique achieved the objectives for decreasing the packet drop and increases the throughput with jumbo frame, which will be led to enhance the scalability and efficiency of jumbo frame networks.

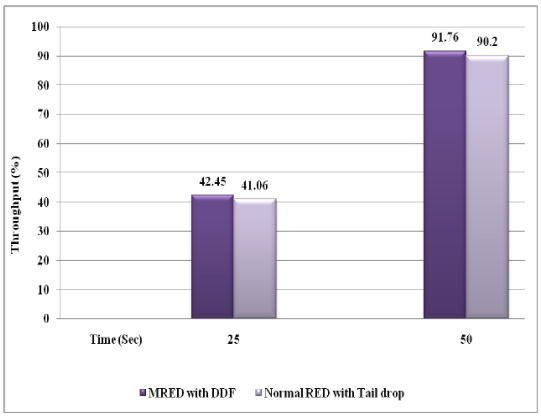


Figure 8. Packet drop rate between MRED with DDF and RED with tail drop in increase of congestion

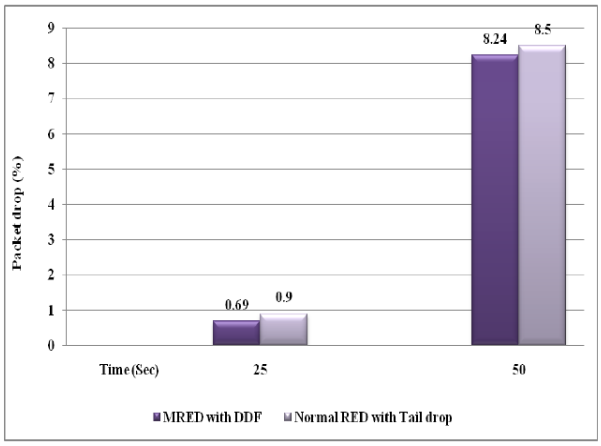


Figure 9. Throughput rate between MRED with DDF and RED wih tail drop in increase of congestion

IV. CONCLUSION

This work has been proposed new scheme in AQM with jumbo frame networks, by combined modified random early detection MRED with developed drop front DDF. The proposed algorithm help to reduce the packet loss in jumbo frame networks, and increase the throughput, by reduce the overhead and enhance the scalability and efficiency for jumbo frame networks. The proposed algorithm has been implemented by NS2 simulator, it have achieved the best results for reducing the packet loss at queue and increase throughput in jumbo frame environments when it compared with a result for applying the normal RED combined with drop tail technique in jumbo frame environments with the same metrics.

REFERENCES

[1] Alteon (1999) "Extended Frame Sizes for Next Generatinon Ethernets", White paper, Lightwave technology journal, pages 66 -73.

[2] Balakrishnan, H. and Padmanabhan, V. N.,Seshan, S.,Stemm, S. and Katz, R. H. (1998) "TCP behavior of a busy interent server: Analysis and improvements", INFOCOM'98. Seventeenth annual joint conference of the computer and communication societies.

[3] Chelsio communication white paper (2007), "Ethernet Jumbo Frames. The Good, the Bad and the Ugly" ITG fachbericht photonische netze journal.

[4] Chung, J. and Claypool, M. (2003), "Analysis Active Queue Management", IEEE international symposium on network computing and applications conference.

[5] Dykstra, P. (1999), "Gigabit Ethernet Jumbo Frames, And why you should care", White paper, WareOnEarth Communications and Available at: http://sd.wareonearth.com/phil/jumbo.html,1999.

[6] Floyd, S. and Jacobson, V. (1993), "Random Early Detection Gateways for Congestion Avoidance", IEEE/ACM Transactions on Networking, pages 397-413.

[7] Floyd, S., Gummadi, R. and Shenker, S. (2001) "Adaptive RED: An Algorithm for Increasing the Robustness of RED's Active Queue Management", Preprint journal.

[8] Gass, R. (2004), "Packet size distribution" ACM SIGMETRICS performance evaluation review journal, pages 373.

[9] Genkov, D. and Llarionov, R. (2006), "Avoiding IP Fragmentation at the Transport Layer of the OSI Reference Model", Proceedings of the international conference on computer systems and technologies – CompSysTech, University of Veliko Tarnovo, Bulgaria.

[10] Lakshmant, T. V., Neidhardt, A., Teunis, J. (1996), "The Drop from Front Strategy in TCP and in TCP over ATM", Proceedings of the fifteenth annual joint conference of the IEEE computer and communications societies conference, pages 1242 - 1250.

[11] Ramakrishnan, K. and Floyd, S. (1999), "A proposal to add Explicit Congestion Notification (ECN) to IP", IETF RFC 2481.

[12] Rekhter, Y. and Li, T. (1995) "A Border Gateway Protocol", IETF RFC 1771.

[13] Sauver, J. S. (2003), "Practical Issues Associated with 9K MTUs", University of Oregon computing center journal.

[14] Sathaye, S. (2009), "Jumbo Framescommons design".

[15] Y n, N. and Hluchyj, M. G. (1990), "Implication of Dropping Packets from the Front of a Queue", IEEE transactions on communications.

Cross Layer QoS Support Architecture with Integrated CAC and Scheduling Algorithms for WiMAX BWA Networks

Prasun Chowdhury, Iti Saha Misra, Salil K Sanyal
Department of Electronics and Telecommunication Engineering
Jadavpur University, Kolkata-700032, India

***Abstract*— In this paper, a new technique for cross layer design, based on present Eb/N0 (bit energy per noise density) ratio of the connections and target values of the Quality of Service (QoS) information parameters from MAC layer, is proposed to dynamically select the Modulation and Coding Scheme (MCS) at the PHY layer for WiMAX Broadband Wireless Access (BWA) networks. The QoS information parameter includes New Connection Blocking Probability (NCBP), Hand off Connection Dropping Probability (HCDP) and Connection Outage Probability (COP). In addition, a Signal to Interference plus Noise Ratio (SINR) based Call Admission Control (CAC) algorithm and Queue based Scheduling algorithm are integrated for the cross layer design. An analytical model using the Continuous Time Markov Chain (CTMC) is developed for performance evaluation of the algorithms under various MCS. The effect of Eb/No is observed for QoS information parameters in order to determine its optimum range. Simulation results show that the integrated CAC and packet Scheduling model maximizes the bandwidth utilization and fair allocation of the system resources for all types of MCS and guarantees the QoS to the connections.**

Keywords- Cross layer QoS support architecture; SINR based CAC algorithm; Queue based Scheduling algorithm; Adaptive Modulation and Coding; WiMAX BWA networks.

I. INTRODUCTION

Successful delivery of real-time multimedia traffic over the IP based networks, like Internet is a challenging task because of the strict requirement of Quality-of-Service (QoS). Traditional IP based best effort service will be unable to fully meet all the stringent requirements. The time-varying nature of channels along with the resource constrained devices associated with wireless networks make the problem even more complicated. In wireless networks in contrast to wired networks, the channel impairments vary rapidly due to the fluctuation of channel conditions even when the transmitter and receiver are stationary. It is, therefore, desirable to adjust the transmitted power, type of modulation and data rate to match the channel condition dynamically in the best possible manner in order to support the highest possible channel capacity even under the worst case condition of the channel. To make it feasible, the cross layer design for wireless multimedia communication networks has gained significant popularity in the research community [1]. In cross layer design, the challenges from the physical wireless medium and the associated QoS demands for the relevant applications are taken into consideration so that the rate, power, and coding at the physical layer can adapt dynamically to meet the stringent requirements of the applications under the current channel and network conditions.

The ever-increasing wireless traffic is a conglomeration of various real-time traffic such as voice, multimedia teleconferencing, games and data traffic such as WWW browsing, messaging and file transfers etc. All these applications require widely varying and diverse QoS guarantees for different types of traffic. Next generation wireless networks such as WiMAX (Worldwide Interoperability for Microwave Access) or IEEE 802.16 [2] have different types of services with varied QoS requirements, but the IEEE 802.16 has not yet defined the standards for QoS guarantees. Wireless channels suffer from bandwidth limitation, fluctuations of the available bandwidth, packet loss, delay and jitter. Real-time media such as video and audio is highly delay sensitive but Non-real time media such as web data is comparatively less delay sensitive but both types require reliable data transmission. Some advanced techniques such as Orthogonal Frequency Division Multiple Access (OFDMA) [3], Adaptive Modulation and Coding (AMC) [4-6] and wide variety of protocols and standards [7] are used to combat the challenges for stringent QoS requirement in wireless networks. Also, the mobile devices are power constrained. Maintaining good channel quality in one side and minimizing average power consumption for processing and communication on the other are two very conflicting requirements. Receivers in multimedia delivery systems are quite different in terms of latency requirements, visual quality requirements, processing capabilities, power limitations, and bandwidth constraints. In view of the above constraints, a strict modularity and protocol layer independence of the traditional TCP/IP or OSI stack will lead to a sub-optimal performance of applications over IP based wireless networks [8]. For optimization, we require suitable protocol architectures that would modify the reference-layered stack by allowing direct communication between the protocols at non-adjacent layers or sharing of the state variables across different layers to achieve better performance. The objective of a cross layer design is to actively exploit this possible dependence between the protocol layers to achieve performance gains. Basically, a cross layer design involves

feedback received from other layers to optimize the relevant parameters in the current layer to achieve the optimum performance. Although the cross layer design is an evolving area of research, considerable amount of work has already been done in this area [1].

In this paper, information from the MAC layer is used to optimize the parameter at PHY layer. Among several alternatives, the IEEE 802.16 PHY layer standard realizes the usage of Orthogonal Frequency Division Multiplexing (OFDM) in order to mitigate the adverse effects of frequency selective multi-path fading and to efficiently contrast Inter-Symbol and Inter Carrier Interferences (ISI and ICI) [9]. In addition, PHY layer also supports variable channel bandwidth, different FFT sizes, multiple cyclic prefix times and different modulation schemes such as QPSK, 16QAM and 64QAM with convolutional encoding at various coding rates [10]. On the other hand, Call Admission Control (CAC) mechanism [11-19] and Scheduling mechanism [17-23] are the important wings of WiMAX QoS framework at MAC layer. A CAC algorithm at base station (BS) will admit a subscriber station (SS) into the network if the BS ensures the minimum QoS requirement of the SS without degrading the existing QoS of other SSs in the network.

Providing CAC in IEEE 802.16 BWA networks guarantees the necessary QoS of different types of connections and at the same time decreases the Blocking Probability (BP) [13], Dropping Probability (DP) [14] and Outage Probability (OP) [24] for all types of connections. These QoS parameters can be used as feedback information to the PHY layer in order to select an appropriate Modulation and Coding Scheme (MCS) for better network performance. Again, scheduling algorithm at SS will distribute the bandwidth of the selected MCS among the real-time and the non-real-time traffic, based on the type of users in the network and their QoS requirements. Scheduling in the uplink direction is tedious because of required bandwidth request, bandwidth grant and transmission of data of each flow which affect the real-time application due to large Round Trip Delay (RTD) [25, 26].

The scheduling algorithm involves in fair allocation of the bandwidth among the users, determining their transmission order and enhancement of bandwidth utilization. Fairness refers to the equal allocation of network resources among the various users operating in both good and bad channel states. In this paper, fairness has been quantified using Jain's Fairness Index [27].

A. Background literature

A significant number of proposals on cross layer designs have been found in recent literature. Many of them refer to feedback mechanisms between the PHY and MAC layers. In [28] authors have introduced a cross layer approach for calculating the QoS indicators (blocking rates, download time and bit error rates) taking into account the PHY layer conditions (modulation and coding, propagation and MIMO), the MAC layer radio resource management algorithms and the higher layer traffic characteristics. Authors have also considered different admission control schemes and studied the impact of adaptive modulation and coding on the performance of elastic traffic. But no packet scheduling algorithms for fair allocation of bandwidth among the traffic sources have been considered. Also optimum range of Signal to Noise Ratio (SNR) for adaptive MCS in the networks has not been determined.

A framework of memory less scheduling policies based on channel quality states, buffer occupancy states and retransmission number states has been provided in [29]. The scheduling policies combine AMC (Adaptive Modulation and Coding) and ARQ (Automatic Repeat Request) in a cross layer fashion that produces a good throughput performance with reduced average delay. However, the authors have not considered any CAC algorithm for connection management in the networks.

Similar to [29], [30] proposes a scheduling algorithm only at the MAC layer for multiple connections with diverse QoS requirements, where each connection employs AMC scheme at the PHY layer over wireless fading channels. A priority function is defined for each connection admitted in the system, which is updated dynamically depending on the wireless channel quality, QoS satisfaction, and services across all layers. The connection with highest priority is scheduled at each time. At the MAC layer, each connection belongs to a single service class and is associated with a set of QoS parameters that quantify its characteristics.

A cross layer optimization mechanism for multimedia traffic over IEEE 802.16 BWA networks has been proposed in [31, 32].The main functionality of the cross layer optimization mechanism resides at the BS part. The authors have used a decision algorithm at the BS part that relies on the values of two major QoS parameters, i.e., the packet loss rate and the mean delay. These QoS parameters activate proper adjustment of the modulation and/or the media coding rate, aiming at improving QoS and system throughput. However, this paper has not considered proper CAC and scheduling mechanisms for the delay sensitive real-time traffic flow at MAC layer. The authors have also not evaluated the optimum range of SNR in order to uniformly control modulation and data encoding rates.

A cross layer scheduler that employs AMC scheme at the PHY layer, according to the SNR on wireless fading channels has been described in [33]. The authors have defined cost function for each kind of multimedia connections based on its service status, throughput or deadline in MAC layer to achieve an optimum tradeoff between the throughput and fairness. Though the authors have included a suitable scheduling algorithm at the MAC layer but the authors have not considered any CAC algorithm.

[34] proposes a cross layer optimization architecture for WiMAX system. It consists of a Cross layer Optimizer (CLO), which acts as an interface between MAC and PHY layers. The CLO gathers and optimizes the parameters from both layers to achieve optimum performance gain. The CLO gets the channel condition information from the PHY layer and bandwidth requests and queue length information from the MAC layer so that it can switch to different burst profile for different modulation and coding schemes. In this case also, the authors have not specified proper CAC and scheduling mechanism at the MAC layer. Moreover, no optimum range of SNR has been taken into account in CLO while switching to different MCS.

Another cross layer approach as provided in [3], analyses several QoS parameters that include the bit rate and the bit error rate (BER) in the PHY layer, and packet average throughput/delay and packet maximum delay in the link layer. Authors show that dynamic OFDMA has a stronger potential to support multimedia transmission than dynamic OFDM-TDMA. Authors have taken care of proper scheduling, throughput and delay guarantee of traffic flows ignoring the CAC mechanism at MAC layer.

[4] describes a polling based uplink scheduling schemes for TCP based applications in a multipoint to-point fixed broadband IEEE 802.16 BWA network. This scheme adapts the transmission rates between the SSs and the BS dynamically using adaptive modulation technique. With adaptive modulation, the uplink scheduling algorithm helps to achieve higher data transmission rate, fairness in slot assignment and in the amount of data transmission.

B. Contribution

The novelty of our work focuses on the dynamic selection of MCS based on Eb/N0 ratio of the concerned connections and relevant feedback information obtained from the MAC layer in terms of New Connection Blocking Probability (NCBP), Hand off Connection Dropping Probability (HCDP) and Connection Outage Probability (COP) to employ the most appropriate MCS in the network. The target values of the feedback information changes the selection of modulation and coding rate dynamically based on optimum range of Eb/N0 ratio of the connections to enable a flexible use of the network resources. The effect of Eb/N0 ratio has been observed on NCBP, HCDP and COP for all the relevant connection types under various MCS by means of exhaustive simulation. The simulation results help to determine the optimum range of Eb/N0 ratio for all the connections with appropriate MCS. It has been observed that the integrated CAC and Scheduling algorithm maximizes the utilization and fair allocation of the system resources for all types of MCS and provide better QoS support with reduced NCBP, HCDP and COP in the IEEE 802.16 BWA networks.

Since there is also a possibility that the wireless channel deteriorates or becomes unavailable at certain instant of time, the CAC and scheduling algorithms at the MAC layer, which has been left undefined in the standard, can get proper knowledge of the channel condition, in terms of SINR of the connections. Thereby, only the connection with a good channel condition is admitted and scheduled for transmission to achieve more gain in bandwidth utilization and fairness index.

An analytical model based on Continuous Time Markov Chain (CTMC) [35] is developed for the more realistic analysis of QoS parameters for the performance evaluation of IEEE 802.16 Broadband Wireless Access (BWA) Networks. As the Markov Chain basically involves state transition analysis, the probabilities for getting admission without degradation of QoS for both real as well as non-real time services is accurately analysed at the state level.

The remainder of the paper is organized as follows: Section II discusses System models. In addition to it, cross layer architecture and a joint algorithm for SINR based CAC and Queue based scheduling are proposed. Detailed description of the analytical model of proposed SINR based CAC algorithm is given in Section III. Section IV sets the parameter for performance evaluation. Section V shows QoS performance results numerically for different MCS. Finally, Section VI concludes the paper.

II. System Model and Cross Layer Architecture

To accommodate a wide variety of applications, WiMAX defines five scheduling services that should be supported by the BS MAC scheduler for data transport. Four service types are defined in IEEE 802.16d-2004 (Fixed) standard [12], which includes UGS (Unsolicited Grant Service), rtPS (Real-time Polling Service), nrtPS (Non Real-time Polling Service), and BE (Best Effort). In addition, one more service i.e. ertPS (Extended Real-time Polling Service) is defined in IEEE 802.16e-2005 (Mobile) standard [14]. The UGS is designed to support real-time service flow that generates fixed-size data periodically, such as T1/E1 and VoIP without silence suppression. On the other hand, the rtPS supports the same with variable data size, such as video streaming services. Similarly, the nrtPS deals with FTP. The BE and ertPS perform tasks related to e-mail and VoIP respectively. The guaranteed delay aspect is taken utmost care in video streaming and VoIP. In the mobile WiMAX environment, the handover procedure begins as soon as the mobile SS moves into the service range of another BS. The relevant QoS parameters specified in the standard are Maximum Sustained Traffic Rate (MSTR), Minimum Reserved Traffic Rate (MRTR), Maximum Latency (ML), Tolerated Jitter (TJ) and Request/Transmission Policy. The Service Flow Identifier (SFID), Connection Identifier (CID) and traffic priority are mandatory for all QoS classes [10].

A. PHY Layer Model

IEEE 802.16 specifies multiple PHY specifications including Single Carrier (SC), SCa, OFDM and OFDMA. Among the several alternatives, we consider only OFDMA in the PHY802.16 layer for the cross layer design because of its limited interference in the network [9]. OFDMA is similar to OFDM using multiple sub-carriers to transmit data. However, while OFDM uses all available sub-carriers in each transmission, different sub-carriers could be arranged to different subscribers in downlink and each transmission could use the available sub-carriers in uplink in OFDMA.

We consider 802.16 PHY layer to be supported with variable channel bandwidth, different FFT sizes, multiple cyclic prefix time and different modulation schemes such as QPSK, 16QAM and 64QAM with convolutional encoding at various coding rates. The raw data rates of the OFDMA are functions of several parameters such as channel bandwidths, FFT size, sampling factor, cyclic prefix time, modulation scheme, encoding scheme and coding rate. It can be up to 70 Mbps by using high-grade modulation scheme with other suitable parameters [10, 36]. For proper design of PHY layer with MCS, we first observe the effect of all the said parameters in terms of raw data rate and spectrum efficiency.

In [36], a method of calculating raw data rate and OFDM symbol duration for different MCS is given. In this paper, we also derive a relevant term, spectrum efficiency (the alternative

term bandwidth efficiency is also frequently used), being defined [37] as the transmitted bit per second per Hertz (b/s/Hz). This normalized quantity is a valuable system parameter. For instance, if data are transmitted at a rate of 1 Mbps in a 0.6 MHz wide baseband system, the spectral efficiency is 1 Mb/s/0.6 MHz, or 1.67 b/s/Hz. In this paper, spectral efficiency is evaluated as the ratio of raw data rate to the channel bandwidth as determined by the particular MCS in IEEE 802.16 BWA network. The calculation for the above parameters is given below. Let, 'R' be the raw data rate;

$$R = N*b*c/T \quad (1)$$

where:

N = Number of used sub-carriers

b = Number of bits per modulation symbol

c = Coding rate

T = OFDM symbol duration

The value of 'N' is a function of the FFT size. Table I lists the values of 'N' for different FFT sizes.

TABLE I. NUMBER OF USED SUB-CARRIERS AS A FUNCTION OF FFT SIZE [36]

FFT Size	Number of Used Sub-carriers (N)
2048	1440
1024	720
512	360
128	72

The OFDM symbol duration is obtained by using the following formulae [36]:

$$Fs = Sampling\ factor * Channel\ Bandwidth \quad (2)$$

$$Tb = FFT\_size / Fs \quad (3)$$

$$Tg = G * Tb \quad (4)$$

$$T = Tb + Tg \quad (5)$$

where:

Fs = Sampling frequency

Tb = Useful symbol time

Tg = Cyclic prefix time

G = Cyclic prefix

T = OFDM symbol duration

Sampling-Factor is set to 8/7, 'G' is varied as 1/32, 1/16, 1/8 and 1/4 and coding rates used are 1/2, 2/3, and 3/4. These are the standard values specified in WiMAX [36].

Based on the given values of Sampling-Factor, Channel bandwidth and 'G', the OFDM symbol duration 'T' is computed by using equations (2) to (5). Raw data rates can be computed using equation (1). We compute raw data rate and spectrum efficiency for different Cyclic Prefix, modulation schemes and coding rates, using FFT_size = 2048 and channel bandwidth= 20 MHz and the computed results are shown in Table II. From Table II, it is observed that raw data rate and spectrum efficiency are maximum when cyclic prefix G =1/32 (i.e. Minimum value of G) which is obvious because this value of G causes lower time spread of the signal which results in the larger data rate and hence provides maximum spectrum efficiency.

B. Medium Access Control (MAC) Layer Model

We consider SINR based CAC algorithm for BS and Queue based Scheduling algorithm for SS in the WiMAX MAC layer for the design of cross layer model. The terminologies used in this section are given in Table III. We make the following assumptions in our proposed algorithm.

1) *ertPS connection requests are considered to be same as rtPS connections, because both connections have the same QoS parameters and differ only by the way of Request/Transmission policy.*
2) *BE connections are not considered in our CAC scheme, because they are designed to support best effort flows which do not need any QoS guarantees.*
3) *The connections of the similar service types have same QoS parameter values.*

TABLE II. LIST OF DATA RATES AND SPECTRUM EFFICIENCY FOR DIFFERENT MCS AND CYCLIC PREFIX

Modulation Type	Coding Rate	Cyclic Prefix=1/32		Cyclic Prefix=1/16		Cyclic Prefix=1/8		Cyclic Prefix=1/4	
		Raw data rate (Mbps)	Spectrum Efficiency	Raw data rate (Mbps)	Spectrum Efficiency	Raw data rate (Mbps)	Spectrum Efficiency	Raw data rate (Mbps)	Spectrum Efficiency
QPSK	1/2	15.5844	0.7792	15.1261	0.7563	14.2857	0.7143	12.8571	0.6429
QPSK	3/4	23.3766	1.1688	22.6891	1.1345	21.4286	1.0714	19.2857	0.9643
16QAM	1/2	31.1688	1.5584	30.2521	1.5126	28.5714	1.4286	25.7143	1.2857
16QAM	3/4	46.7532	2.3377	45.3782	2.2689	42.8571	2.1429	38.5714	1.9286
64QAM	1/2	46.7532	2.3377	45.3782	2.2689	42.8571	2.1429	38.5714	1.9286
64QAM	2/3	62.3377	3.1169	60.5942	3.0252	57.1429	2.8571	51.4286	2.5714
64QAM	3/4	70.1299	3.5065	68.0672	3.4034	64.2857	3.2143	57.8571	2.8929

GLOSSARY

B	Total amount of bandwidth available at the BS for uplink Connections (Mbps)
W	Bandwidth in Hz
B_U	Minimum Reserved Rate for UGS connections (kbps)
B_r^{min}	Minimum Reserved Rate for rtPS connections (kbps)
B_r^{max}	Maximum Sustained Rate for rtPS connections (kbps)
L	Maximum Latency for rtPS connections (ms)
B_n^{min}	Minimum Reserved Rate for nrtPS connections (kbps)
B_n^{max}	Maximum Sustained Rate for nrtPS connections (kbps)
$MRTR_i$	Minimum Reserved Traffic Rate of connection type i
f	Duration of a timeframe which includes downlink and uplink subframes (ms)
r_i	Token arrival rate of a connection type i (kbps)
b_i	Token bucket size of a connection type i (kbits)
m_i	L/f , m_i must be an integer
$E_b/N_{0,i}$	Signal energy per bit to noise density of connection type i
$SINR_i$	Measured signal to interference plus noise ratio of a connection type i
$SINR_{th,i}$	Calculated threshold signal to interference plus noise ratio of a connection type i
B_{rem}	Amount of bandwidth left after bandwidth allocation to admitted connection
$B_{req,i}$	Bandwidth request of a connection type i
C_{NRT}	Total amount of bandwidth allocated to non real-time connections.
C_{rtPS}	Total amount of bandwidth allocated to rtPS connections.
n_u	Number of UGS connections admitted into the network
n_r	Number of rtPS connections admitted into the network
n_n	Number of nrtPS connections admitted into the network
d_r	Current Degraded Bandwidth of rtPS connections
d_n	Current Degraded Bandwidth of nrtPS connections
B_{poll}	Remaining uplink bandwidth allotted for the polling services
B_{total}^r	Remaining uplink bandwidth for real-time traffic flows
B_{total}^n	Remaining uplink bandwidth for non real-time traffic flows

C. Description of the SINR based CAC algorithm

In the proposed CAC mechanism, the QoS requirements under consideration are indicated by SINR value, transmission delay and bandwidth availability of the connections. The proposed CAC algorithm uses the Signal to Interference plus Noise Ratio (SINR) to check for the availability of enough OFDMA sub-carriers for the new connection request. When a new connection request arrives, the BS calculates the SINR threshold (SINRth) for the requesting connection. The BS then updates the measured SINRi for each connection type, with the assumption that requesting connection of service type 'i' is in the system by occupying an OFDMA sub-carrier. If the measured SINRi is greater than or equal to the corresponding SINRth, and the transmission delay requirement is guaranteed for every real time connection in the system and also the required bandwidth is available for the requesting connection, the connection is admitted into the system. However, if the former two conditions are satisfied but bandwidth availability is not met, the requesting connection is admitted in the system through adaptive bandwidth degradation. Otherwise, the connection request is rejected. The detail of CAC mechanism is shown in the Figure 1.

D. Calculation of SINRth and SINRi

The SINRth for the requesting connection should be determined in such a way so that the wireless transmission error remains below the acceptable threshold for the connection type 'i'. To find SINRth, the effective bit rate of the requesting connection type is first determined. The effective bit rate is defined as the minimum bandwidth required by traffic source to meet the transmission error threshold. In this paper, we have considered that the effective bit rate of a connection type 'i', is its MRTR, as defined in the WiMAX standard. Hence, SINR threshold of a connection type 'i' can be calculated as given in equation (6).

$$SINR_{th,i} = \left(E_b/N_0\right)_i * \left(MRTR_i/W\right) \quad (6)$$

Let the requesting connection be the Cpth connection and the requesting connection's traffic type be 'i' and the connections which are already admitted in the particular WiMAX cell is denoted by 'Cj', where $1 \leq j \leq p-1$ and $j \neq p$. To find the effects of the requesting connection on all other admitted connections, the BS measures the total power 'ψ' received from all existing connections 'Cj' in the system. The received power from all existing connection acts as interference to the requesting connection. Let 'Sj' be the received power from the existing connection which acts as interference to the OFDMA sub-carrier of the requesting connection and 'Si' be the signal power of the ith type of the Cpth requesting connection in an OFDMA sub-carrier as shown in figure 2. Hence,

$$\psi = \sum_{j=1}^{p-1} S_j \quad (7)$$

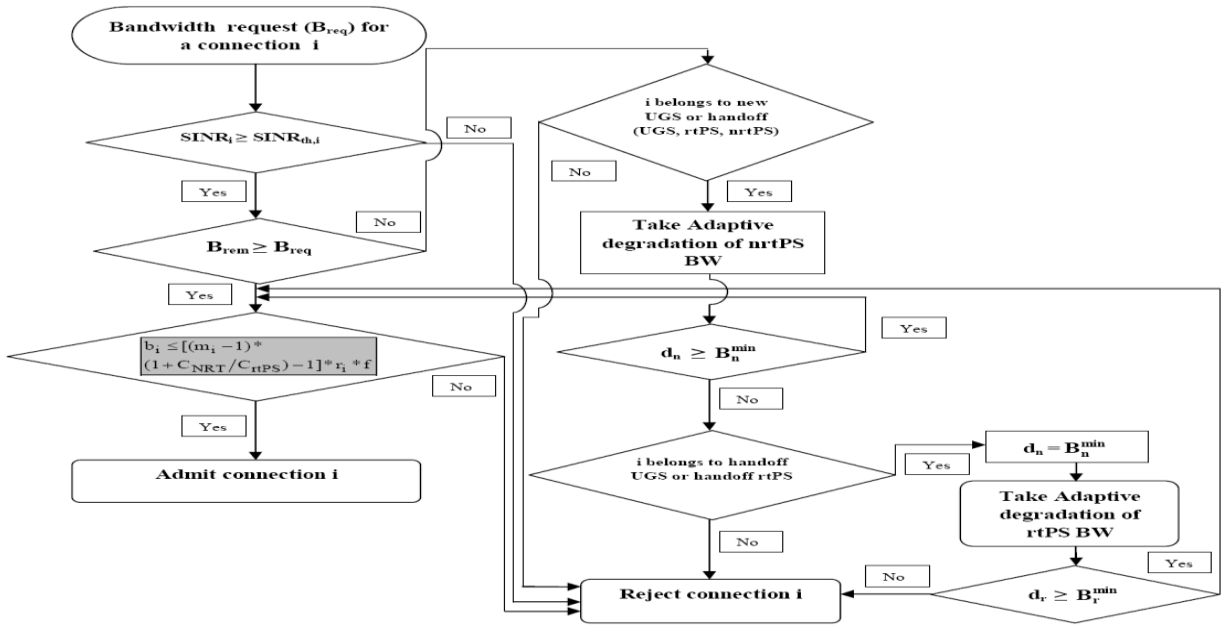


Figure 1. Flow chart of proposed SINR based CAC algorithm for WiMAX BWA Networks

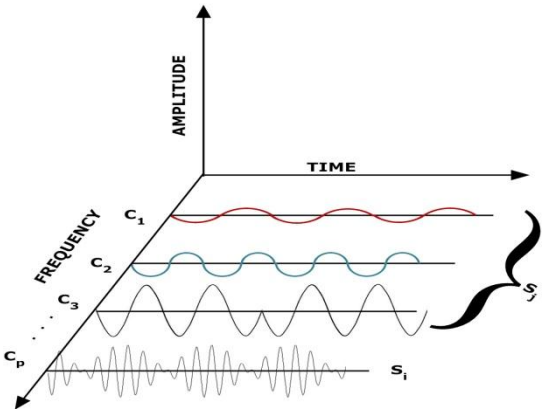


Figure 2. OFDMA sub-carriers in time and frequency domain

Let SINRi be the new measured SINR for a connection when a new connection type 'i' wants to get admitted in the network and 'η' be the thermal noise in the network. SINRi is updated by the BS using equation (10) as derived below [10].

$$SINR_i = \frac{S_i}{\psi + \eta}$$

$$\Rightarrow SINR_i = \frac{S_i / \eta}{\psi / \eta + 1}$$

$$\Rightarrow SINR_i = \frac{S_i / \eta}{\sum_{j=1}^{p-1} S_j / \eta + 1} \quad (8)$$

Again S_i / η for a connection can be calculated as given in [37],

$$\frac{S_i}{\eta} = \left(\frac{E_b}{N_0}\right)_i * \frac{r_i}{W} \quad (9)$$

Replacing $\frac{S_i}{\eta} = \left(\frac{E_b}{N_0}\right)_i * \frac{r_i}{W}$ in (8) we get,

$$SINR_i = \frac{\left(E_b/N_0\right)_i * \left(r_i/W\right)}{\sum_{j=1}^{p-1} S_j / \eta + 1} \quad (10)$$

Where 'r_i' is the transmission data rate of *i*th connection type.

Again $\sum_{j=1}^{p-1} S_j / \eta$ can be calculated using equation (9) for all connection type,

$$\sum_{j=1}^{p-1} S_j / \eta = (n_u * (\frac{E_b}{N_0})_U * \frac{r_U}{W}) + (n_r * (\frac{E_b}{N_0})_r * \frac{r_r}{W}) + (n_n * (\frac{E_b}{N_0})_n * \frac{r_n}{W}) \quad (11)$$

Where 'r_u', 'r_r' and 'r_n' are the bit rates and 'n_u', 'n_r', 'n_n' are the number of admitted connection of UGS, rtPS and nrtPS services respectively.

E. Description of delay guarantee condition

A condition has been provided in [18] to satisfy the Delay Guarantee required by the rtPS connection type 'i'. Since token bucket mechanism is used to schedule the packets, maximum number of packets (in terms of arriving bits) that arrive in a time frame of duration f is bi+ri*f. These arriving bits must be scheduled in next (mi-1) time frame (see Table III) to avoid delay violation. The required condition is given below, (for proof of the condition kindly see [18])

$$b_i \leq [(m_i - 1) * (1 + C_{NRT,i}/C_{rtPS,i}) - 1] * r_i * f \quad (12)$$

Where $C_{NRT,i} = B - (n_U * B_U) - (n_r * B_r)$ and $C_{rtPS,i} = n_r * B_r$.

Here $C_{NRT,i}$ and $C_{rtPS,i}$ are calculated taking the new connection type *'i'* to be admitted. Here, *'i'* refers to the rtPS connection type.

F. Adaptive bandwidth degradation of rtPS and nrtPS connections

In this scheme, bandwidths of lower priority connections are degraded as per minimum bandwidth requirement of the newly admitted or handoff connections. To admit more new UGS connections, the bandwidth is degraded only from nrtPS connections. In addition, degradation is performed on both rtPS and nrtPS connections to allow more UGS, rtPS and nrtPS handoff calls. The advantages of adaptive bandwidth degradation [38] over fixed bandwidth degradation [12- 16] are:

- No fixed step size degradation.
- No need to assign initial arbitrary step size.
- Instead, the minimum required bandwidth is calculated and then degraded.
- The degradation is adaptive to the required bandwidth.
- Bandwidth utilization of the system is greatly improved.

G. Need for rescheduling of bandwidth at SS

According to IEEE 802.16 standard, BS is responsible for allocating the uplink bandwidth based on the request from SSs. As the SS may have multiple connections, the bandwidth request message should report the bandwidth requirement of each connection to BS. All packets from application layer in the SS are classified by the connection classifier based on CID and are forwarded to the appropriate queue. Scheduler at SS will retrieve the packets from the queues and transmit them to the network in the appropriate time slots, as defined by the uplink map massage (UL-MAP) [18] sent by the BS. The UL-MAP is determined by the uplink packet scheduling module based on the BW-request messages that report queue size of each connection in SS. But the scheduler inside the BS may have only limited or even outdated information about the current state of each uplink connection due to large Round Trip Delay (RTD) [25] as shown in Figure 3.

So a scheduling algorithm is needed in each SS to reassign the received transmission bandwidth among different connections. Since the uplink traffic is generated at SS, the SS scheduler is able to arrange the transmission based on the up-to-date information and provide tight QoS guarantee for its connections.

Hereunder, we integrate the distributed scheduling algorithm [26], with our proposed SINR based CAC scheme. The scheduling algorithm is based on current queue size and queue delay at SS.

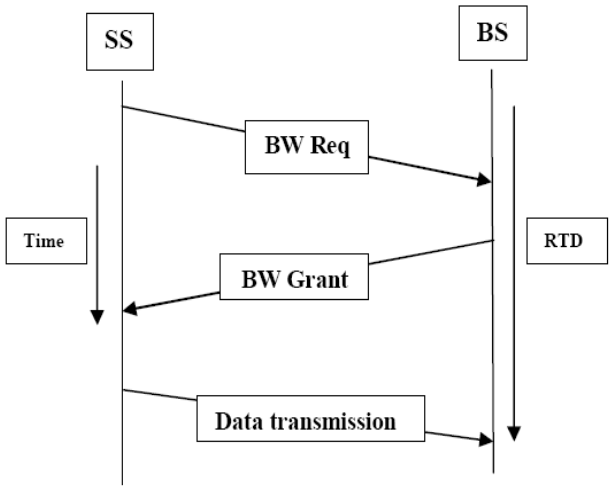


Figure 3. RTD for Uplink transmission in IEEE 802.16

H. Scheduling algorithm at SS

Since UGS service has a critical delay and delay jitter requirement and its transmission cannot be deferred or interrupted by other flows, SS scheduler firstly guarantees the bandwidth for UGS queue. This generally takes a fixed chunk of bandwidth. The remaining uplink bandwidth 'Bpoll' that is allotted for the polling service will be

$$B_{poll} = B - n_U * B_U \quad (13)$$

A parameter α, proposed in [26], is defined as the ratio of the maximum time a rtPS or nrtPS MAC Protocol Data Unit (MPDU) can wait in the queue (i.e. max_mpdu_delay) to the maximum latency specification of the real-time flows.

$$\alpha = \frac{max\_mpdu\_delay}{max\_latency\_of\_rtPS\_flow} \quad (14)$$

This 'α' can be considered as a design parameter which provides the present status of the queue. It also controls the QoS of the real and non-real time services. The remaining uplink bandwidth after allocation to UGS, are divided among 'nr' and 'nn' numbers of real-time and non real-time flows as follows [1, 26].

The real-time traffic flows are allotted an uplink bandwidth of

$$B^r_{tot} = B_{poll} * \frac{n_r * \alpha}{n_r * \alpha + n_n} \quad (15)$$

The non real-time traffic flows are allotted an uplink bandwidth of

$$B^n_{tot} = B_{poll} * \frac{n_r}{n_r * \alpha + n_n} = B_{poll} - B^r_{tot} \quad (16)$$

Every SS repeats this process at the beginning of every uplink. Since the bandwidth request/grant will take some time due to RTD, the rtPS traffic can actually vary within this period. This algorithm is not affected by the rtPS delay bound traffic because the granted bandwidth is always redistributed.

In this way this algorithm provides fair distribution of the bandwidth.

I. Cross layer Architecture

Figure 4 shows the cross layer architecture proposed in this paper. At PHY layer FFT Size =2048 and Cyclic Prefix=1/32 have been selected because these parameters provide highest raw data rate and spectrum efficiency for all types of MCS as shown in Table II. In addition, various MCS like QPSK-1/2, QPSK-3/4, 16QAM-1/2, 16QAM-3/4, 64QAM-1/2, 64QAM-2/3 and 64QAM-3/4 have been selected at PHY layer to evaluate their individual performance in IEEE 802.16 networks.

Further, our proposed CAC algorithm and the Scheduling algorithm have been implemented at MAC layer of BS and SS respectively. The input parameters for the CAC algorithm are the SINR for all the connections, a condition for providing delay guarantee to the rtPS connection and availability of bandwidth. Based on these input parameters, CAC algorithm takes the proper decision to admit or reject an incoming connection request along with its required bandwidth. The output performance parameters obtained from the CAC algorithm are NCBP, HCDP and COP. These output performance parameters act as feedback information, in order to select appropriate MCS dynamically at PHY layer to support the required QoS guarantee.

This dynamic allocation of MCS at PHY layer based on feedback information from MAC layer employs Adaptive Modulation and Coding (AMC) technique in the network. Scheduling algorithm takes current queue size and queuing delay requirement as the input parameters to reschedule the bandwidth shared between real-time and non real-time traffic. It can be noticed that bandwidth which has already been granted or scheduled by the CAC algorithm at BS is again rescheduled by the Scheduling algorithm at SS among the traffic sources. Bandwidth utilization and fairness index are the output performance parameters obtained from the Scheduling algorithm which reveals efficient utilization and fair allocation of the system resources for all kind of MCS. Thereby, when AMC technique selects a MCS in the network, the scheduling algorithm distribute the bandwidth provided by that MCS among the various traffic sources in the network. In this way, Scheduling algorithm helps to improve utilization and fair allocation of the system resources for all MCS.

III. ANALYTICAL MODEL

In this paper, the performance evaluation of the CAC mechanisms is obtained by using the Continuous Time Markov Chain (CTMC) Model [35]. The Markov model has been opted because it examines the probability of being in a given state at a given point of time, the amount of time a system is expected to spend in a given state, as well as the expected number of transitions between states.

A single WiMAX BS in isolation is considered. This BS will receive the bandwidth requests from the SS within the coverage area of a BS. Three types of services UGS, rtPS, nrtPS need QoS guarantees and request for connection admission. BS changes state from one to another upon the admission or rejection of a connection. Further, it is assumed that the BS either admits or rejects only one connection at a particular instant of time. So the next state of the BS depends only on the present state of the BS but does not depend on the previous states of the BS.

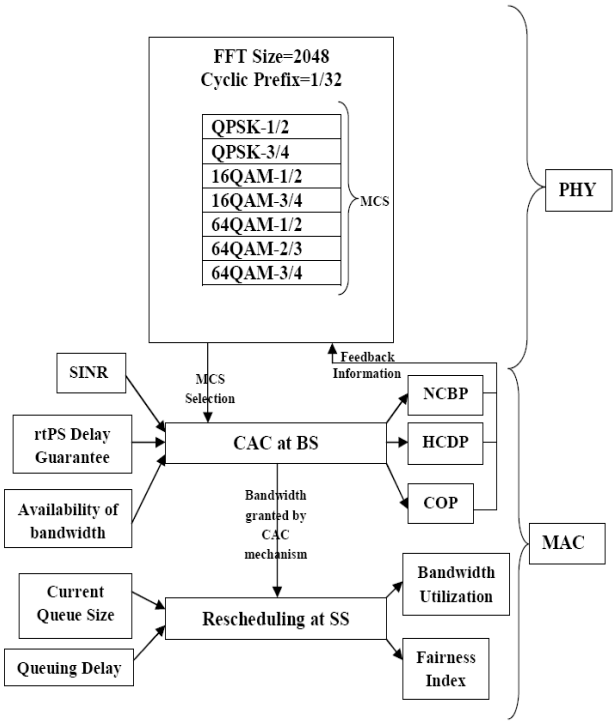


Figure 4. Cross layer architecture for IEEE 802.16 BWA network

Therefore, the states of the BS form a Markov Chain, and accordingly the BS can analytically be modeled as shown in Figure 5. In this scenario, the BS can uniquely be represented in the form of a five dimensional Markov Chain $(n_u, n_r, d_r, n_n, d_n)$ based on the number of admitted connections of each type.

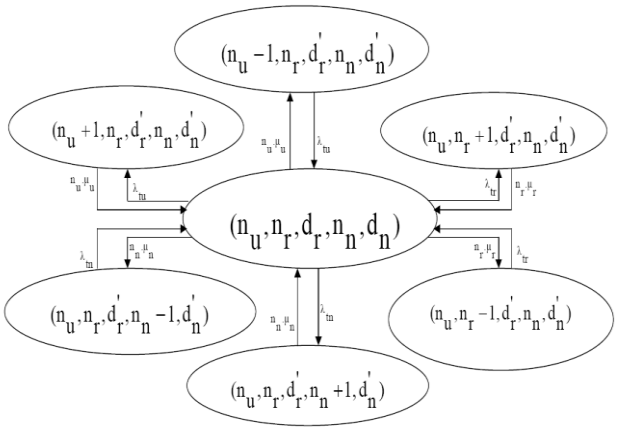


Figure 5. State Transition Diagram of the Markov Chain model for proposed CAC Algorithm

State $s = (n_u, n_r, d_r, n_n, d_n)$ represents that the BS has currently admitted 'n_u', 'n_r', and 'n_n' number of UGS, rtPS and nrtPS connections respectively into the network. In Figure

5, '$d_r^{'}$' and '$d_n^{'}$' are the degraded bandwidth of rtPS and nrtPS connections respectively after state transition and may have a different value from that of 'd_r' and 'd_n'. The BS will be in a particular State $s = (n_u, n_r, d_r, n_n, d_n)$ until a new connection of one of them i.e. UGS, rtPS, nrtPS is admitted into the network or an ongoing connection is terminated. The arrival process of the handoff and newly originated UGS, rtPS, and nrtPS connections is Poisson with rates λ_{hu}, λ_{hr}, λ_{hn}, λ_{ou}, λ_{or}, and λ_{on} respectively. Therefore Total arrival rate λ_{tu}, λ_{tr}, λ_{tn} corresponding to UGS, rtPS and nrtPS connections have been evaluated as given below,

$$\lambda_{tu} = \begin{cases} \lambda_{hu} + \lambda_{ou}, \text{ if } (n_u + 1)B_U + n_r B_r^{max} + n_n B_n^{min} \le B \\ \lambda_{hu}, \quad \text{otherwise} \end{cases} \quad (17)$$

$$\lambda_{tr} = \begin{cases} \lambda_{hr} + \lambda_{or}, \text{ if } n_u B_U + (n_r + 1)B_r^{max} + n_n B_n^{min} \le B \\ \lambda_{hr}, \quad \text{otherwise} \end{cases} \quad (18)$$

$$\lambda_{tn} = \begin{cases} \lambda_{hn} + \lambda_{on}, \text{ if } n_u B_U + n_r B_r^{max} + (n_n + 1)B_n^{max} \le B \\ \lambda_{hn}, \quad \text{otherwise} \end{cases} \quad (19)$$

The service times of UGS, rtPS and nrtPS connections are exponentially distributed with mean $1/\mu_u$, $1/\mu_r$ and $1/\mu_n$ respectively.

The state space **S** for our proposed CAC scheme is obtained based on the following equation.

$$\mathbf{S} = \{ s = (n_u, n_r, d_r, n_n, d_n) \mid (n_u . B_U + n_r . d_r + n_n \cdot d_n) \le B \wedge (B_r^{min} \le d_r \le B_r^{max}) \wedge (B_n^{min} \le d_n \le B_n^{max}) \} \quad (20)$$

From Figure 5, it is observed that every state $s = (n_u, n_r, d_r, n_n, d_n)$ in state space '**S**' is reachable from every other state i.e. each state communicates with other states in the state space '**S**'. As for example, state $(n_u, n_r, d_r, n_n, d_n)$ communicates with state $(n_u, n_r + 1, d_r^{'}, n_n, d_n^{'})$ and also communicates with state $(n_u, n_r, d_r^{'}, n_n - 1, d_n^{'})$. Hence state $(n_u, n_r + 1, d_r^{'}, n_n, d_n^{'})$ and $(n_u, n_r, d_r^{'}, n_n - 1, d_n^{'})$ can also communicate to each other. In this way, each state can communicate with other state in the state space 'S'. Therefore, the state space 'S' forms a closed set and the Markov chain obtained is irreducible [35].

Let the steady state probability of the state $s = (n_u, n_r, d_r, n_n, d_n)$ is represented by $\pi_{(n_u, n_r, d_r, n_n, d_n)}$. As the Markov chain is irreducible, thereby observing the outgoing and incoming states for a given state 's', the state balance equation of state s is shown in equation (21).

$$\begin{aligned} &(\lambda_{tu}\varphi_{(v+1,w.x.y.z)} + \lambda_{tr}\varphi_{(v,w+1,x,y,z)} + \lambda_{tn}\varphi_{(v,w,x,y+1,z)} + \\ &v\mu_u\varphi_{(v-1,w,x,y,z)} + w\mu_r\varphi_{(v,w-1,x,y,z)} + y\mu_n\varphi_{(v,w,x,y-1,z)})\pi_{(v,w,x,y,z)} = \\ &\lambda_{tu}\varphi_{(v-1,w,x',y,z')}\pi_{(v-1,w,x',y,z')} + \lambda_{tr}\varphi_{(v,w-1,x',y,z')}\pi_{(v,w-1,x',y,z')} \\ &+ \lambda_{tn}\varphi_{(v,w,x',y-1,z')}\pi_{(v,w,x',y-1,z')} + (v+1)\mu_u\varphi_{(v+1,w,x',y,z')}\pi_{(v+1,w,x',y,z')} \\ &+ (w+1)\mu_r\varphi_{(v,w+1,x',y,z')}\pi_{(v,w+1,x',y,z')} \\ &+ (y+1)\mu_n\varphi_{(v,w,x',y+1,z')}\pi_{(v+1,w,x',y+1,z')} \end{aligned} \quad (21)$$

Where v, w, x, y and z represent n_u, n_r, d_r, n_n and d_n respectively.

$\varphi_{(v,w,x,y,z)}$ represents the characteristic equation as shown below.

$$\varphi_{(v,w,x,y,z)} = \begin{cases} 1, & (v,w,x,y,z) \in \mathbf{S} \\ 0, & \text{otherwise} \end{cases}$$

By using equation (21), the state balance equations of each state in the state space 'S' are obtained. Solutions of these equations provide the steady state probabilities of all states in the space S with the normalized condition imposed by equation (22).

$$\sum_{s \in S} \pi_{(n_U, n_r, d_r, n_n, d_n)}(s) = 1 \quad (22)$$

From the steady state probabilities we can determine various QoS parameters of the system as given under.

A. *New Connection Blocking Probability (NCBP)*

The new connection blocking probability is the probability of rejecting a new connection request for admission into the network. Conditions for blocking a new connection request have been included in Table IV.

- *Estimation of NCBP for UGS, rtPS, nrtPS connections*

While admitting a new connection, if the next state is not allowable Markov chain state in the state space 'S' due to the conditions as given in Table IV, then the next state is

considered to be a blocked state for that connection i.e. the new connection is blocked.

Let 'S_{UB}', 'S_{rB}' and 'S_{nB}' form a state space for the states whose next state are not allowed in the Markov chain due to connection blocking.

The summation of the steady state probabilities of the states in the state space 'S_{UB}', 'S_{rB}' and 'S_{nB}' give the NCBP of UGS, rtPS and nrtPS connections respectively. Hence,

$$NCBP\text{-}UGS = \sum_{s \in S_{UB}} \pi_{(n_u,n_r,d_r,n_n,d_n)}(s) \quad (23)$$

$$NCBP\text{-}rtPS = \sum_{s \in S_{rB}} \pi_{(n_u,n_r,d_r,n_n,d_n)}(s) \quad (24)$$

$$NCBP\text{-}nrtPS = \sum_{s \in S_{nB}} \pi_{(n_u,n_r,d_r,n_n,d_n)}(s) \quad (25)$$

B. Handoff Connection Dropping Probability (HCDP):

The handoff connection dropping probability is the probability of rejecting a handoff connection request for admission into the network. Conditions for dropping a handoff connection request are summarized in Table IV.

- *Estimation of HCDP for UGS, rtPS and nrtPS connections*

Again, for the handoff connection if the conditions as given in Table IV occur, then the handoff connection is dropped and the next state is not allowed in the Markov chain due to connection dropping.

Similar to NCBP, HCDP can also be estimated as below, where 'S_{UD}', 'S_{rD}' and 'S_{nD}' form a state space corresponding to UGS, rtPS and nrtPS connection for the states whose next state is not allowed in the Markov chain due to connection dropping. Hence,

$$HCDP\text{-}UGS = \sum_{s \in S_{UD}} \pi_{(n_u,n_r,d_r,n_n,d_n)}(s) \quad (26)$$

$$HCDP\text{-}rtPS = \sum_{s \in S_{rD}} \pi_{(n_u,n_r,d_r,n_n,d_n)}(s) \quad (27)$$

$$HCDP\text{-}nrtPS = \sum_{s \in S_{nD}} \pi_{(n_u,n_r,d_r,n_n,d_n)}(s) \quad (28)$$

C. Connection Outage Probability (COP)

The connection outage probability is the probability that the SINR of the connections in the network drops below a certain threshold when a new connection wants to get admitted in the network. Condition for connection outage in the network is summarized in Table IV.

- *Estimation of COP for UGS, rtPS and nrtPS connection*

A new or handoff connection is not admitted in the network if SINR of the connection falls below its threshold value as given in Table IV.

Similar to NCBP and HCDP, COP can be estimated as shown below.

$$COP\text{-}UGS = \sum_{s \in S_{UO}} \pi_{(n_u,n_r,d_r,n_n,d_n)}(s) \quad (29)$$

$$COP\text{-}rtPS = \sum_{s \in S_{rO}} \pi_{(n_u,n_r,d_r,n_n,d_n)}(s) \quad (30)$$

$$COP\text{-}nrtPS = \sum_{s \in S_{nO}} \pi_{(n_u,n_r,d_r,n_n,d_n)}(s) \quad (31)$$

As E_b/N_0 ratio of the connections has the direct relationship (e.g. Equation (6), (10) and (11)) to the SINR and $SINR_{th}$, so from the proposed CAC mechanism and above defined parameters it is observed that the E_b/N_0 ratio provides significant influence on NCBP, HCDP and COP for all the connection types. The effect of E_b/N_0 ratio on NCBP, HCDP and COP will help to determine the optimum range of E_b/N_0 ratio for all the connections, in order to select an appropriate MCS to improve the channel condition. This dynamic selection of MCS based on E_b/N_0 ratio of the connections, employs Adaptive Modulation and Coding (AMC) technique in the network.

D. Bandwidth Utilization (BU):

The Bandwidth Utilization (BU) is defined [16] as the ratio of total used bandwidth to the available bandwidth of the system. BU of the system is encountered to estimate whether SS lying in a bad channel state wastes precious bandwidth. BU can be obtained as follows.

$$BU = \frac{\sum_{s \in S} (n_u.B_U + n_r.d_r + n_n.d_n)\pi_{(v,w,x,y,z)}}{B} \quad (32)$$

Where, 'B' is the total bandwidth.

Fairness Index

Fairness refers to the equal allocation of network resources among the various users operating in both good and bad channel states. Fairness is quantified using Jain's Fairness Index (JFI) [27] as given below.

$$JFI = \frac{(\sum_{i=1}^{n} r_i)^2}{n * \sum_{i=1}^{n} r_i^2} \quad (33)$$

Where, r_i is the data rate of connection type i.

TABLE III. CONDITION FOR CONNECTION BLOCKING, DROPPING AND OUTAGE IN THE NETWORK

Current State	Next State	Condition	Status
$(n_u, n_r, d_r, n_n, d_n)$	$(n_u+1, n_r, d_r, n_n, d_n')$	$d_n' < B_n^{min}$ or $(n_u+1)*B_U + n_r * B_r^{max} + n_n * B_n^{min} > B$	New UGS blocked
		$d_n' < B_n^{min}$ or $d_r' < B_r^{min}$ or $(n_u+1)*B_U + n_r * B_r^{min} + n_n * B_n^{min} > B$	Handoff UGS dropped
		$(E_b/N_0)_u/(n_u*(E_b/N_0)_u + n_r*(E_b/N_0)_r + n_n*(E_b/N_0)_n) < SINR_{th,u}$	UGS outage
	$(n_u, n_r+1, d_r, n_n, d_n')$	$d_n' < B_n^{min}$ or $(n_u*B_U + (n_r+1)*B_r^{max} + n_n*B_n^{min} > B$	New rtPS blocked
		$d_n' < B_n^{min}$ or $d_r' < B_r^{min}$ or $n_u*B_U + (n_r+1)*B_r^{min} + n_n*B_n^{min} > B$	Handoff rtPS dropped
		$(E_b/N_0)_r/(n_u*(E_b/N_0)_u + n_r*(E_b/N_0)_r + n_n*(E_b/N_0)_n) < SINR_{th,r}$	rtPS outage
	$(n_u, n_r, d_r, n_n+1, d_n)$	$n_u*B_U + n_r*B_r^{max} + (n_n+1)*B_n^{max} > B$	New nrtPS blocked
		$d_n' < B_n^{min}$ or $n_u*B_U + n_r*B_r^{max} + (n_n+1)*B_n^{min} > B$	Handoff nrtPS dropped
		$(E_b/N_0)_n/(n_u*(E_b/N_0)_U + n_r*(E_b/N_0)_r + n_n*(E_b/N_0)_n) < SINR_{th,n}$	nrtPS outage

IV. PARAMETERS FOR PERFORMANCE EVALUATION

The frame duration (f) is taken as 1ms because it provides more bandwidth utilization of the system compared to higher length of frame duration as analyzed in our earlier work [19]. Total channel bandwidth (W) is taken as 20 MHz because this is the maximum bandwidth supported by the WiMAX network [36]. The FFT size is taken as 2048 with cyclic prefix 1/32 as this combination provides highest raw data rate and spectrum efficiency for all types of MCS as calculated in Table II.

As the IEEE 802.16 standards have not specified values for the QoS parameters, we have considered the system parameters for performance evaluation as given in Table V. The arrival rates of all the connections are assumed to be same i.e. $\lambda_{hU} = \lambda_{oU} = \lambda_{hr} = \lambda_{or} = \lambda_{hn} = \lambda_{on}$. The service time for UGS, rtPS and nrtPS connections is exponentially distributed with mean $1/\mu_U, 1/\mu_r, 1/\mu_n$ respectively and we assume $\mu_U = \mu_r = \mu_n = 0.2$. That is, the mean service time of the connections is taken as 5 sec. Service time of the admitted connections should be as minimal as possible [16] in order to provide access into the network by other users. Degradation of rtPS and nrtPS connections is adaptive to the required bandwidth.

TABLE IV. QOS PARAMETERS

Service	Max.Sustained traffic rate (kbps)	Min. reserved traffic rate (kbps)	Bucket Size (Bits)	Delay (ms)	Eb/N0 (dB)
UGS	256	256	64	-	3.6
rtPS	1024	512	10240	21	6.3
nrtPS	1024	256	10240	-	8.1

V. NUMERICAL RESULTS

To evaluate the effectiveness and efficiency of the proposed cross layer architecture, the IEEE 802.16 PHY and MAC layer protocols are analyzed using MATLAB under version 7.3. Exhaustive simulations are carried out.

Firstly, the performance of the CAC algorithm is verified by using the CTMC model under various MCS. The effect of E_b/N_0 ratio on our proposed CAC mechanism is observed. The optimum range of E_b/N_0 ratio is determined to select an appropriate MCS in order to employ AMC in the network.

Next, the impact of queue based scheduling algorithm is examined on bandwidth utilization and Jain's fairness index for different connection arrival rates under various MCS. Justifications behind all the numerical results have been provided.

A. Impact of CAC Mechanism on QoS informationParameters

Figure 6 shows the blocking probability of UGS, rtPS and nrtPS connections for different MCS. Total arrival rate of the connections have been taken from equation (17), (18) and (19). It is observed that 64QAM-3/4 provides least blocking probability and QPSK-1/2 provides highest blocking probability to the new connections to be admitted in the network as compared to the other MCS. This is because of the highest raw data rate (70.1299 Mbps) and spectrum efficiency (3.5065) obtained by the 64QAM-3/4 among all other MCS (Reference Table II). Again, it is observed that performances of 16QAM-3/4 and 64QAM-1/2 have been merged together because of their same raw data rate.

Figure 7 shows the dropping probability of UGS, rtPS and nrtPS connections for different MCS. Here, it is also observed

that, 64QAM-3/4 provides least dropping probability and QPSK-1/2 provides highest dropping probability to the ongoing hand off connections as compared to the other MCS because of the same reasons as mentioned previously.

Next, connection outage probability of UGS, rtPS and nrtPS are examined under different MCS and results are shown in Figure 8. Outage probability is the probability that the SINR of the connections in the network drops below a certain threshold when a new connection wants to get admitted in the network and thereby required QoS guarantee is not met to the admitted connections. It is observed that 64QAM-3/4 provides highest outage probability and QPSK-1/2 provides least outage probability as compared to other MCS. Since QAM scheme has the higher data rate and spectrum efficiency than the other MCS, so it has the capability of accommodating more number of connections. Thereby the SINR value of each admitted connections in the network gets decreased and falls below the threshold level, which results in the higher outage probability of the connections in the system.

Therefore, QAM schemes provide good QoS in terms of blocking probability of the newly admitted connections and dropping probability of the hand off connections but fail to provide good QoS in terms of connection outage probability of the network. On the other hand, QPSK schemes provide good QoS in terms of connection outage probability but fail to provide good QoS in terms of blocking and dropping probability.

Hence, there is a need for Adaptive Modulation and Coding (AMC) technique whose performance is based on feedback information of the MAC layer QoS parameters like connection blocking probability, connection dropping probability and connection outage probability. By dynamically changing the allocation of modulation and coding rate based on Eb/N0 ratio as well as SINR of the connections, AMC enables a flexible use of the network resources that can support nomadic or mobile operations.

B. Impact of Eb/N0 ratio on QoS information Parameters

To comprehend how modulation and coding schemes could be made adaptive in a region under WiMAX cell boundary, the proposed SINR based CAC scheme is again simulated under various MCS. We consider Eb/N0 ratio of all the connections because it has the influence on both SINR and SINRth of the admitted connections in the network. Eb/N0 ratio is varied between 1 dB to 20 dB because within this span we are able to find out the adaptive range of all MCS. Connection arrival rate is kept constant at 10 calls per second.

Figure 9 shows the blocking probability under various MCS with different Eb/N0 ratio. Table VI shows the suitable range of Eb/N0 ratio where different MCS show zero blocking probability for all the connections. As for example, 64QAM-3/4 provides zero blocking probability in the range 4 dB to 20 dB for UGS and rtPS connection but for nrtPS connection the range is 9dB to 20 dB. So the optimum range of Eb/N0 ratio in which 64QAM-3/4 provides zero blocking probability is 9dB to 20 dB for all connections types. Again, we have taken 20 dB as the maximum value of the Eb/N0 ratio in our simulation so it is regardless to mention the maximum value of the Eb/N0 ratio because the MCS can provide zero blocking probability beyond this range.

Next, Figure 10 shows the dropping probability under various MCS with different Eb/N0 ratio, whereas Table VI also shows the range of Eb/N0 ratio where different MCS show zero dropping probability for all the connections in our proposed SINR based CAC scheme.

Figure 11 shows the outage probability under various MCS with different Eb/N0 ratio and also Table VI shows suitable range of Eb/N0 ratio in which different MCS show zero outage probability for all the connections.

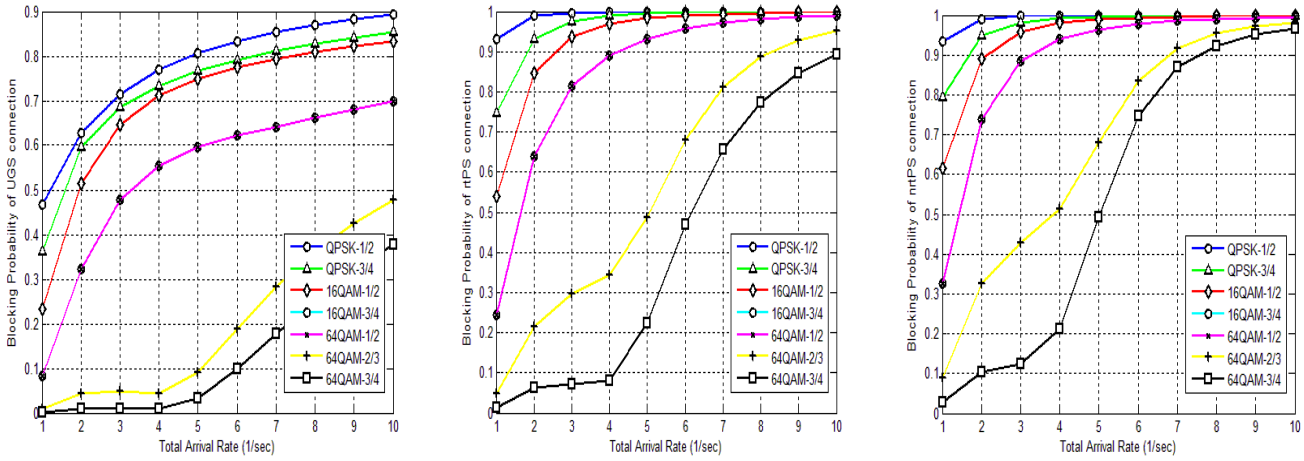
TABLE V. RANGE OF EB/N0 RATIO

Modulation and Coding Scheme	Minimum Eb/N0 ratio(dB) for zero blocking probability obtained from figure 9	Minimum Eb/N0 ratio(dB) for zero dropping probability obtained from figure 10	Maximum Eb/N0 ratio(dB) for zero outage probability obtained from figure 11
64QAM-3/4	9	4	1
64QAM-2/3	10	4	1
64QAM-1/2	14	6	2
16QAM-3/4	14	6	2
16QAM-1/2	20	9	3
QPSK-3/4	-	14	4
QPSK-1/2	-	-	7

By analyzing Table VI, it can be observed that, QAM schemes perform well at higher E_b/N_0 ratio in terms of zero blocking probability (range is above 9 dB) and zero dropping probability (range is above 4 dB) and QPSK schemes perform well at lower E_b/N_0 ratio (range is below 7 dB) in terms of zero outage probability. However, zero probability is not achievable in the practical scenario owing to high intense traffic arrival rate in comparison to 10 calls per second which we have taken for our simulation. Despite of it, the evaluated adaptive range of E_b/N_0 ratio will provide minimum blocking, dropping and outage probability in the practical scenario.

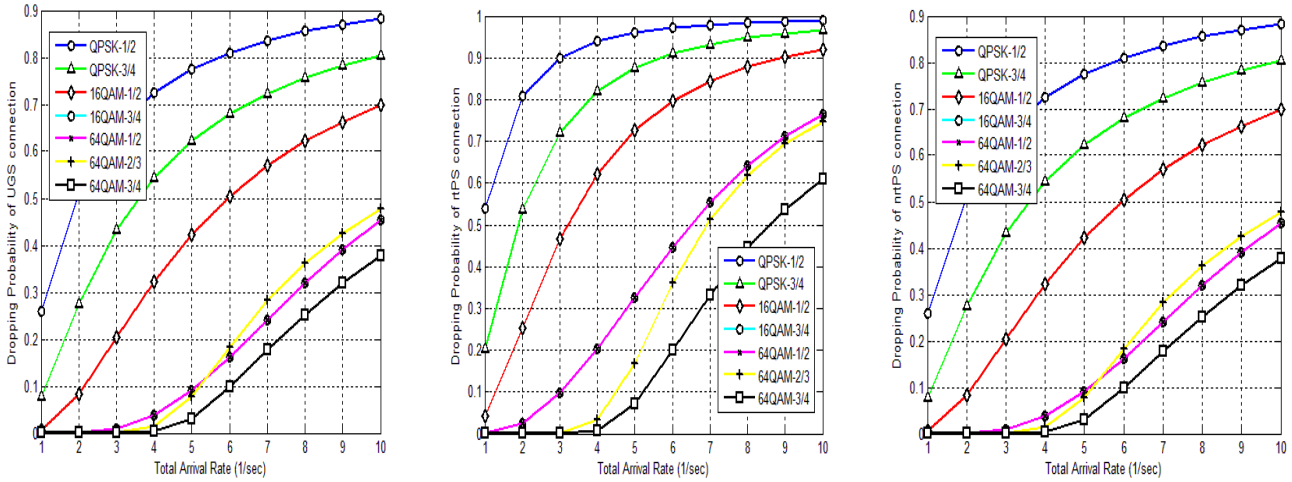
Hence, it is better to select higher order MCS like QAM near to the WiMAX base station where SINR of the connections as well as population density of the users are more and lower order MCS like QPSK far away from the base station where SINR of the connections as well as population density of the users are less. This adaptive allocation of MCS makes the user to experience less blocking, less dropping and less outage in their connection request. Thus, AMC is employed in the network that enables dynamic use of the network resources to support nomadic or mobile operation. Therefore, cross layer adaptation of the different modulation capability at the PHY layer with the QoS requirement at the MAC layer is obtained.

Impact of CAC Mechanism on QoS Information Parameters



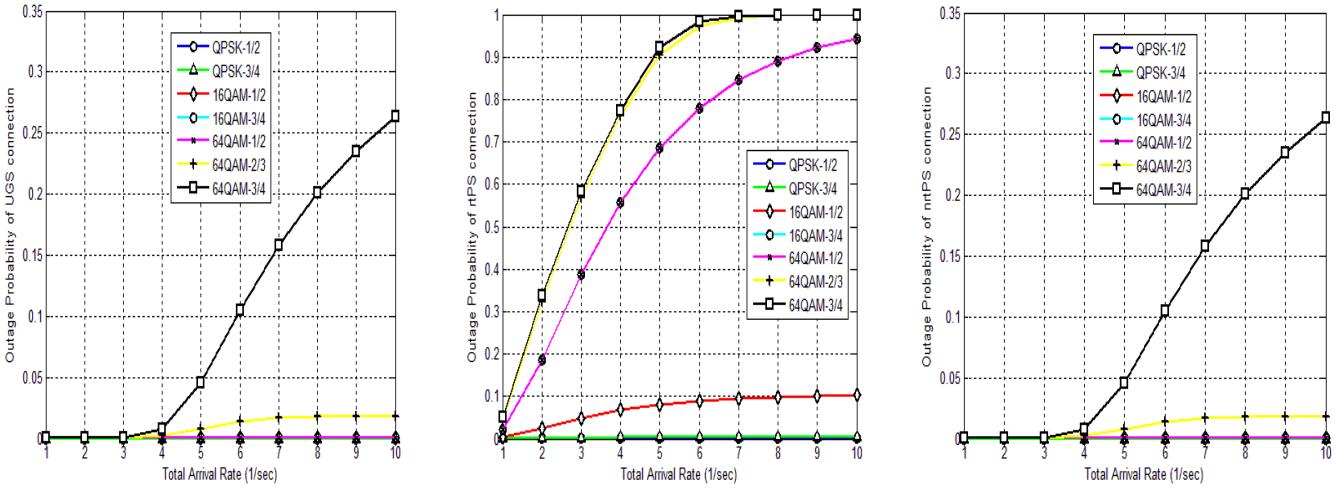
(a) (b) (c)

Figure 1. Blocking probability of UGS connections, (b) Blocking probability of rtPS connections and (c) Blocking probability of nrtPS connections



(a) (b) (c)

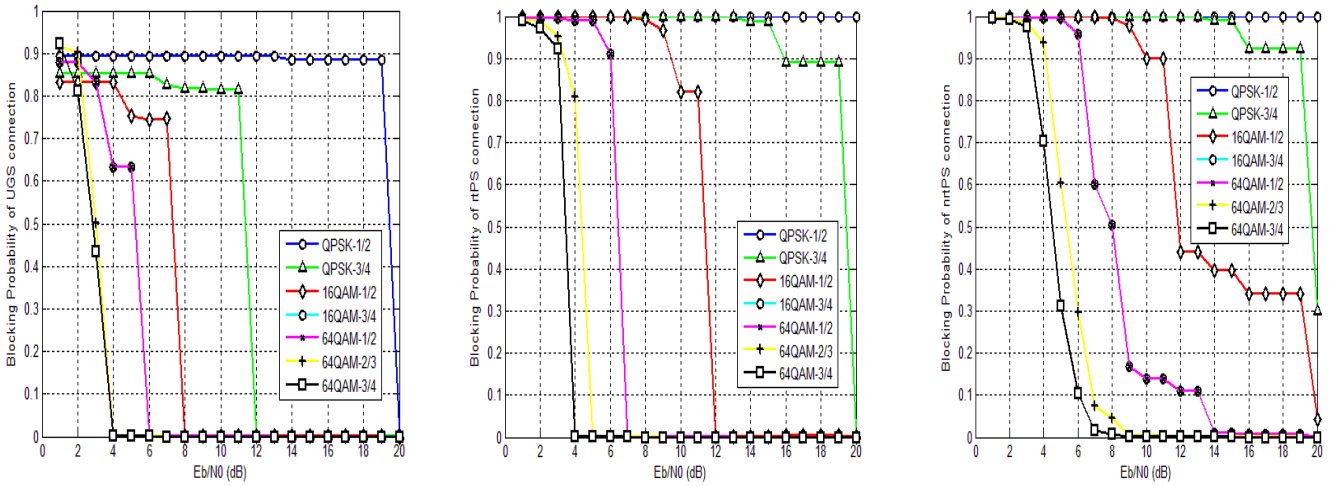
Figure 7(a) Dropping probability of UGS connections, (b) Dropping probability of rtPS connections and (c) Dropping probability of nrtPS connections



(a) (b) (c)

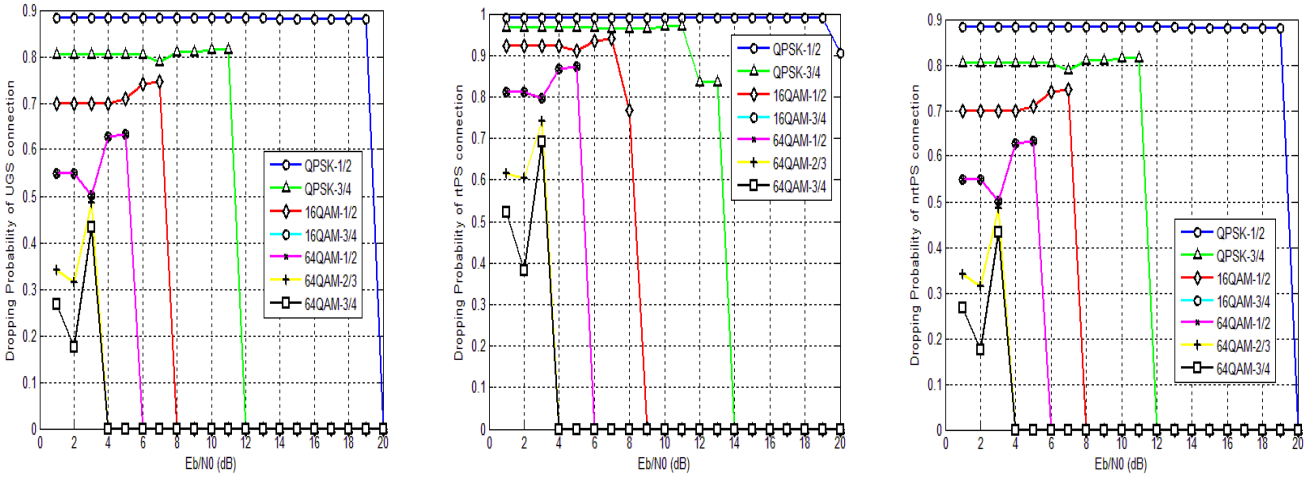
Figure 8(a) Outage probability of UGS connection, (b) Outage probability of rtPS connections and (c) Outage probability of nrtPS connection

Impact of E_b/N_0 ratio on QoS Information Parameters



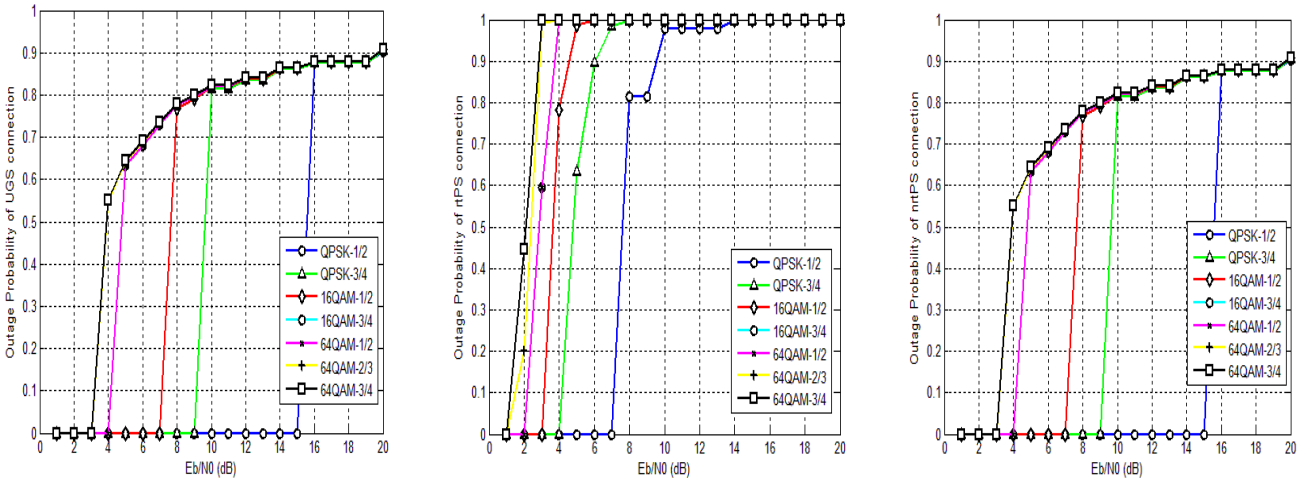
(a) (b) (c)

Figure 9 (a) Blocking probability of UGS connections, (b) Blocking probability of rtPS connections and (c) Blocking probability of nrtPS connection



(a) (b) (c)

Figure 10(a) Dropping probability of UGS connections, (b) Dropping probability of rtPS connections and (c) Dropping probability of nrtPS connections



(a) (b) (c)

Figure 11(a) Outage probability of UGS connections, (b) Outage probability of rtPS connections and (c) Outage probability of nrtPS connections

C. Impact of Queue Based Scheduling on QoS Parameters

To demonstrate the advantage of rescheduling of bandwidth in our proposed cross layer architecture, further simulation has been carried out for different connection arrival rates under various MCS.

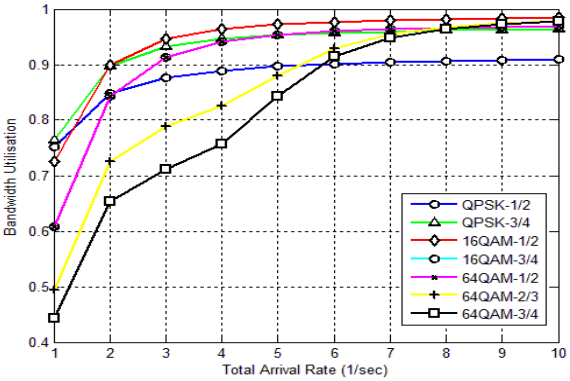


Figure 12(a). Average Bandwidth utilization before queue based rescheduling

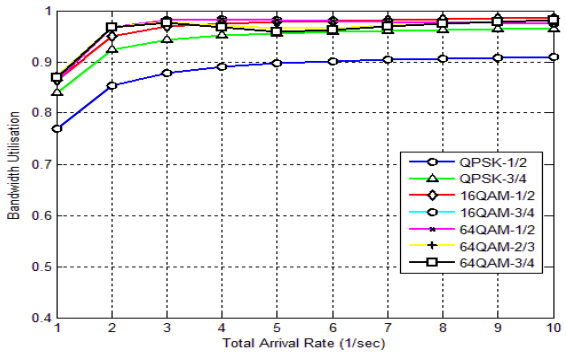


Figure 12 (b). Average Bandwidth utilization after queue based rescheduling

Figure 12(a) and 12(b) show the bandwidth utilization before and after queue based scheduling scheme respectively under different MCS. It shows the significant performance improvement in the bandwidth utilization due to rescheduling of the resource sharing between real-time and non real- time traffic based on their current queue size and latency requirements. From Figure 12(b), it can be observed that bandwidth utilization seems to be independent of MCS i.e. High bandwidth utilization for all MCS. It is because of the influence of scheduling algorithm at SS. As bandwidth is considered to be a limited resource in the network, rescheduling of bandwidth at SS will automatically improve the revenues of the service providers.

On the other hand, Figure 13(a) and 13(b) show the Jain's fairness index before and after the queue based scheduling scheme respectively under different MCS. It is also observed that substantial performance improvement in the fair allocation of resources between real-time and non real-time traffic is achieved because of the rescheduling of the resources. Hence, joint performance of the CAC and rescheduling algorithms also provides better resource utilization at lower traffic arrival rate.

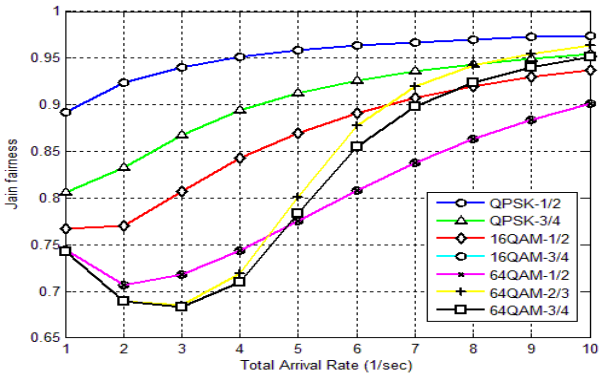


Figure 13(a). Jain's fairness index before queue based rescheduling

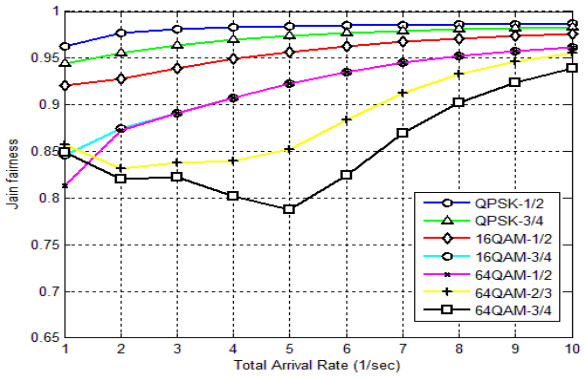


Figure 13(b). Jain's fairness index after queue based rescheduling

Percentage improvements of the performances after rescheduling of bandwidth for all MCS at traffic arrival rate = 1 are shown in Table VII. Highest performance improvement is observed in case of 64QAM-3/4 as compared to the other MCS. This is because 64QAM-3/4 has highest data rate and spectrum efficiency as calculated in Table II.

TABLE VI. PERCENTAGE IMPROVEMENT IN QOS AFTER RESCHEDULING OF BANDWIDTH WHEN TOTAL ARRIVAL RATE = 1

Modulation and Coding Scheme	Percentage improvement in the BU	Percentage improvement in the JFI
QPSK-1/2	2.3813	7.4288
QPSK-3/4	10.0498	20.1068
16QAM-1/2	18.8653	29.4374
16QAM-3/4	43.2356	43.3220
64QAM-1/2	42.2646	45.8475
64QAM-2/3	76.5587	46.6461
64QAM-3/4	96.6757	49.1656

When AMC technique selects an MCS in the network based on the feedback information obtained from the CAC algorithm, the scheduling algorithm redistribute the bandwidth of that selected MCS among the various traffic sources in the network. In this way, scheduling algorithm helps to improve utilization and fair allocation of the system resources for all MCS in IEEE 802.16 BWA networks.

VI. CONCLUSION

Cross layer adaptations are essential for guaranteeing QoS supports in real-time multimedia traffic over wireless networks. In this paper, a cross layer architecture for adapting different MCS in PHY layer has been considered by incorporating MAC layer information in terms of New Connection Blocking Probability (NCBP), Hand off Connection Dropping Probability (HCDP) and Connection Outage Probability (COP) for WiMAX BWA systems. In literatures, many researchers have proposed various cross layer mechanisms for providing better QoS support to the system, but so far no such comprehensive cross layer design considering the parameters stated above, has yet been reported in the literature. In this work, SINR based CAC integrated with the Queue based Scheduling has been analysed with Markov Chain model. The effect of E_b/N_0 ratio is observed on NCBP, HCDP and COP for all the connection types under various MCS by means of exhaustive simulation. Also optimum range of E_b/N_0 ratio is determined, in order to select an appropriate MCS which may be used as the threshold parameters for SINR in any adaptive modulation scheme. Moreover, the joint performance of the CAC and Scheduling algorithms has been proved to be good enough to meet the QoS requirements in terms of bandwidth utilization and Jain's fairness index.

ACKNOWLEDGEMENT

The authors deeply acknowledge the support from DST, Govt. of India for this work in the form of FIST 2007 Project on "Broadband Wireless Communications" in the Department of ETCE, Jadavpur University.

REFERENCES

[1] Y. Zhang, "WiMAX Network Planning and Optimization", CRC Press, Taylor and Francis Group, 2009.

[2] WiMAX Forum, http://www.wimaxforum.org/home/.

[3] Y.J. Chang, F.T. Chien, C.C.J. Kuo, "Cross layer QoS Analysis of Opportunistic OFDM-TDMA and OFDMA Networks", IEEE Journal on Selected Areas in Commun. Vol. 25(4) pp. 657 – 666, 2007.

[4] H.K. Rath, A. Karandikar, V. Sharma, "Adaptive Modulation-Based TCP-Aware Uplink Scheduling in IEEE 802.16 Networks", IEEE International Conference on Commun., pp. 3230-3236, 2008.

[5] D. Marabissi, D.Tarchi, R. Fantacci, F. Balleri, "Efficient Adaptive Modulation and Coding Techniques for WiMAX Systems " IEEE International Conference on Communications, ICC '08, pp. 3383 – 3387, 2008.

[6] V. Nitnaware, S. Lande, S. Balpande, "Comparative Study of Various Modulation Techniques used in Point to Multipoint Mode for WiMAX" Published in International Journal of Advanced Engineering & Applications, pp. 46-56,Vol.1, Jan. 2010.

[7] B. Rong, Y. Qian, H. H. Chen, "Adaptive power allocation and call admission control in multiservice WiMAX access networks" IEEE Wireless Commun., Vol. 14(1), pp. 14 – 19, 2007.

[8] F. D. Priscoli, T. Inzerilli, L. Munoz, "QoS Provisioning in Wireless IP Networks", Wireless Personal Communication, Springer, pp. 23-39, 2006.

[9] L. Litwin, M. Pugel, "The Principles of OFDM", www.Rfdesign.com, Jan 2001.

[10] J. G. Andrews, A. Ghosh, R. Muhamed, "Fundamentals of WiMAX-Understanding Broadband Wireless Networking", PEARSON Education, March 2007.

[11] M. H. Ahmed, "Call Admission Control in Wireless Networks: A comprehensive Survey", IEEE Communications Surveys & Tutorials, First Quarter, Vol.7 (1) pp. 50-69, 2005.

[12] H. Wang, W. Li and D. P. Agrawal, "Dynamic admission control and QoS for IEEE 802.16 Wireless MAN", Proc. of Wireless Telecommunications Symposium (WTS 2005), pp. 60-66 April 2005.

[13] F. Hou, P. H. Ho, X. Shen, "Performance Analysis of Reservation Based Connection Admission Scheme in IEEE 802.16 Networks", Global Telecommunications Conference, GLOBECOM'06, pp. 1-5, November 2006.

[14] L. Wang, F. Liu, Y. Ji, N. Ruangchaijatupon, "Admission Control for Non-preprovisioned Service Flow in Wireless Metropolitan Area Networks", Universal Multiservice Networks, (ECUMN-07). Fourth European Conference, pp. 243 – 249, Feb. 2007.

[15] Y. Ge, G.S. Kuo, "An Efficient Admission Control Scheme for Adaptive Multimedia Services in IEEE 802.16e Networks", IEEE 64th Vehicular Technology Conf., pp.1 – 5, Sept. 2006.

[16] K. Suresh, I. S. Misra and K. Saha (Roy), "Bandwidth and Delay Guaranteed Call Admission Control Scheme for QOS Provisioning in IEEE 802.16e Mobile WiMAX" Proceedings of IEEE GLOBECOM, pp.1245-1250, December 2008.

[17] C. H. Jiang, T.C. Tsai, "CAC and Packet Scheduling Using Token bucket for IEEE 802.16 Networks", Consumer Communications and Networking Conf., 2006. CCNC 2006. 3rd IEEE Vol. 1, pp. 183-187, Jan. 2006.

[18] K. Wongthavarawat, A. Ganz, "Packet scheduling for QoS support in IEEE 802.16 broadband wirelessaccess systems", International Journal of Communication Systems, Vol. 16(1), 81-96, 2003.

[19] P. Chowdhury, I. S. Misra, S. K. Sanyal, "An Integrated Call Admission Control and Uplink Packet Scheduling Mechanism for QoS Evaluation of IEEE 802.16 BWA Networks", Canadian Journal on Multimedia and Wireless Networks, Vol.1 (3), 2010.

[20] P. Chowdhury, I. S. Misra, "A Comparative Study of Different Packet Scheduling Algorithms with Varied Network Service Load in IEEE 802.16 Broadband Wireless Access Systems" IEEE Proc. Int. Conf. Advanced Computing & Communications, Bangalore, Dec. 2009.

[21] J.Lin, H.Sirisena, "Quality of Service Scheduling in IEEE 802.16 Broadband Wireless Networks", Proceedings of First Int. Conf. on Industrial and Information Systems, pp.396-401, August 2006.

[22] C.Cicconetti, A.Erta, L.Lenzini, E.Mingozzi, "Performance Evaluation of the IEEE 802.16 MAC for QoS Support", IEEE Transactions on Mobile Computing, Vol. 6(1), pp. 26-38, 2007.

[23] P. Chowdhury, I. S. Misra, "An Efficient Quality of Service Scheduling Strategy for IEEE 802.16 Broadband Wireless Access Systems", proceedings of Int. conf. NeCoM-2010, Recent Trends in Netwoks and communication, springer, Vol. 90, pp. 306-315, July 2010.

[24] W. Ye, A. M. Haimovich, "Outage Probability of Cellular CDMA Systems with Space Diversity, Rayleigh Fading, and Power Control Error", IEEE Communications Letters, Vol.2 (8), pp. 220-222, 1998.

[25] C. S. Chu, D. Wang, S. Mei, "A QoS architecture for the MAC protocol of IEEE 802.16 BWA system", IEEE communications, Vol.1, pp. 435-439, 2002.

[26] K. R. Raghu, S. K. Bose, M.Ma, "Queue based scheduling for IEEE 802.16 wireless broadband", proceedings of ICICS, Vol. 5(6), pp.1-5, 2007.

[27] A. Haider, R. Harris, "A novel proportional fair scheduling algorithm for HSDPA in UMTS networks", Proceedings of 2nd IEEE AusWireless, pp.43-50, 2007.

[28] S. E. Elayoubi, B. Fourestie, "Performance Evaluation of Admission Control and Adaptive Modulation in OFDMA WiMax Systems" IEEE/ACM Transaction on Networking, Vol. 16 (5), pp. 1200-1211, 2008.

[29] M. Poggioni, L. Rugini, P. Banelli, "QoS Analysis of a Scheduling Policy for Heterogeneous Users Employing AMC Jointly with ARQ" IEEE Transaction on Communication, Vol. 58(9), pp. 2639-2652, 2010.

[30] Q. Liu, X. Wang, G.B. Giannakis, A cross layer scheduling algorithm with QoS support in wireless networks. IEEE Transactions on Vehicular Technology, Vol. 55(3), pp. 839-84 ,2006.

[31] D. Triantafyllopoulou, N. Passas, A.K. Salkintzis, A. Kaloxylos, "A cross layer optimization mechanism for multimedia traffic over IEEE 802.16 networks. International Journal of Network Management, Vol. 17(5), pp. 347-361, 2007.

[32] D. Triantafyllopoulou, N. Passas, A.K. Salkintzis, A. Kaloxylos, "A Heuristic Cross layer Mechanism for Real-Time Traffic in IEEE 802.16 Networks" , IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1-5, September 2007.

[33] X. Li, X. Wu, W. Li, X. Wang, "An Adaptive Cross layer Scheduling Algorithm for Multimedia Networks" International Conference on Intelligent Information Hiding and Multimedia Signal Processing, , pp. 52-55, 2008.

[34] K. A. Noordin, G. Markarian, "Cross layer Optimization Architecture for Wimax Systems", IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1-4, 2007.

[35] S. M. Ross. "Probability Models for Computer Science", Elseveir, June 2001.

[36] "QualNet 4.5 Advanced Wireless Model Library" , Scalable Network Technologies, Inc., http://www.scalable-networks.com.

[37] K. Feher, "Wireless Digital Communications: Modulation and Spread Spectrum Application", Prentice-Hall, 1995.

[38] C. T. Chou, K.G. Shin, "Analysis of adaptive bandwidth allocation in wireless networks with multilevel degradable quality of service" IEEE Transactions on Mobile Computing, Vol. 3, pp. 5 – 17, 2004.

AUTHORS PROFILE

1. *Mr. Prasun Chowdhury* (prasun.jucal@gmail .com) has completed his Masters in Electronics and Telecommunication Engineering from Jadavpur University, Kolkata, India in 2009. Presently he is working as Senior Research Fellow (SRF) in the Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata, India. His current research interests are in the areas of Call Admission control and packet scheduling in IEEE 802.16 BWA Networks. He has authored some journals and international conference papers.

2. *Dr. Iti Saha Misra* (itimisra@cal.vsnl.net.in) received her PhD in Microstrip Antennas from Jadavpur University (1997). She is presently a Professor in the Department of Electronics and Telecommunication Engineering, Jadavpur University, India. Her current research interests include Mobility and Location Management, Next Generation Wireless Network Architecture and protocols, Call Admission control and packet scheduling in cellular and WiMAX networks. She has authored more than 100 research papers in refereed Journal and International Conference and a book on Wireless Communications and Networks. She is an IEEE Senior Member and founder Chair of the Women In Engineering, Affinity Group, IEEE Calcutta Section.

3. *Dr. Salil K. Sanyal* (s_sanyal@ieee.org) received his Ph.D from Jadavpur University, India (1990). He is currently a Professor in the Department of Electronics and Telecommunication Engineering, Jadavpur University. He has authored more than 130 Research Papers in refereed Journals and International/National Conference Proceedings and also co-authored the Chapter "Architecture of Embedded Tunable Circular Microstrip Antenna" in the book entitled "Large Scale Computations, Embedded Systems and Computer Security". He is a Senior Member of IEEE and past Chair of IEEE Calcutta Section. His current research interests include Analog/Digital Signal Processing, VLSI Circuit Design, Wireless Communication and Tunable Microstrip Antenna.

A Conceptual Design Model for High Performance Hotspot Network Infrastructure (GRID WLAN)

Udeze Chidiebele. C, Okafor Kennedy .C
R & D Department, Electronics Development Institute
(FMST-NASENI), Awka, Nigeria.

Prof. H. C Inyiama, Dr C. C. Okezie,
Electronics and Computer Engineering Department,
Nnamdi Azikiwe University, Awka, Nigeria

Abstract—The emergence of wireless networking technologies for large enterprises, operators (service providers), small-medium organizations, has made hotspot solutions for metropolitan area networks (MAN), last mile wireless connectivity, mobile broadband solutions, IP-based cellular phones (VOIP) and other event-based wireless solutions in very high demand. Wireless radio gateways (routers and access points) with its wide-spread deployment has made Wi-Fi an integral part of today's hotspot access technology in organizational models. Despite its role in affecting performance for mobility market segments, previous research has focused on media access control (MAC) protocol techniques , carrier sense multiple access with collision avoidance (CSMA/CA), fair scheduling, and other traffic improvement methodologies without detailed consideration for virtual switch partitioning for load balancing, reliable fragmentation capacities, buffer size dependencies, load effects, queuing disciplines as well as hotspot access control framework. This paper proposes a conceptual high performance hotspot solution (GRID WLAN) and through simulations with OPNET modeler, presents efficient performance metrics for its deployment. Considering the GRID WLAN access points in context, the results shows that with the design model, a careful selection of buffer sizes, fragmentation threshold, network management framework with network load intensity will guarantee an efficient hotspot solution.

Keywords: Hotspot solutions, IP-based cellular phones, Buffer size dependencies, GRID WLAN.

I. INTRODUCTION

The mobility market segments has made modern computing very attractive as such creating a platform for application developers INTEGRATE solutions in laptops, cloud computing environments as well as mobile devices. In this research, a conceptual high performance hotspot solution termed GRID WLAN is shown in Fig. 1. Anyone with a PC, notebook, mobile phone or PDA that is Wi-Fi-enabled can use the proposed GRID WLAN hotspot services.

From the user perspective, the network can be selected, browser launched and the user signs up by entering an authentication ID or by using a signed credit card for access control. The network comprises the infrastructure gateway (ADSL modem and GRID WLAN Gateway), Hypervisor layer (Virtualization firmware), GRID WLAN switch/load balancer, GRID WLAN Access points (AP1…….APn), Management framework, Wimax base station, VOIP, and hotspot nodes.

The network model adopted in this research is an extended service set GRID WLAN mode. The APs establishes connections to other users who are directly connected to the hotspots. Basically, FTP traffic and HTTP traffic take place in the GRID WLAN hotspot zone sites while the GRID WLAN gateway access point creates a gateway link to the IP cloud (internet). The scope of this paper is majorly on the capacity of the various buffer sizes, loads as well as fragmentation threshold of the GRID WLAN APs as the principal metrics of interest in the proposed GRID WLAN design. This paper defines capacity in this context to be the maximum number of simultaneous, bidirectional traffic flow that can be supported in the hotspot environment.

The paper is organized as follows: In section II, the literature review was discussed, the general system model and assumptions for GRID WLAN hotspot was presented in III. In IV, GRID WLAN mechanism is presented. Section V gives the simulation results to support our propositions. The paper ends with the conclusions and future directions.

II. LITERATURE REVIEW

A detailed study on various works on hotspot WLAN solutions was carried out. However, a close attention was given to the work in [1] which is the most related work. The authors focused on WLAN performance issues viz: tuning the physical layer related parameters, MAC protocol techniques, tuning the IEEE 802.11 parameters, fair scheduling, and Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) and other traffic improvement methodologies.

A good introduction to the 802.11 standard, followed by a performance study of both the Distributed Coordination Function (DCF) and Point Coordination Function (PCF) is presented in [2]; this study suggests that an IEEE 802.11 network may be able to carry traffic with time-bounded requirements using the PCF. In Point Coordination Function (PCF) mode, polling occurs with a point coordinator determining which station has the right to transmit [3]. In the DCF mode, a BSS operates as an ad-hoc network in which any station can communicate with any other station in the BSA without the intervention of a centralized access point (AP) by contending for a shared channel. All stations have equal priorities and hence an equal chance of getting the channel [3].

Basically, the current use of WLANs for Internet access from wireless stations is dominated by downlink TCP traffic; however, the optimal performance of WLAN setups have not been sufficiently taken into account for data transport over WLANs devices like APs. WLAN parameters can be fine tuned to optimize performance. This has not been addressed in the previous literature reviews [4], [5].This paper shall focus on metrics that will enhance performance in a hotspot setup as shown in our system model in Fig.1.

III. SYSTEM DESIGN AND ASSUMPTIONS

A. System Model

The high performance GRID WLAN hotspot solution is an extended service set wireless configuration shown in Fig. 1. It assumed that the GRID WLAN gateway is a composite powerful gateway combined with radio which acts as a complete hotspot. At the core of the proposed GRID WLAN is a switch load balancer (bridge network) which acts as a distribution medium for all indoor access points, wired LAN, outdoor hotspots nodes, wimax base station, back office server (management framework). This paper practically details the implementation methodology as well as discusses the GRID WLAN mechanism in the section IV.

B. Design Goals

In designing our proposed GRID WLAN hotspot solution, the main goals are to maintain throughput and to create a wireless hotspot solution that is based on robust network infrastructure, hence scalable, flexible and quick to deploy at anytime with less administrative overhead. Fig. 1 shows the proposed system model.

As shown in Fig 1, the core LAN switch with virtual OS connects the ADSL modem, management framework ,GRID WLAN access points (for indoor coverage site), the Wimax base station (outdoor coverage site), local hotspot nodes (hsn1.....hns6) , and VOIP. For our large hotspots including mesh nodes, the high powered GRID WLAN gateway will be capable of supporting over 500 simultaneous user sessions. In this context, we assume the GRID WLAN hotspot to be a combination of wired access points, Wimax base station, and indoor and outdoor mesh access point nodes dependent on the local site.

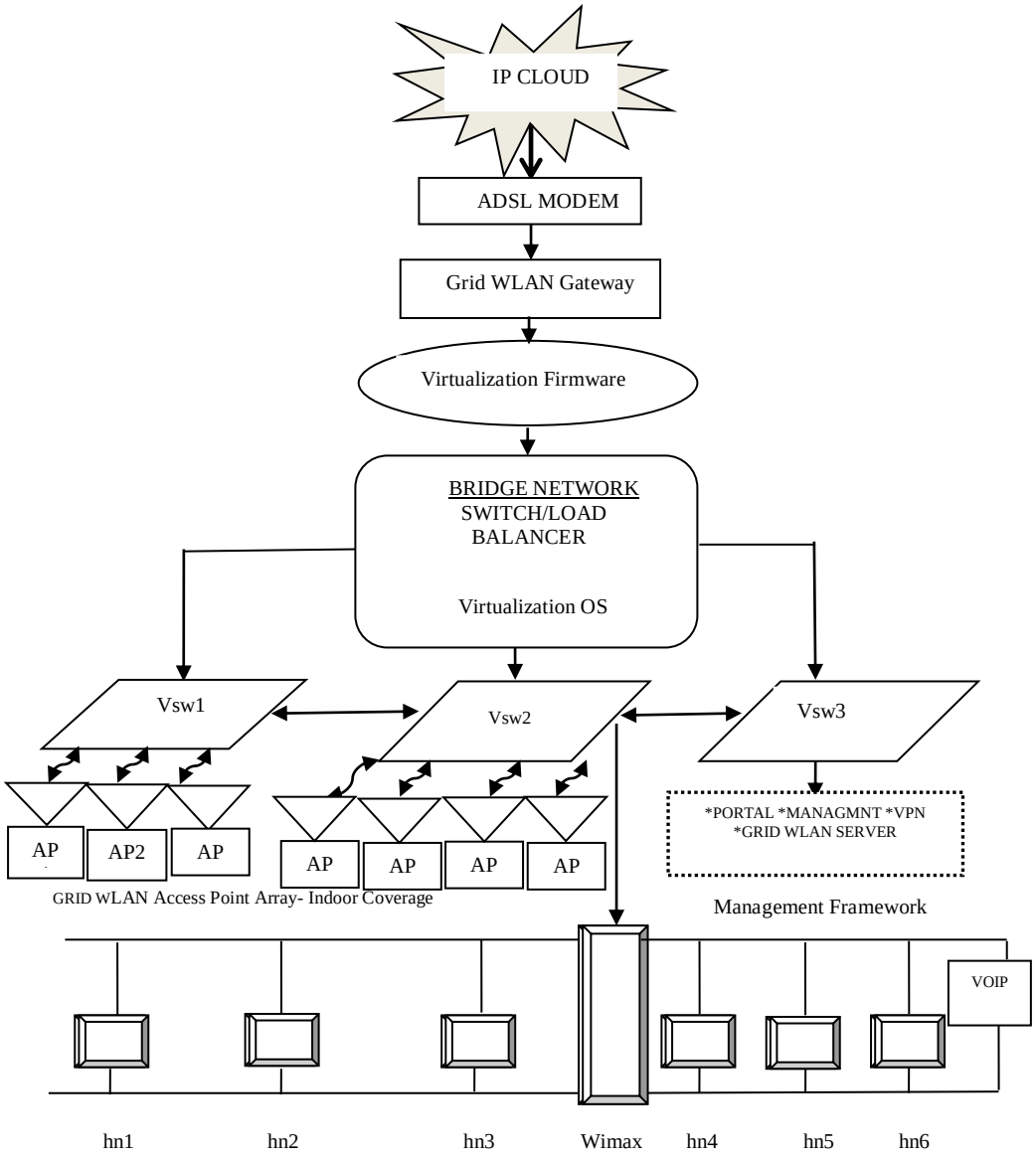


Figure 1. The proposed GRID WLAN Hotspot system model

C. *Experimental Test Bed*

In this paper, to measure the system performance of the proposed GRID WLAN hotspot network infrastructure, OPNET Modeler [6] was used to achieve the objective. OPNET Modeler is a graphical network simulator mainly used for simulation of both wired and wireless communication networks and information systems. For our proposed test bed, the infrastructure components include:

i. 5 Hotspots nodes for outdoor Coverage
ii. Bridge Switch/Load balancer with virtual OS (Extended service mode).
iii. 5 Access Points for indoor coverage
iv. 40 WLAN client Stations
v. Management Server (Enterprise Red hart Server)
vi. Back office Applications (Network Monitoring, Configuration, Authentication, Authorization and Accounting) :
- Grid Control Management server for central network configuration & Monitoring.
- Portal Server for Network Services and management.
- VPN tunnel Terminator server

In Fig 1, the proposed GRID WLAN Hotspot system model utilizes the infrastructure components outline above. A management server has the role authenticating and monitoring the overall network for efficient service delivery. The extended service mode of the bridge switch enables both indoor and outdoor connectivity. The virtual switching by the switch bridge interfaces the management server, hotspot nodes, access points and client stations in our context as shown in Fig. 2a. The Access points were setup in the various subnets with the client base stations (Fig. 2b). The bridge switch is a speed redundancy layer supporting virtualization and load balancing. Access to the IP internet cloud is completed by the gateway and modem. The network model, node model and process models were accomplished in our test bed using OPNET modeller. After setting up the model, a simulation run was carried out to generate our graphical plots shown in this work (Fig 3). Also, a consistency test carried out shows that the design model is stable and consistent before the simulation execution.

IV. DISCUSSION

From Fig. 1, the GRID WLAN hotspot solution employs the concept of mesh routing to create a self configurable, centrally controlled hotspot solution. Every GRID WLAN access point (AP) is capable of relaying traffic coming from any of its neighbours. The mesh network reconfigures itself when a node is added or removed. By adding a GRID WLAN access point (AP), the wireless network range is expanded. Owing to the self-reconfigurable mesh routing algorithms, the network has no point of failure, hence, a reliable wireless platform for connectivity

The GRID WLAN gateway router supports the following functionalities: Automatic Configuration, Bandwidth Control/Optimization, Firewall, Multiple Services, and Virus Checking with Extensive Access list control.

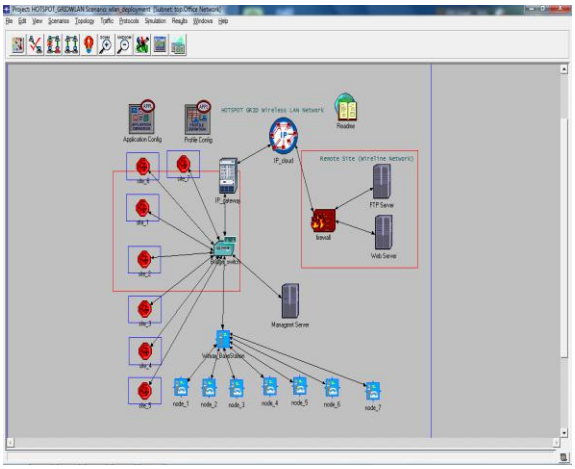


Figure 2. GRID WLAN testbed Implementation with subnetted clients

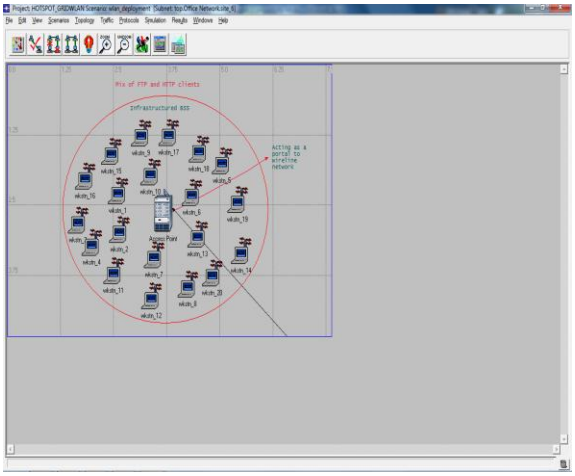


Figure 2b: GRID WLAN nodes

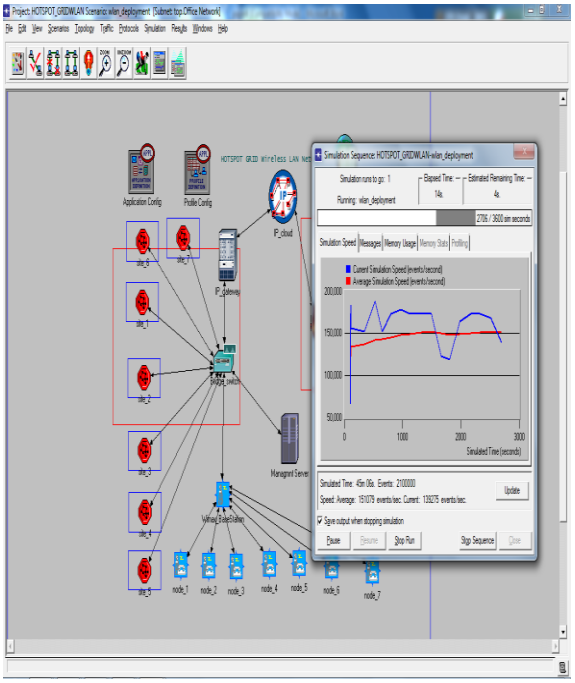


Figure 3. shows the GRID WLAN Simulation run

V. SIMULATION PARAMETERS

In this section, we provide simulation results that support section III. The GRID WLAN APs including the gateway router connects wireless network (nodes) to wired network to wired networks via the bridge switch as depicted in our model.

A. *Simulation Configurations*

We used OPNET modeler [6] was used to generate the parameters for various case scenarios in the simulations. Traffic attributes are listed in Table 1, while Table 2 show the GRID WLAN parameters used and the various load intensities from OPNET Modeler.

TABLE 1. OPNET TRAFFIC ATTRIBUTES

	CONFIGURATIONS	Values
1.	Simulation Duration for Each Scenario	60mins
2.	Link Propagation Delays	0.5usecs
3.	Switch Output Buffer	100 packets
4.	Simulation Seed	128
5.	Update Interval	50000 Events
6.	Simulation Kernel	Optimized

TABLE 2. GRID WLAN SIMULATION PARAMETER

WLAN PARAMETERS	5 Sources WLAN BUFFER-64K	10 Sources WLAN BUFFER-128K	15 Sources WLAN BUFFER-256K	20 Sources WLAN BUFFER-512K	30 Sources WLAN BUFFER-1024K	40 Sources WLAN BUFFER-2048K
RTS-Threshold (Bytes)	256 Bytes	256Bytes	256Bytes	256Bytes	256Bytes	256Bytes
Fragmentation Threshold (Bytes)	256 Bytes	512Bytes	1024Bytes	2048Bytes	4096Bytes	8192Bytes
Data Rate (bps)	54	54	54	54	54	54
Physical Characteristics	DSSS (D3S)	DSSS	DSSS	DSSS	DSSS (D3S)	DSSS (D3S)
Packet-Reception Power Threshold(W)	7.33e-14	7.33e-14	7.33e-14	7.33e-14	7.33e-14	7.33e-14
Short Retry Limit	7	7	7	7	7	7
Long Retry Limit	4	4	4	4	4	4
AP Functionality	Active	Active	Active	Active	Active	Active
Buffer Size (bits)	64K	128k	256k	512k	1024K	2048K
Max-Receive Lifetime (Sec)	0.5	0.5	0.5	0.5	0.5	0.5
Large-Packet Processing	Drop	Drop	Drop	Drop	Drop	Drop

TABLE 3. GRID WLAN SCENARIO TABLE

Fragmentation Threshold (Bytes)	Load Intensities/Sources	Buffer Sizes (Bytes)	Scenarios/Cases
256	5	64K	CASE-1
512	10	128k	CASE-2
1024	15	256k	CASE-3
2048	20	512k	CASE-4
4096	30	1024k	CASE-5
8192	40	2048k	CASE-6

Subnet1
Subnet2
Subnet3
Subnet4
Subnet5
Subnet6
Subnet7

B. *Performance Evaluation*

The simulation was run accordingly and the statistics collected for various cases and analyzed. As claimed before, the proposed GRID WLAN hotspot is a high performance solution that is robust and scalable. The paper used a case-based scenario for various load intensities as well as

fragmentation threshold and buffer sizes as depicted in Table 3.0

To demonstrate the effects of Fragmentation Threshold, we employed the 5 node simulation scenarios with combinations of values for Fragmentation Threshold and RTS as well as other parameters. After the first case simulations, the throughput was collected for analysis. The throughput is the bit rate sent to the higher layer. It represents the rate of data successfully received from other stations. At the offered load from one node source to five node sources, it was shown that for a buffer size of 64 bytes for the FTP and HTTP traffic, the throughput had a slight deviation at 89% showing a significant network performance.

Fig 7: shows the media access delay for offered Load time against fragmentation threshold Fragmentation threshold is an important parameter that affects WLAN performance. It is used to improve the WLAN performance when the media error rate is high.

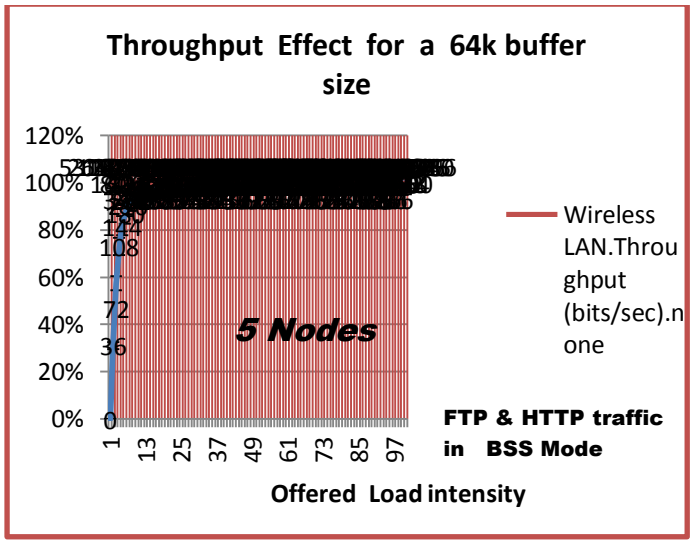


Figure 6: A plot of Offered Load Versus Processed Throughput for a 64k buffer size

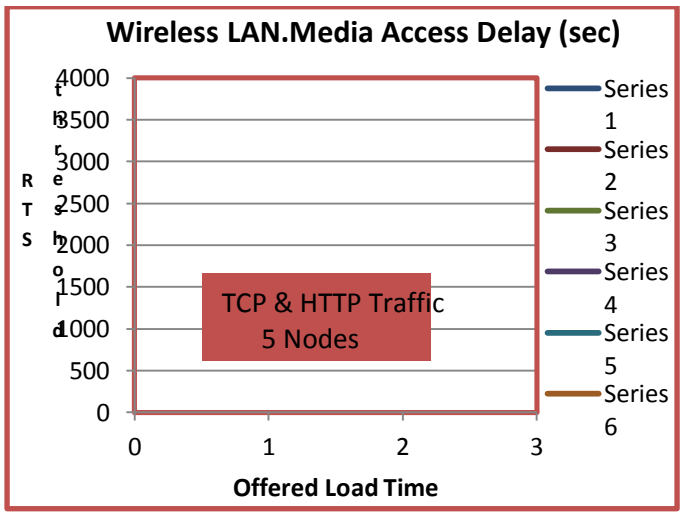


Figure 7: Offered Load time against fragmentation threshold

From Fig. 7, the effect of Fragmentation threshold on the WLAN performance was seen to be insignificant considering the offered load times. This could result from the selected threshold in the APs and client machines. For the first five simulation scenarios, parameters for Fragmentation Threshold are listed in Table 2. The simulation results indicate that for low bit error rates (2x10-5), various fragmentation thresholds (256 bytes, 512 bytes, or no fragmentation limit) have no significant effect on the WLAN performance.

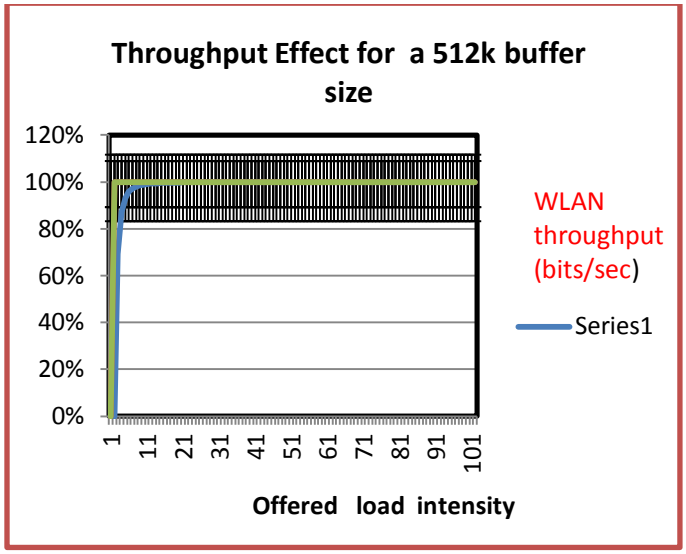


Figure 8: A plot of Offered Load Versus Processed Throughput for a 512k buffer size

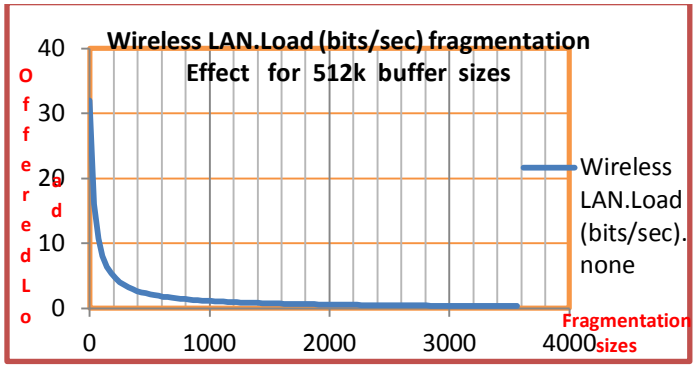


Figure 9: A plot of WLAN Load Against fragmentation threshold for a 512k buffer size.

Fig. 8 shows the throughput effect on offered load intensity for a 512k buffer size. From one to five workstations, there was a gradual exponential response until at a saturation point of about 87% , then it now begin to maintain a constant throughput with additional loads. For, every network, traffic optimization can only be achieved when the network throughput is significantly acceptable or high.

The simulation results in Fig. 9 illustrates that with an adaptive back-off mechanism, the offered load between WLAN stations can be greatly decreased with increase in the

fragmentation sizes while throughput can still maintain the same or achieve a slightly higher value. The reduction of WLAN load is important for power reduction in wireless devices. The Fragmentation threshold parameters used for the simulation run indicate that when the bit error rate.

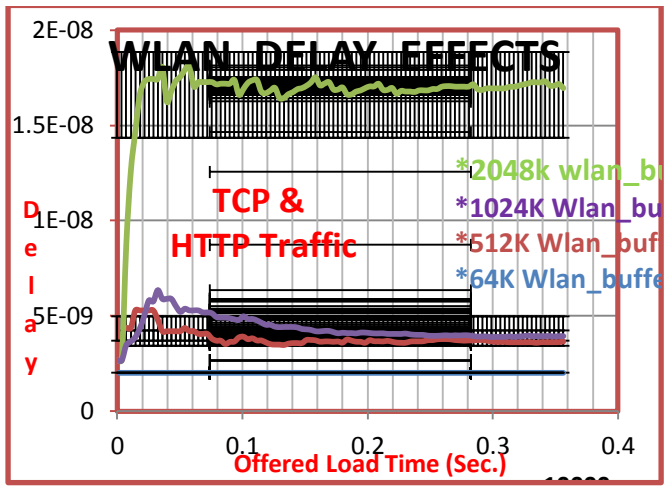


Figure 10: A Plot Of WLAN Delay Effects For 64k,512k,1024k,&2048k

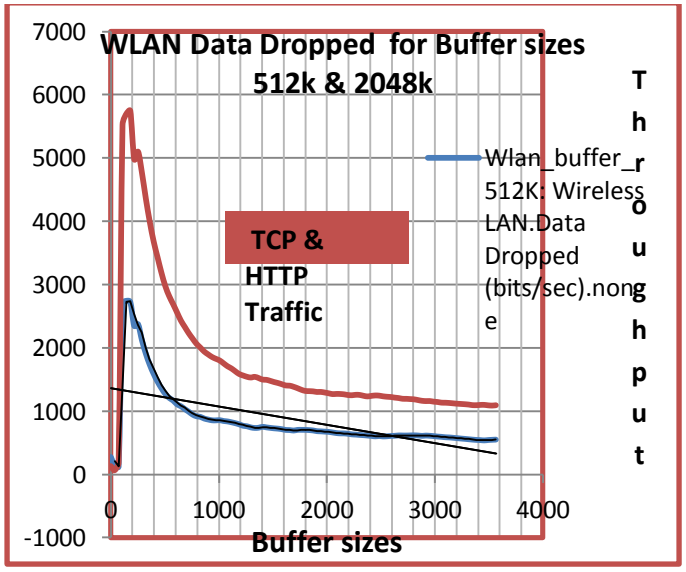


Figure 11: A plot of WLAN Data Dropped for 512K & 2048K buffer Sizes

During the simulations, the media access delay i n the four cases above was collected. The media access delay is the sum of queue and contention delays of data packets received by the WLAN MAC layer from the higher layer. For each packet, the delay is recorded when the packet is sent to the physical layer.

The simulation results (Fig.10) indicate that smaller buffer size values can decrease the average media access delay, while larger buffer sizes increases average media delay. Also the average media access delay is increased with increase in load intensities. This shows that performance of a wireless LAN is reduced with an increase in load intensity even when the buffer size is as well increased.

Fig. 11 shows simulated DCF throughput as a function of offered load for 5–40 wireless stations. Offered load in this context depicts the WLAN client nodes. First, we note that for the 10 and 20 stations, throughput is higher for 2048k buffer size compared with 512k buffer size. As offered load is increased, throughput grows linearly until a saturation point at which throughput ceases to increase; in fact, it decline. When the number of stations is large, throughput decrease and hence packet drop is inevitable. The drop in peak and saturation throughput as a function of the number of wireless stations is shown in Fig. 11.

At a moderate load of 16-20 wireless nodes, peak throughput decreases for buffer sizes 512k-2048k. Hence, for efficient network performance, the number of workstations should be carefully chosen and the AP device properly selected for high throughput in all cases.

VI. CONCLUSION

This research has presented a conceptual GRID WLAN hotspot model with performance analysis which can be further expanded to address the security challenges of the Wimax base station. The work investigates the effects of GRID WLAN APs buffer size distributions with respect to load intensities as well as fragmentation capacities. The analysis and measurement results from OPNET simulation shows that a careful selection of buffer sizes, fragmentation thresholds in GRID WLAN hotspot model can help to optimize network performance. Although some of these results are characteristic of the designed network, I can conclude that the GRID WLAN hotspot network performance is mainly based on the simplification of the traffic in addition to mobility and the buffer sizes.

REFERENCES

[1] Jiaqing Song and Ljiljana Trajkovic: Enhancements and performance evaluation of wireless local area networks.

[2] B. P. Crow, I.Widjaja, J. G. Kim, and P. T. Sakai. IEEE 802.11 wireless local area networks.IEEE communications Magazine, September 1997.

[3] S. Khurana, A. Kahol, S. K. S. Gupta and P. K. Srimani :Performance evaluation of distributed co-ordination function for IEEE 802.11 Wireless LAN protocol in presence of mobile and hidden terminals (Unpublished).

[4] LAN MAN Standards Committee of the IEE Computer Society, "Part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications," ANSI/ IEEE Standard 802.11, 1999 Edition.

[5] L. bononi, m. conti, and L. donatiello, "design and performance evaluation of a distributed contention control (dcc) mechanism for ieee 802.11 wireless local area networks," in proceedings of first Acm international workshop.

[6] C.-H. Ng, J. Chow, and Lj. Trajkovic, "Performance evaluation of the TCP over WLAN 802.11 with the snoop performance enhancing proxy," OPNETWORK 2002, Washington, DC, Aug. 2002.

[7] .S. A. Bawazir, S. H. Al-Sharaeh, "Performance of infrastructure mode wireless LAN access network based on OPNET Simulator" Department of Computer Science, Normal, AL 35762,USA

[8] P. Jarmo, "OPNET - Network Simulator", seminar presented at VTT Technical Research Centre of Finland, March, 2006

[9] J.Singh, Quality Of service in wireless Lan Using OPNET Modeller, Available at http://dspace.thapar.edu.

[10] M. S. Gast, 802.11 Wireless networks: the definitive guides, O'Reilly, 2002.

[11] A. Banchs, "Analysis of the delay distribution in 802.11 DCF: A step towards delay guarantees in WLANs," LNCS, vol. 3266, Sept. 2004, pp. 64–73.

[12] IEEE 802.11e/D13.0, Part 11, "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: medium access Control (MAC) enhancements for Quality of Service (QoS)," draft supp. to IEEE 802.11 std., Jan. 2005.

[13] R. Litjens et al., "Performance analysis of wireless LANs: an integrated packet/flow level approach," Proc. 18th Int'l. Teletraffic Cong., Sept. 2003, pp. 931–40.

[14] B. Liu, Z. Liu, and D. Towsley. On the capacity of hybrid wireless networks. In Proc. IEEE INFOCOM '03, pages 1543–1552, 2003.

[15] T. Hansen, P. Yalamanchili and H-W. Braun, "Wireless measurement and analysis on HPWREN", Proceedings of Passive and Active Measurement Workshop, Fort Collins, Co, pp. 222-229, March 2002.

[16] M. H.Manshaei and T. T. INRIA, "Simulation-based performance analysis of 802.11a Wireless LAN". Route des Lucioles, BP-93, 06902 Sophia-Antipolis Cedex, France ,2004.

[17] J. Song and L. Trajkovic, "Enhancements and performance evaluation of wireless LAN". Communication Networks Laboratory Simon Fraser University, Burnaby, BC, Canada.

[18] Kaur1, S. Vijay, S.C. Gupta, "Performance analysis and enhancement of IEEE 802.11 wireless LAN". Global Journal of Computer Science and Technology Vol. 9 Issue 5 (Ver 2.0),130 ,January 2010 .

An enhanced Scheme for Reducing Vertical handover latency

Mohammad Faisal, Muhammad Nawaz Khan
Department of Computing, Shaheed Zulfikar Ali Bhutto Institute of Science & Technology (SZABIST),
Islamabad, Pakistan.

Abstract- Authentication in vertical Hand over is a demanding research problem. Countless methods are commenced but all of them have insufficiencies in term of latency and packet loss. Standard handover schemes (MIPv4, MIPv6, FMIPv6, and HMIPv6) also practice these shortages when a quick handover is desirable in several genuine circumstances like MANETs, VANETs etc. This paper will evaluate the literature of the work done in past and present for undertaking the authentication concerns in vertical handover and will put down a basis for building the latency and packet loss more effective in such a huge shared situation that can produce to an extremely bulky level. This effort will mostly focus on the existing tendency in vertical handover mostly with the authentication, latency and packet loss issues.

Keywords: Vertical hand over, Authentication, FMIPv6, HMIPv6, reactive, proactive, latency and packet loss.

I. INTRODUCTION

Future generation heterogeneous wireless networks (FGHWNs) are facing with service continuity challenge for which vertical handoff is indispensible. So its assurance by means of rapid and efficient vertical handover that recognize service connection and seamless mobility maintaining the security and QoS is demand of the day [4]. Administered by diverse operators like WiMax,WiFi ,UMTS,GSM etc each in the run to achieve the highest quality and data rates, dynamically to choose for the always best connected network. Same technology networks switching are called horizontal while different technology networks switching are called vertical handover.

As shown:

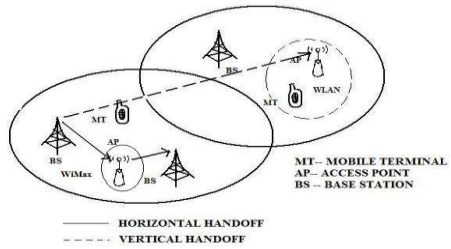


Figure 1. Vertical and Horizontal Handoff

[K.Savitha et al:1]

Four stages are accomplished during handover mechanism

- Handoff Initiation stage: Signal strength, link quality etc initiated the handover procedure.
- System discovery stage: Discovering the neighbor networks, sharing Quality of Service (QoS) information offered by these networks.
- Handover Decision stage: Comparing QoS of all available networks leads the user toward best network selection.
- Handoff Execution stage: Relinquishing old and establishing new connection and security services is done in this phase [1].

FGHWN are the result of overlapping of diverse wireless networks. The heterogeneity of FGHWN needs the realization of a vertical authentication method that trim down the handover delay whereas safeguarding the security perspective of the enduring communication. To uphold the continuing connections still after the home network is no longer accessible, vertical handover methods must be put into practice. Today, subscribers turn into additional challenges in conditions of omnipresent services availability and security. Furthermore, the installed fourth generation (4G) networks are illustrated by several wireless access networks expertise. To assure availability at the same time as preserving the security level, operators and service providers be supposed to harmonize and facilitate a mobile subscriber steadily using the services he is subscribed to regardless of his position and the access network expertise that is obtainable in his region, for which novel inter-networks roaming and vertical handover means ought to be cleared. To provide effective mobility with security all generation technologies (2G, 3G, and 4G) should be interoperable to facilitating the subscribers with ubiquitous wireless communication with high data rates. In case of loose coupling between WiMAX and WiFi networks operators have to describe supplementary mechanisms to implement network intelligence be capable to execute vertical handovers, as WiFi networks are incapable of management. In case of tight coupling, WiMax apparatus will be utilized to tackle subscribers' mobility and seamless connection shift, as they are capable of intelligent management [5].

II. LITERATURE REVIEW

In this paper the author's K.Savitha et al illustrated three different designs namely centralized, distributed and trusted

distributed for vertical handoff decision, by means of three parameters; Throughput, Decision Processing and End-End delay [1]. The strength of the paper is that, it validates that among the each outcome of all the three designs the trusted distributed for vertical handoff decision having best performance among all. The limitation of the paper is that, a lot of delay time both in decision processing and End-End transmission. Secondly, there are many inappropriate judgments attempted by the mobile nodes to decide their target network due to data flooding which leads to network congestion.

In this paper the author's Bao Guo et al compared and contrast two schemes (Encrypt-and-MAC, Encrypt- then-MAC) to achieve the confidentiality and integrity for the Short Message Services Networks [2]. The Encrypt-and-MAC scheme using CTR and CMAC modes while Encrypt-then-MAC scheme using EAX mode, both characterize by AES. Strength of the paper is that, output of each scheme in terms of operation time is almost similar but the Encrypt-then-MAC scheme offer more assurance, because appending MAC of the encrypted plaintext. Secondly Encrypt-then-MAC scheme also demonstrating best performance during online processing's. The limitation of paper is that, implementation of transmitter flow (SE/ES) will create the I/O obstruction if input data is increased. Secondly, Python 2.6 executables codes functionality is platform restricted.

In this work the author's Jakub Szefer et al suggested a flexible hardware mechanism (FPGA) that facilitates in time operations for encryption and hashing algorithms [3]. Strength of the paper is that, the Design is compatible with any portable devices and sensor nodes. Secondly, its implementation doesn't compromise its performance at the cost of its area and time [3]. Thirdly, the design can be used for other security services not just ciphering and hashing. Limitations of the paper are that, the mechanism of the FPGA policy is fixedly predefined. Secondly mechanism is adjustable with a limited and fixed quantity of Parameters. Thirdly, its efficiency is variable for cipher-to-cipher or hash-to-hash; maximum for whirlpool and block ciphers.

In this paper the author's wafaa bou diab et al proposed a new Seamless vertical handover solution for real-time data transfer with fast authentication. while analyzing the results with conventional schemes in terms of signaling cost (handover and authentication) and packet loss [4] It proved better results than its predecessor (IMS, MIPv6, FMIPv6) although quality and security services both maintained. Strength of the paper is that, it decreased the number of authentication messages, because both handover and authentication amalgamated in a single message, which reduced handoff latency as well. Resultantly the chances of packet loss also diminished. The limitation of the paper is that, this model is very sensitive to packet loss. Any loss of packet will lead to miss-synchronization, resultantly either to reinstate a new security association or roll back to the initial state.

In this paper the author's Neila KRICHENE et al discussed a vertical authentication method among the GSM, UMTS, WiFi and WiMAX technologies, regardless of any previous subscription to the visited network [5]. Strength of the paper is that, the authors proposed a global authentication protocol authorizing vertical handover between 4G architecture networks and proved the strength of the protocol against man-in-the-middle and denial of service attacks during vertical handover. The limitation of the paper is that, this protocol is only compatible with mesh topology 4G networks and does not bother for Quality of Service.

In this paper the author's Ahmed H. Zahran et al described a broker-based design for integrated heterogeneous networks to enhance the vertical handoff management, utilizing novel resource query method for Media Independent Handover [6]. The strength of the paper is that, it focuses on the decline in signaling load, MT configuration time, power consumption, user authentication delay, VHO delay, and the probability of VHO interruption, although guarantying seamless transitions as well. The limitation of the paper is that, almost all of its work is least cited which degrade its authenticity.

In this paper the author's Ali Al Shidhani et al put forward two re-authentication protocols (FUAR,LFR) comparing them to existing protocols(EAP-AKA, UMTS-AKA) in the 3G Home Networks [7]. The strength of the paper is that, It considerably spawning less signaling traffic and enduring less delay, accomplishing secured key management and mutual authentication, demanding no adjustments to the interworking architecture. The limitations of the paper is that,FAUR's security and performance depends upon the lifetime of security keys (TK, nCK, nIK).while security keys neither can be shared nor it can be reused, one time the session expired. The paper only focused on 3G networks

In this paper the author's Hoyeon Lee et al, have done efforts to enhance SIP, for vertical handover design (WiFi-to-UMTS) to reduce the IMS delay, comparing the results with conventional SIP [8].The strength of the paper is that, it enabling make-before-break vertical handover for IMS, successfully defined and operated two new headers in SIP, as sustaining for delay-sensitive real time applications, presenting compatibility with traditional. The limitation of the paper is that, it does not bother about the Layer 2 and 3 handover latency, secondly neglecting the forecasted troubles due to increase in header size.

In this paper the author's Jaeho Jo et al have focused on vertical handover via layer 2 and layer 3 signaling messages connecting mobile WiMAX and 3G networks [9].

The strength of the paper is that, it sort out L2 and L3 signaling messages as merging them resultantly limiting handover latency and UDP packet loss ratio while enhancing the TCP throughput. The limitation of the paper is that, it implemented the idea of predictive mode fast handover in which chances of failure are always open.

In this paper the author's chan-Kyu Han et al, assessed the signaling load, authentication procedures and compatibility of vertical handover in the EPS architecture networks [10]. The strength of the paper is that, it introduces the EPS network release 8 concentrating on security, secondly allowing random processes in authentication arrival, thirdly discovering new authentication generation processes other than routine, fourthly enhancing authentication signals as required. Limitations of the

paper are that, mandatory parameters like mobility organization, security strategies, and a variety of haphazard arrival processes are ignored in the current model.

In this paper the author's Liming Hou et al, introduced a pre-authentication design, based on the EAPTLS protocol, between WiFi and WiMAX hybrid networks, comprises of two stages; pre-authentication and re-authentication [11]. Strength of the paper is that, authentication delay is considerably reduced, while introducing pre-authentication stage. The design can be used for real time services as well. Limitation of the paper,EAPTLS is based on a public key infrastructure for its authentication, which reduces its portability.

In this paper the author Mario Marchese, integrate the interconnection and (Delay Tolerant Networks) DTN gateways architectures for wireless ubiquitous networks [12]. The strength of the paper is that, it shared successfully the functionalities of QoS mapping, and resource control from interconnection Gateways while extended delays and momentary link unavailability from DTN Gateways. The limitation of the paper is that, essential functionalities like well-organized mobility operates on broad territory coverage and guarantee of end-to-end data deliverance for elongated delay corridor and momentary link distraction are ignored.

In this paper the author Shih-Jung Wu specify (Host Identity Protocol) HIP-based vertical handover scheme via integrated architecture for seamless mobility of diverse wireless networks [13]. The strength of the paper is that, it used effectively HIP to tackle the problems of mobility and multi-homing while Diameter Protocol for registered users authentication. Limitations of the paper is that, as the local scope identifier (LSI) is a 32-bit identifier of host identity which is incompatible for IPv6 so collision probability is enviable, and its scope will be also controlled. Lastly results are not simulated or proved.

In this paper the author's Gabriele Tamea et al have provided an algorithm which makes its vertical handover decision on the basis of probability to avoid ping-pong effect [14]. Strength of the paper is that, the suggested algorithm evaluating the (WDP) wrong decision probability on the basis of: improving the collective goodput, and dropping of redundant and needless vertical handovers (PPE), was empowering the Mobile Terminal (MT) for the commencement and restriction of VHO that is why it is distinct as a Mobile Terminal-Controlled Handover scheme. Limitations of the paper is that, it didn't bother other estimations of good put , and analysis of its correlation at different intervals.

In this paper the author's Pedro J. Fern´andez Ruiz et al, portrayed a testbed setup installed over inter and intra technology vertical handover (WiFi,WiMAX and UMTS) to testify the secure mobility on vehicular networks [15]. Strength of the paper is that, it successfully experienced the authentication process on vehicular networks using IKEv2 and EAP3 for dynamic IP allocation and authentication respectively, exclusive of losing connectivity. Limitation of the paper is that, due to vehicular network there must be overlapping and no overlapping zones for which the selection and authentication procedures are not discussed.

III. CRITICAL ANALYSIS

TABLE 1.

Author	Working	Problems	Solution
K.Savitha et al	Validates that the trusted distributed for vertical handoff decision having best performance among all.	Delay time both in decision processing, End-End transmission, many inappropriate judgments which lead to network congestion.	By replacing the reactive routing protocol, Ad hoc On Demand Distance Vector (AODV) with a proactive routing protocol like, Highly Dynamic Destination-Sequenced Distance Vector routing protocol (DSDV), to reduce the delay time.
Bao Guo et al	Encrypt-then-MAC scheme offer more assurance, because appending MAC of the encrypted plaintext.	Implementation of transmitter flow (SE/ES) will create the I/O obstruction if input data is increased. Python 2.6 executables codes functionality is platform restricted.	If we implement MAC-then-encrypt scheme, will enhance security services. And if python 2.6 is replaced by PERL or ruby programming language then the platform independence issue will also be resolved
Jakub Szefer et al	The Design is compatible with any portable devices and sensor nodes. Its implementation doesn't compromise its performance at the cost of its area and time. The design can be used for other security services not just ciphering and hashing.	The mechanism of the FPGA policy is fixedly predefined. Mechanism is adjustable with a limited and fixed quantity of Parameters. Its efficiency is variable for cipher-to-cipher or hash-to-hash; maximum for whirlpool and block ciphers.	We can design FPGA user defined while assigning unlimited parameters according to the need and situation dynamically.
wafaa bou diab et al	It decreased the number of authentication messages, because both handover and authentication amalgamated in a single message, which reduced handoff latency as well. Resultantly the chances of packet loss also diminished.	This model is very sensitive to packet loss. Any loss of packet will lead to miss-synchronization, resultantly either to reinstate a new security association or roll back to the initial state.	A mechanism can be introduced so that, on each and every mobile node SCID (Session Context ID) routing table should be fully updated with full header packets rather only destination IP addresses.
Ali Al Shidhani et al	It considerably spawning less signaling traffic and enduring less delay, accomplishing secured key management and mutual authentication,	FAUR's security and performance depends upon the lifetime of security keys (TK,nCK,,nIK).while security keys	Consequence of additional security keys for each session is required to be investigated. Proper confirmation of the security and

	demanding no adjustments to the interworking architecture.	neither can be shared nor it can be reused, one time the session expired. The paper only focused on 3G networks	performance of FUAR will be carried out with constraint of life time.
Hoyeon Lee et al	It enabling make-before-break vertical handover for IMS, successfully defined and operated two new headers slots in SIP.	It does not bother about the Layer 2 and 3 handover latency, secondly neglecting the forecasted troubles due to increase in header size.	Large bandwidth allocation to solve the header size problem.
Jaeho Jo et al	It sort out L2 and L3 signaling messages as merging them resultantly limiting handover latency and UDP packet loss ratio while enhancing the TCP throughput.	It implemented the idea of predictive mode fast handover in which chances of failure are always open.	predictive mode can be replaced with reactive mode fast handover
chan-Kyu Han et al	Allowing random processes in authentication arrival, discovering new authentication generation processes other than routine, enhancing authentication signals as required.	Parameters like mobility organization, security strategies, and a variety of haphazard arrival processes are ignored in the current model.	Several features like haphazard walk mobility model, associations with every authentication activation and different random process generation are to be examined for the compatibility with the real world scenarios.
Liming Hou et al	Authentication delay is considerably reduced, while introducing pre-authentication stage. The design can be used for real time services as well.	EAPTLS is based on a public key infrastructure for its authentication, which reduces its portability.	we can replace the protocol by any other of EAP family member like EAP-MD5, EAP-OTP, EAP-GTC, EAPTLS and EAP-SIM etc.
Mario Marchese	It shared successfully the functionalities of QoS mapping, and resource control from interconnection Gateways while extended delays and momentary link unavailability from DTN Gateways.	Essential functionalities like well-organized mobility operates on broad territory coverage and guarantee of end-to-end data deliverance for elongated delay corridor and momentary link distraction are ignored.	A novel illumination relating to architectures and protocols is obligatory
ih-Jung Wu	It used effectively HIP to tackle the problems of mobility and multi-homing while Diameter Protocol for registered users authentication.	As the local scope identifier (LSI) is a 32-bit identifier of host identity which is incompatible for IPv6 so collision probability is enviable, and its scope will be also controlled. Lastly results are not simulated or proved.	Its results should be simulated so that to assess its presentation and should also be compared with its predecessor work.
Gabriele Tamea et althe	Improving the collective good put, and dropping of redundant and needless vertical handovers (PPE),empowering the Mobile Terminal (MT) for the commencement and restriction of VHO	It didn't bother other estimations of good put, and analysis of its correlation at different intervals.	To include the other goodput parameters and enhance correlation among them.
Pedro J. Fern´andez Ruiz et al	it successfully experienced the authentication process on vehicular networks using IKEv2 and EAP3 for dynamic IP allocation and authentication respectively, exclusive of losing connectivity.	As it is vehicular network there must be overlapping and no overlapping zones, for which the selection and Authentication procedures are not discussed.	To set pre-established security and mobility policy that will allow the user to opt the appropriate interface in each and every situation.

IV. PROPOSED SOLUTION

The above observations reveal that many schemes were introduced but all of them have deficiencies. Hand over process takes time in address acquisition and on time calculations which increases latency and hence leads to packet loss. For fast and smooth hand over between homogeneous and heterogeneous networks, a modified and fast hand over mechanism always needed. Hand over is more critical in some conditions like in ad hoc networks such as MANETs and VANETs. In ad hoc networks the hand off need very few time for changing point of attachment to the access point and with routers as mobile nodes move very fast. The horizontal handover between different network such as WiFi, WIMax, UMTS also leads latency and packet loss. Latency created in the hand over process because the difference in the band width. When a node leaves from one network and inters into the premises of another network such as from WiFi to WiMax, the difference between band widths leads latency. The queue becomes full at the bridge of the network and waiting packets feel delay in term of latency and as a result all this lead to packet loss. In horizontal hand over on time calculations also takes timing to change from point to point.

The Standard IEEE 802.11 Handover process of consist into following six phases: triggering, discovery, authentication, association, IP address acquisition, and home agent (HA) registration [16]. The first four phases related to the data link layer and called layer two (L2) while the last two phases done it the network layer and called layer three (L3) handover [17]. In the whole process there is on time calculation and awake the moment of the mobile node. The moment of the mobile node from one network premises to another continuously monitor from its Received Signal Strength Indicator (RSSI) value. When RSSI value decrease in area the mobile node start scanning for other network having better RSSI value. IEEE 802.11 had over two types of scan modes are define: active scan and passive scan. In active scanning, the MN transmits a probe request massage on a channel and then waits for a while. If no response from other side, mobile node continues to sends a probe massage on the next channel and wait for response from other access router. While in passive scanning, the mobile node sequentially listens to beacons on different channels instead of transmitting massages for request. When mobile node listen all the channel, it choose a channel with better RSSI value and start massages to shift from one access router to another.

In our proposed scheme passive scanning mode, the mobile node start hand over procedure rather network. Nearly all parameters are set prior starting the connection establishment of new connection with another access router. While some very necessary calculation still performed at time of changing access routers.

In the proposed scheme total focus is on proactive mechanism rather than reactive all time. Proactive means much of the work been done by the system before actually hand over been starting. First, when mobile node experience weak RSSI value, it starts searching for better RSSI network value. The mobile node starts discovery and section for new access routers. Once the mobile is performed authentication to entering at premises of the new network. Then no need for re-authentication, when changing from one access point to other (horizontal handover). At the association process the mobile node define itself prior to association and almost all massages shared between new router and mobile node.

Therefore for a mobile node tack very little time to associate with access router. After association the mobile node trying to get new IP address. Two variations either Router Advertisement messages are periodically broadcasted by new router or the mobile node sends Router Solicitation messages to new router to obtain the Router Advertisement messages. Care of Address (CoA) has been assign to mobile node for time being, duplication address detection (DAD) process ensure unique addresses in the same network premises. When a mobile node obtains its CoA in new network then mobile node informs it's Home Agent about the new CoA by sending a binding update message to home agent. The home agent also sends a binding acknowledgement massages to the mobile node to complete the process [18].

The standard scheme shows that latency turn outs because of the on board calculation. Nearly all schemes needs on time calculations for completing the hand over procedure. But these calculations can be decrease by dividing the whole process into two parameters. Some necessary parameters are still calculated on time but most of the parameters should set before the actual process being start.

The proactive mechanism reduces the delay in term of reducing latency. Reactive mechanisms are applied when some parameters need some on line calculating values otherwise the proactive mechanism. Proactive mechanisms reduce the hand over latency by applying pre-define parameters values. Proactive parameters including pre-define calculations by the mobile node for smooth roaming in the same network with different access points or roaming between different routers of the different networks, pre-define pool of address for mobile nodes which decrease on time calculation for address acquisition and duplication address detection.

V. Conclusion And Future Work

While evaluating all the above mentioned work it can be concluded that almost all of the time reactive protocols are used when a node shift from one entrance point to a new one, while executing estimations in terms of handover and authentication. As handover massages are exchanged, the actual data packets experiences extensive average delay, which amplify latency and lastly leads to packet loss in few cases. In our proposed scheme, a few parameters will be pre-defined, which will assist in calculation at hand over process. The proposed scheme will be helpful as it is based on preceding standard approaches with improvements. The proposed elucidation will be a hybrid approach of reactive and proactive.

Aknowledgment

We are very thankful to Almighty Allah; whose grace and blessed mercy enabled us to complete this work with full devotion and legitimacy. We are grateful to Mr. Mohammad Ibrahim, Lecturer in Computer Science, Virtual University of Pakistan, for their invaluable support and guidance throughout this research work. We also want to thank our friends and family for their encouragement; without whose support we could not have lived through this dream of ours.

References

[1] K.Savitha and Dr.C.Chandrasekar "VERTICAL HANDOFF DECISION SCHEMES IN HETEROGENEOUS WIRELESS NETWORK USING MADM", JGRCS 1(5), December 2010, 16-20.

[2] Bao Guo and William Emmanuel Yu "Comparison between Encrypt-and-MAC Composite (CMAC CTR) and Encrypt-then-MAC Composite (AES EAX) Modes of Operation in Cryptography Systems for Use in SMS-based Secure Transmission". Proceedings of the international multi conference of Engineers and Computer Scientists 2011 Vol I, IMECS 2011 March 16-18, 2011, Hong Kong.

[3] Jakub Szefer, Yu-Yuan Chen and Ruby B. Lee "General-Purpose FPGA Platforms for Efficient Encryption and Hashing"

[4] Wafaa Bou Diab and Samir TohmeSeamless "Handover and Security Solution for Real-Time Services" 2009 11th IEEE International Symposium on Multimedia 978-0-7695-3890-7/09 $26.00 © 2009 IEEE

[5] Neila KRICHENE and Noureddine BOUDRIGA "Securing roaming and vertical handover in fourth generation networks "Third International Conference on Network and System Security" 978-0-7695-3838-9/09 $26.00 © 2009 IEEE

[6] Ahmed H. Zahran and Cormac J. Sreenan "Extended Handover Keying and Modified IEEE 802.21 Resource Query Approach for Improving Vertical Handoff Performance" 978-1-4244-8704-2/11/$26.00 ©2011 IEEE

[7] Ali Al Shidhani and Victor C. M. Leung "Reducing Re-authentication Delays during UMTS-WLAN Vertical Handovers "978-1-4244-2644-7/08/$25.00 ©2008 IEEE

[8] Hoyeon Lee, Bongkyo Moon, and A. H. Aghvami "Enhanced SIP for Reducing IMS Delay under WiFi-to-UMTS Handover Scenario The Second International Conference on Next Generation Mobile Applications, Services, and Technologies "978-0-7695-3333-9 /08 $25.00 © 2008 IEEE

[9] Jaeho Jo and Jinsung Cho "A Cross-layer Vertical Handover between Mobile WiMAX and 3G Networks"978-1-4244-2202-9/08/$25.00 © 2008 IEEE

[10] Chan-Kyu Han, Hyoung-Kee Choi,Jung Woo Baek, Ho Woo Lee "Evaluation of Authentication Signaling Loads in 3GPP LTE/SAE Networks"2009 IEEE 34th Conference on Local Computer Networks (LCN 2009) Zürich, Switzerland; 20-23 October 2009 978-1-4244-4487-8/09/$25.00 ©2009 IEEE [11] LimingHou and Kai X.

[11] Miao "A Pre-authentication Architecture in WiFi & WiMAX Integrated System.

[12] Mario Marchese, "Wireless Pervasive Networks for Safety Operations and Secure Transportations" IEEE 2010 International Symposium on Wireless Pervasive Computing (ISWPC)

[13] Shih-Jung Wu "A New Integrated Mobile Architecture for Heterogeneous Wireless Networks 2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing" 2010 IEEE.

[14] Gabriele Tamea, Anna Maria Vegni, Tiziano Inzerilli, Roberto Cusani "A Probability based Vertical Handover Approach to Prevent Ping-Pong Effect" ISWCS 2009.

[15] Pedro J. Fern´andez Ruiz Cristian A. Nieto Guerra Antonio F. G´omez Skarmeta "Deployment of a Secure Wireless Infrastructure oriented to Vehicular Networks 2010 24th IEEE International Conference on Advanced Information Networking and Applications "

[16] Daehan Kwak, Jeonghoon Mo, Moonsoo Kang, "Investigation of Handoffs for IEEE 802.11 Networks in Vehicular Environment".

[17] Yao-Tung Chang, Jen-Wen Ding, Jen-Wen Ding, Ing-Yi Chen. "A Survey of Handoff Schemes for Vehicular Ad-Hoc Networks".

[18] R. Koodli, Ed. "Mobile IPv6 Fast Handovers", Request for Comments: 5268 Starent Networks Obsoletes: 4068 June 2008.

AUTHORS PROFILE

Mohammad Faisal received his B.S. degree in Computer Science (Hons) from University of Malakand at Chakdara, Dir lower, KPK, Pakistan. He is currently pursuing his MS in Information Security management system at Shaheed Zulfiqar Ali Bhutto Institute of Science & Technology (ZABIST) Islamabad Pakistan. Since November 2006, he is working as Network Engineer in Pakistan Revenue Automation Limited (PRAL), Motorway Department.

His research areas focuses on Handoff, forensics, cryptography and networks (SENSOR, WIRELSS, MANETS) security. He intended to precede his studies (PhD) in any of the above mentioned fields.

Muhammad Nawaz Khan is lecturer in Computer Science in Govt. College of Management Science. In 2008, he received Silver Medal in B.S. (Hons) degree in Computer Science from University of Malakand, K.P.K. Pakistan. He partially completed MS in Computer Communication Security at School of Electrical Engineering & Computer Science (SEECS), National University of Science & Technology (NUST) Islamabad, Pakistan. In 2010, he worked as a Research Assistant in a project on "Distributed Computing" supported by Higher Education Commission of Pakistan. Currently he is working as Research Assistant at Shaheed Zulfikar Ali Bhutto Institute of Science & Technology Islamabad. His research is focused on Computer Information Security especially Computer Communication Security. He has also showed keen interest in Ad-hoc networks (MANETs, VANETs), wireless communications security and security related issues in distributed computing. He intended to proceed his studies (PhD) in any of the above mentioned fields.

A Feasible Rural Education System

Lincy Meera Mathews
Assistant Prof, Department of ISE,MSRIT
M S R Nagar, Bangalore,
Karnataka, India

Dr Bandaru Rama Krishna Rao
Senior Professor, School of Information Technology and Engineering, Vellore Institute of Technology University, Vellore, Tamil Nadu

***Abstract*— The education system in rural and semi-rural areas of developing and underdeveloped countries are facing many challenges. The limited accessibility and challenges to the education are attributed mainly to political, economic and social issues of these underdeveloped countries. We propose a "Feasible Rural Education System (FRES)" based on Ontology and supported by Cloud to enhance the accessibility to education in rural areas. The system has been proposed incorporating the FOSS approach.**

Keywords- FRES; Education; FOSS; Semantic Web; Ontology; Natural language processing; Cloud computing.

I. INTRODUCTION

In developing countries like India, the rural and village schools form 87% of the total schools, of which 90% are run by the government with financial aid and rest of the schools are unaided [1]. These schools offer education in four phases: *Primary, Upper-primary, Secondary and the Higher Secondary.*

Though the literacy rates in developing countries have been steadily increasing in higher education system (which includes educational and vocational training), the same is not noticed in the Primary and Upper-primary education in rural and semi-urban areas. Various initiatives had been taken up by the Government of India to support Education in rural areas by bringing in various technologies and through the usage of internet. The role of Information and Communications Technologies in education can be dispersed into manners [6].

- Alternative instructional delivery systems such as radio, educational TV and audio visual communication.
- Computers and computer based system for instructional delivery and management; use of multimedia and internet/web based education.

As an initiative for the former mode of communication, EDUSAT was launched in September 2004, at a cost of USD 20 million. India's first Education satellite was dedicated. However it has failed to provide any speak able impact in rural areas. The problems attributed to this were previously recorded lectures, non-interactive nature of sessions, not formed for as per students learning capability.

As for the latter, the government has initiated an Internet Based Continuing Education. The components are comprised of online learning materials, online Academic counselling, online assignments and projects and online collaboration. To help back all the state government ICT education initiatives, in May 20, 2006, the government of India, Ministry of HRD, Department of Secondary and Higher Education had issued an order for the implementation of Broadband Connectivity in all secondary schools. To accelerate the progress of education in rural areas, we take the help of the free and open source software (FOSS) movement. The FOSS has originally begun as an intellectual movement for development, modification, ownership and redistribution of software managed by a community of programmers. We suggest that the principles of FOSS namely transparency, collaboration, openness and co-ownership can be introduced in our education system in providing accurate, accessible and free education resources.

II. LACUANES IN RURAL EDUCATION SYSTEM

A typical rural education system contains three main components namely: *Teacher*, *Student* and *Infrastructure* [2].

A. Student

In rural scenario, student speaks his native language more fluently, demanding communication for learning in local language. A student in rural area faces more challenges in daily life as compared to his counterpart in urban areas, to name few (i) lack of commutation to schools situated far off from villages, (ii) playing in streets, want of setup area meant for play (iii) studying under street lamps, want of electricity at home, (iv) lack of good curriculum or educational material, (v) lack of full-pledged laboratory or library facilities etc. In essence, priority for rural student is to survive for the day rather than look and work for a better future. The governments of developing nations aim to bring quality education to these rural students so that they could compete with their urban counterpart. However, this becomes difficult, since, the urban schools have access to top-class teachers, laboratory and library facilities. Therefore, the needs to be addressed for rural students are as follows:

- An interesting and challenging representation of free education material, so that attention of the student holds.
- Communication in learner's language.
- Limited dependence on teacher's instruction and on non-availability of other modes of learning.
- More interaction with students of other schools of the same curriculum.

- Up to date and accurate representation of curriculum

B. Teacher

A teacher plays the most impacting role in the rural education system. A teacher forms the root of the system on which students, learning resources and technologies are dependent. Without the teacher, the student stands to gain a very limited knowledge. The learning resources and the new technologies cease to play any role if teachers do not use the facilities. The learning resources will not be reviewed if the teacher fails play an active role in the education system. The more effective the teacher becomes, the better the quality of learning by students. However, the challenges faced by teachers of rural and semi urban in terms where technology can play roles are as follows:

- The large gap between rural and urban technology in facilitating teaching and learning process.
- Limited availability of facilities like study materials, language friendly multimedia.
- Very less or non- availability of resources for accelerated and quality learning.
- Huge cost involved in procuring and maintenance of infrastructure- mainly software and hardware.
- Lack of communication between schools handling the same curriculum, hence shortage of updated information and mentoring programs.

C. Infrastructure

The basic infrastructure required can be divided into consumables products and non-consumable products. The consumable products mainly are stationery items such as exercise books, pens, pencils etc and the non-consumable materials are the curriculum, learning aids, visual aids and technological equipment's. The dearth of non-consumable products can be met by present day technologies. The various factors that need to be addressed in this context are as follows:

- Free availability of resource materials in the form of text, video or audio classes.
- An effective teaching guide to support the teachers and learners.
- Challenge based or game based learning curriculum that can hold the attention of learners.
- An effective educator system that can take care of mentoring and supporting of rural educators.
- Availability of technology (software and hardware) without the maintenance and cost issues.
- A user friendly interface preferably in the local language of the user
- Low cost systems and lab facilities.

Various other factors like political, male to female ratio, economic reasons exist [3]. However factors where technology can play a role and help alleviate the hurdles should be addressed in this current era.

III. FOSS IMPACT IN EDUCATION

The Free and Open source movement was begun by a community of programmers to remove or decrease restrictions on collaborative development, distribution, ownership and redistribution of software [5]. This however got noticed and the principle is being applied by various researchers and academic communities in various domains. The values that the FOSS movement promotes are openness, transparent collaboration and co-ownership. All these values need to be adapted into the education scenario.

A. Openness

Openness in FOSS environment is to have open access to the source code. In education, openness leads to free access to curriculum and pedagogy. Currently information related to curriculum and pedagogy is available only to the selected few. Though governments provide may Open-Courseware resources such as Indira Gandhi National Open University (IGNOU), NPTEL (National Programme on Technology Enhanced Learning) and the CEC (Centre for Education and Communication) repository in India, they are not entitled for changes by others to suit their curriculum, and faces the following challenges:

- Information retrieval works on the basis of queries. The relevance of data is directly proportional to the completeness of a query. The information retrieved is dependent on the learner's query.
- Trust factor of information retrieved is yet a widely researched area. The user has to decide on the accuracy and completeness of information.
- There is no guidance in the flow of information or in the method of learning process. The user or reader must decide on the flow of information to be read as there is no connection between the information. If the student begins on a certain topic, he must be guided to its prerequisite information.
- No guidance is available when the student scours through lengthy material or tutorial. The online tutor is absent. She/he does not have any interaction with subject matter expert or any answering system for his queries & doubts.
- The maturity of the user is also of utmost importance. Her/his understandability of the instruction, level of language proficiency, and the grasping power of an individual all plays a role in effectiveness of the information.

B. Collaborative development

This in terms of software relates to the development of code by several professionals interested in the tool or software. They decide to do changes so that it helps in their application.

With education, collaborative development has been interpreted as a transparent process for formulation of educational policy, syllabi and content related to the curriculum or course. The hurdle here lies with the transparency process. The limitations of applicability of

transparent collaboration in education may be attributed to the following reasons.

- There exist no common policy and decision makers. There are different methods of learning tools. Some schools believe in sending work home. Some work on completing at school premises
- There is always a change in curriculum delivery for secondary education. Textbook variation, selection of portions and time of completion causes difference in the delivery of curriculum.
- Many universities have created their digital repositories for the benefit of students. This however has been a good initiative for the students of the respective schools. All different kinds of tools are being used for updating like PPT, text, audio, video or word. Updating of resources requires understand ability of the underlying software. It poses restriction to authors of various domains.

C. Co-ownership

In FOSS, Software can be freely distributed without licensing fees. In education the main stakeholders are the students, teachers, parents and the institution. Ownership relates to the freedom of access and an equal influence over all intellectual artefacts used in the education system. Here again the access and empowerment of curriculum lies in the hands of the select few, who are eligible for the resources and not those few who are in search of knowledge. Faculties do not have access to better materials from various guides. They lack support from domain proficient members and no knowledgeable tools. To bring co ownership to education would mean able to modify and reuse the repositories.

Creativity also has been the other important aspect where the users who have laid the FOSS foundation have encouraged. In all its forms, creativity is the spirit in FOSS. This also must be followed in areas of education too. Tools that enable creativity must be created and used to provide various challenged based learning. The user plays a very important role in creativity. Incorporating the Bloom's taxonomy is one of the major hurdles that are faced by curriculum developers.

IV. Semantic web Based Education System

A. Use Case Scenario

Understanding the shortcomings and the difficulties of a student from a rural background is important aspect to consider in the teaching learning process. Since the system can be categorized as an e-learning space, all the characteristics of classroom learning if not more must be incorporated into the system. The rural education system has basically three entities: *Domain model*, *student model* and *Facilitator's mode*l. Three requirements of the Semantic Web Education System for three different entities are discussed below.

1) Student requirement:

- Student needs a guide as he sits through the learning system. She needs an interactive system that guides him through his learning process
- Her preference of the learning style :Audio, Video, Text or Incorporation of all formats
- The dynamic flow of learning material that suits his understanding capability. She does not have to go through the entire material but that suits his need of knowledge
- Learning material presented depending upon his cognitive learning style
- A good and friendly user interface
- Student data to be stored with respect to his interest, his learning mode etc
- High quality information
- Student community

2) Domain requirement

- Storing of content's metadata or its relevance in the learning repository
- Structure definition: Organization of each content material or the learning sequence of the learning material
- Format of the resource material like video. Website , Audio etc
- Reusability of the content
- Dynamic updating of the content
- Quality assurance
- Automatic of display of Content depending on user's input

3) Facilitator's requirement:

- Uploading of resource material in any format
- Updating of content
- Questionnaire in relevance with the content
- Evaluation procedures
- Guiding of students online or offline
- A secured and an user friendly interface

B. Ontology Support

Ontology can be used for knowledge representation with respect to each domain [4]. Domain ontology has its advantages. Firstly it allows reuse, updating and sharing of ontology between ontology friendly environments. Secondly it has its formal structure which makes it easy to extract and obtain knowledge representations. The courses or resource material can be divided into *declarative* or procedural knowledge. The declarative knowledge represents the factual and conceptual knowledge while the procedural knowledge represents action sequences and problem solving procedures. The objective behind using ontology is to facilitate a structured knowledge which helps facilitate reasoning and help provide the best learning session as per learners need.

We present here the architecture (Figure 1) of ontology supported education system. The ontology supported web based education system has basically three components. They are the facilitator, learner and the data repository. The data repository forms the ontology supported data for the student and the learning documents. The facilitator's role is to support in content design and providing student support. The learners form the student as well as faculties of rural schools. They are the actual beneficiaries of the system. The web based system is then deployed on to a dedicated cloud for education purpose only. The web based application is in the form of centralized web-based applications. One instance of an application hosted in the cloud is cheaper and easier to manage. Upgrading of the application which is to be done often due to development of educational content has to be done often.

To define the ontology, OWL-DL (Protégé Editor has been used. The curriculum developed was included for a single course. The proof of concept will be deployed in Amazon, where it has its Elastic Compute Cloud (EC2), a web service that provides cloud-based resizable computing capacity for application developers.

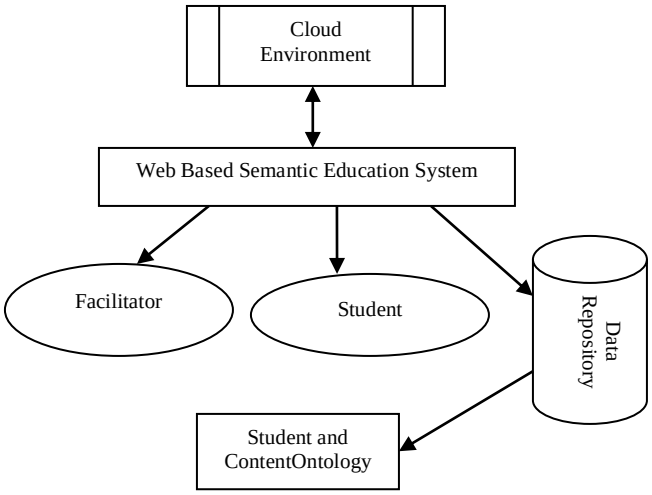


Figure 1: Architecture of Rural Education System

1) Domain Model

The courses are being represented in the form of ontology. The ontology as shown in figure 2 has been organized as follows. The definition class has been subdivided into mainly three sub classes. The definition class is being defined by the source, Content author, last_modified_date and the learning objectives. The definition class exists for contents updated into the domain model. The learning objectives compulsorily follow the objectives as per Washington accord. The source of content will be updated as a pair value of Source and contented. The content_ID acts as the sequence no which denotes the sequence no in which the content is to be presented.

The format class is linked by its various subclasses by the relationship of "for_a_particular". The individual contents could be an image, audio, video or text. The relationship between these various subclasses follows an OR relationship. The contents can be of more than one format but not necessarily all formats. The text data can be further categorized as PPT, .doc or Pdf formats. The PPT might be supported with audio also.

The structure denotes the actual content and details for rendering of content. The subclasses are the content, FAQ's, and test. The content is further divided by its purpose and aim. It can be classified as definition, principals, advantages, application etc. The content is also rated depending on its difficulty level. The content can be rated as topics that come under surface learning and deep learning. The surface learning can be described as the using low level cognitive skills and minimum effort to complete the course requirements. The deep learning involves understanding, engaging in higher level cognitive skills. The student should be able to think conceptually about a topic. The content is being identified by its content _id which will be given as input for the sequence _no subclass. The sequence _no subclass will be related to the prerequisite subclass. The prerequisite states that content to be mastered before the actual content is to be read. The test subclass relates to the content that is to be used to evaluate the students. The evaluation content could be objective or subjective type of questions. The structure class also maintains a FAQ's. The FAQ's will be referred when a student poses a question. The question will be matched against the FAQ pool.

OWL-DL (Protégé Editor) has been used for defining the ontology .The document ontology is being represented as an Onto Graph. Figure 3 states the object properties between various classes. The relationships indicate the functional dependence between various classes. All classes have direct dependence to its subclasses. However certain classes have dependencies to other disjoint classes. One of the functional dependence indicated is the prerequisite. This states that content will have a prerequisite. The inverse relationship is built automatically. In the above example, a prerequisite exists for all content.

2) Student Ontology

The lack of physical educator must be compensated by the system. Said in other words, the system must understand the need of the students and present the course that suits his learning style. The student background information is stored to help navigate and select the right learning style for the student. The profile includes his pre schooling, his language competency, his reading skills, his attention span etc. The preference of his learning style is also recorded for the benefit of the student. The different learning style can be categorized into visual learners, auditory learners, reading /writing learners, and kinaesthetic learners.

The cognitive learning style defines the cognitive characteristics that the student possesses. The Cognitive characteristics can be categorized as brainstorming followed by session or vice versa. It is believed that certain students need facts as to why certain content must be mastered. The evaluation procedure is recorded in the subclass in evaluation metric.

The metric store scores as against to each content for each student. As they complete the test in accordance to the content referred, the metric is updated. The description of the individual can be specified by using the individual description view.

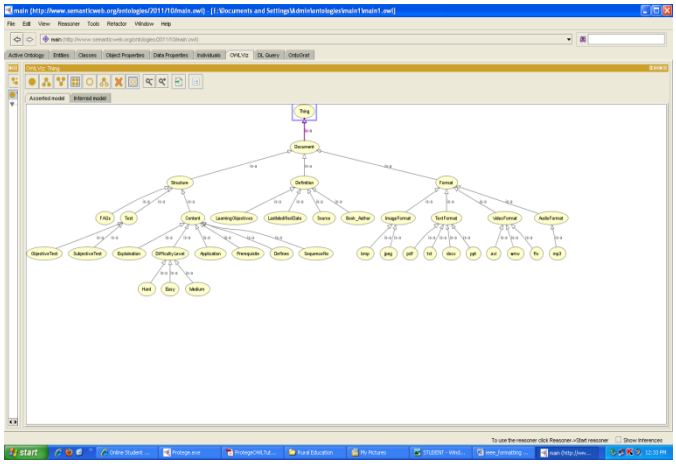


Fig. 2 OWL Viz representation of Document ontology

Domain	Object Property	Range
Definition	Depends_on	Content
Content	Requires	Prerequisite
Sequence no	Is_based_on	Content
Format	For_a_particular	Content
Test	Evaluation	Content
FAQ's	For_each	Content

Fig. 3 Relationship between various classes

The year to which he belongs can be added. Restrictions can be added by selecting the property assertion view. For example, he belongs to any year but value for year must be less than four. Using the equivalent class tab, conditions can be asserted. For example: Student and belongsto some integer [,="4"^^integer] hasdetail some profile. The subclass feedback provides the mentors or facilitators a timely statement so that the facilitators can improve the course. The student ontology has been shown in Figure 4. The relationship defined between the various subclasses has been shown in figure 5.

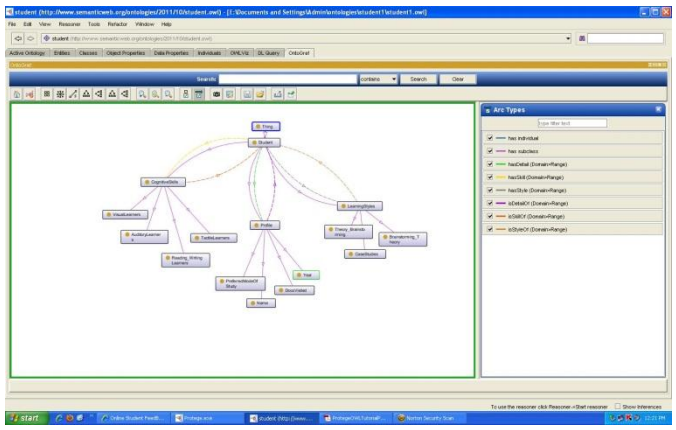


Fig. 4 Onto Graph representation of Student ontology

Domain	Object Property	Range
Profile	Has _skill	Cognitive Skills
Profile	Has_style	Learning Style
Profile	Has_some	feedback

Fig. 5 Onto Graph representation of Student ontology

3) Facilitators Model

Facilitators should be able to enter any of the content in any formats as per convenience. The tutorials can be used by faculties also for their benefit. The tutor when online will have the capability of clearing the doubts raised by students. The questions with its answers will enter the GAQ's pool. The domain specialized author will be authorized to verify and rank the content. The evaluation metric can be for assessment of student's performance. It also indicates to which level of learning the student has progressed.

A domain expert is involved to develop the course into various modules. With those various modules, ontology has to be constructed. Each module or content must be mapped with the learning objectives. The FAQ's must be evaluated for it content by the domain expert.

C. *Design of Flow*

Student:

- As a student logs onto his system, his profile will be considered.
- As he request for certain content, his preferred learning style and cognitive style will be taken into account.
- Depending on his completion of previous sessions, topics will be presented to the user.
- Brainstorming session will be followed by content if his preferred style of cognitive learning is concrete generic type
- If questions are asked, he can directly pose the question to the facilitator
- If facilitators are offline, the question will be matched to the existing FAQ's

Facilitator

- The author can upload and modify the existing content
- Answers to questions of users should be completed
- Domain experts must be able to rank the content and verify the same.
- Provide the right sequence of rendering of content

V. CLOUD COMPUTING FOR RURAL EDUCATION

Once the system has been developed, it should be accessible to the rural community. For this, the Web based system must be deployed on the Cloud. The lacunae in infrastructure, student and teachers in the rural community can be offset by the benefits of deploying on Cloud [7].

- As the application is run in the Cloud and not on the desktop PC, the desktop PC doesn't need the processing power or hard disk space demanded by traditional desktop software. Hence the client computers, i.e. the rural schools systems can be lower priced, with smaller hard disks, less memory, more efficient processors, and the like. There is no requirement of a CD or DVD drive, because no software programs have to be loaded and no document files need to be saved at the user side.

- Computers in a Cloud computing system will boot up faster and run faster, because they'll have fewer programs and processes loaded into memory
- Cloud computing greatly reduces both hardware and software maintenance. With less hardware (fewer servers) necessary in the schools, maintenance costs are immediately lowered. As to software maintenance, as cloud applications are based elsewhere, there's no software on the school's computers for the IT staff to maintain.
- The need to access any particular software in the cloud now becomes less expensive. Software licenses need not be purchased for each system. Even if it costs the same to use web-based applications as it does similar desktop software (which it probably won't), the maintenance staff have saved the cost of installing and maintaining those programs on every desktop.
- When the app is web-based, updates happen automatically and are available the next time the user logs in to the cloud. Whenever the student accesses a web-based application, the updated version of the application is being downloaded.
- All contents are instantly available from whichever community schools have been registered with the education system
- The documents can be accessed simultaneously; the modifications done by one user is automatically reflected by which the other users can see onscreen.

Phase	*Activity*
Requirements	• *Define Objectives* • *Feasibility study* • *Validate scope of the project* • *Understand Key processes* • *Study the Application Landscape (Environment)* • *Understand the key requirements* • *List the risks, dependencies & impacted components*
Planning & Analysis	• *Design a Rollout plan* • *Define test cases* • *Design Document* • *Validate the capacity planning of the servers*
Implementation	• *Setup POC environment for pilot application and build platform* • *Test the Upgrade & amend the target design as per results* • *Assess any performance and availability related issues.* • *Revisit the design and update as needed* • *Prepare the Master rollout schedule*

Fig. 6 Life Cycle of Deployment of system on Cloud

The web based education system will be implemented through various stages as shown in figure 6.

In the requirements phase, the feasibility study is to be carried out. The number of schools, users including the students and faculty should be assessed. The key processes such as content loading and updating; student updating of assignment downloading of contents, availability of software pertaining to the format of the content to be downloaded should be assessed. The scope of the project should also be analyzed.

This includes the content will be formed for how many courses, the formats available for a particular content, the no of users and the number of schools accessing the web based application. The risks also has to be accessed which can occur due to disagreement in the content of a course. It can also occur due to loss of information, change of author etc .In the analysis and design phase, the various components for students, faculties and authors have to be tested. Phase wise testing will take place. The web based education system will be uploaded on a server however the capacity required as per users, data should be considered and decided.

The implementation phase will be for deciding the scope of proof of concept. The border of environment will be formed. Here the number of users, schools, and the content will be worked for the initial POC. The entire POC will be revisited depending on the feedback and issues that arises in the design and planning phase.

VI. CONCLUSION

The paper aims to present in details the components of a rural education system. The drawbacks and the challenges have briefly discussed. The technological solution to fill the lacunas of the present rural education scenario was briefly explained.

The ontology has been used for representing the knowledge repositories with respect to students and content. By deploying the web based semantic education system on to cloud help address several other issues regarding the rural education scenario.

REFERENCES

[1] T.Coladarce, Improving the yield of Rural Education Research: An Editors Swan Song, Journal of Research in Rural Education, 2007, 22(3)

[2] S.Chakraborty and Bhatacharya, Shishak: An Intelligent tutoring System Authoring tool for Rural Education", Information and Communication Technologies & Development, 2007 ICTD 2007,pp1-10

[3] M. Hall , The Challenges for India's Education System, Chatham House, Asia Programme, April 2005, ASP BP 05/05

[4] BiswanathDutta and Devika P Madalli, Ontology Supported Personalized E-learning repositories, University of Trento, August 2009, Technical report, #DISI-09-052

[5] S.Babu, FOSS Movement and its impact education, CSI Communications, Knowledge Digest for IT community, Issue No 6, Sept 2011, Pp 19-22

[6] A Kumar and R Rajendra, Perspectives on ICT Education in India",

[7] Micheal Miller, Cloud Computing, 2008, Edition

Efficient Threshold Signature Scheme

Sattar J Aboud
Department of Information Technology
Iraqi Council of Representatives
Baghdad-Iraq

Mohammad AL-Fayoumi
Faculty of Computer and Information Systems,
Umm Al-Qura University
Saudi Arabia

***Abstract*— In this paper, we introduce a new threshold signature RSA-typed scheme. The proposed scheme has the characteristics of un-forgeable and robustness in random oracle model. Also, signature generation and verification is entirely non-interactive. In addition, the length of the entity signature participate is restricted by a steady times of the length of the RSA signature modulus. Also, the signing process of the proposed scheme is more efficient in terms of time complexity and interaction.**

Keywords- Shamir secret sharing; threshold signature; random oracle model.

I. INTRODUCTION

Disclosure of a private key for non-cryptography purposes for example a compromise of the basic system, human mistake or insider attacks, is actually the highest threat to many cryptography schemes. The most generally suggested solution is distribution of the private key over multiple servers by secret sharing. For digital signature, the primitive we deal with in this paper is the main direction of this thought threshold signature scheme.

However, the interesting type of secret sharing scheme contains threshold scheme with a set of n participants. Their access structure contains all subgroup of t or more participants. Such schemes are called t out of n threshold schemes or just (t,n) schemes. Threshold scheme was independently presented by Shamir [1]. This scheme is relied on polynomial interpolation over a finite field. Suppose $K = GF(q)$ is a finite field with q elements. To build a (t,n) threshold scheme a dealer D selects n distinct nonzero numbers of $GF(q)$ indicated by $x_1,\ldots,x_n$, and passes x_i to P_i upon a public key channel $(i=1,\ldots,n)$. For a secret $K \in GF(q)$, D arbitrarily selects $t-1$ set $a_1,\ldots,a_{t-1}$ from $GF(q)$ and builds a polynomial $f(x) = K + \sum_{i=1}^{t-1} a_i * x^i$ .The share for participant P_i is $s_i = f(x_i)$ .The degree of $f(x)$ is at most $t-1$. It is documented that Shamir scheme is perfect. That is, when a collection of fewer that t participants work together, their original doubt about K is not reduced. Suppose that any subset of r players out of R generate a signature, but reject the generation of a valid signature when less than r players involve in the scheme. This unforgeability characteristic must keep even when certain subgroup of fewer than r players are cheated and act mutually.

For a threshold scheme to be practical if certain players are cheated, it must also be strong, meaning that cheated players must not be capable to stop honest players from creating signature. In this paper, we will consider suggested scheme which face at least one of the following difficulties:

a) *With no accurate security proof, even with a random oracle model.*
b) *Signature generation and verification is not interactive.*
c) *The length of an entity signature explodes linearly in the number of players.*

To enhance this, we will introduce a new threshold RSA-based signature scheme which faces these difficulties. We will highlight that the signature outcome is an entirely invert RSA signature, meaning that the generation and verification algorithms are the same as for common RSA signature. But, there are certain limitations on the public key which should be a prime and the modulus should be the result of two strong prime numbers. The suggested scheme is easy to calculate, and has not previously suggested. However, preceding schemes of threshold signature have that $r = w+1$. This generalization is practical in situations where the honest players is not necessity choose what they are signing, but capable to verify that a big number of them have authorized a specific signature. In specific, threshold signatures with $r = R - w$ and $w < R/3$ is used to decrease the lengths of the messages pass in coordinated network agreement scheme [1]. The use to coordinated network agreement was in fact the original purpose for this study. Almost all preceding work on threshold signatures supposes with a coordinated network, and any players in some way simultaneously agree to commence the signing scheme on a known document. Obviously, we cannot act in such a system when we desire to employ coordinated network agreement protocol.

We also highlight that the idea of a twin parameter threshold scheme gives robust security than one parameter threshold scheme; such scheme is actually more challenging to build and to discuss. The proposed idea of a twin parameter threshold scheme must not be confused with a vulnerable idea that from time to time seems in a threshold cryptosystem research [2]. For this vulnerable idea, there is a parameter $r' > w$ where the rebuilding algorithm needs r' shares, but the security is lost when only one truthful player discloses a share. In proposed idea, no security is lost unless $r - w$ truthful players disclose their shares. We work with a static cheating

system; the opponent should select which players to cheat at the start of the attack. This is in line with preceding studies into threshold signatures, which also suppose static cheating. The proposed system can be verified if $r = w+1$ in the random oracle model using the RSA signature.

II. Related Work

In 1989, Desmedt and Frankel [3] describe the difficulty with threshold signature scheme. This appear from the truth that polynomial interpolation by a coefficient ring $Z_{\theta(n)}$ such that n the RSA modulus and θ is the Euler phi. Also, Desmedt and Frankel in 1991[4] return again to the difficulty of threshold, and introduce a non-robust threshold scheme that is non-interactive but with small share length and without security discussion. Frankel and Desmedt in 1992 [5] introduce approach that providing a proof of security for a non-robust threshold scheme with small share length, but which needs coordinated interaction. Harn in 1994 [6] introduces a robust threshold scheme with small share length that also needs coordinated interaction. Gennaro et al. in 1996 [7] describe a robust threshold scheme with small share length, but again needs coordinated interaction. Actually, Gennaro et al. scheme can be examined with no reconstruction of random oracle. But this will have some practical disadvantages, demanding a particular relationship between the sender and receiver about the share of a signature. It appears that the security of these systems needs carefully examination by an acceptable approach. However, the above schemes are interactive and any threshold signature scheme relied on integer factoring seems inevitable to be interactive, because such signature schemes are randomized, and thus the signers have to create random values, which actually needs coordinated interaction.

But, in 1996 De Santis et al. [8] introduce a variant scheme that uses interaction for large share length. This scheme prevents the difficulties of polynomial interpolation over $Z_{\theta(n)}$ by working with $Z_{\theta(n)}(i)/(\theta_q(i))$, such that $\theta_q(i)$ is the q^{th} polynomial taken $\bmod \theta(n)$, and q is a prime larger than l. This is suitable, as standard secret sharing method can be directly used, but guides to a more difficult scheme that need coordinated interaction. In 1998, Rabin [9] suggests a strictly robust threshold scheme that has small share length, but need coordinated interaction. This scheme takes a diverse line of the interpolation over $Z_{\theta(n)}$ problem, avoiding it by presenting an additional layer of secret sharing and a lot more interaction. In 2006, Jun et al [10] described a non-interactive verifiable secret sharing scheme built by Shamir secret sharing scheme for secure multi-party communication scheme in distributed networks. In 2007, Li et al. [11] they introduce a secure threshold signature scheme without trusted dealer. In the meantime, the signature share generation and verification algorithms are non-interactive. In 2010 Gu, et al. [12] discuss the security of Jun et al. scheme and show that their scheme cannot withstand the misleading performance as they claimed.

$z_A^{'} = z_A \oplus w_A \oplus w_A^{'}$.

III. Scheme Requirments

There are three entities the player R, the dealer and an opponent. There are also a signature verification phase, a share verification phase and a share combination phase. In addition, there are two other variables, w represent number of cheated players; and r denote the number of signatures required to get a signature. The only restrictions are that $r \geq w+1$, and $R - w \geq r$.

The opponent chooses a subset of w players to cheat. In the dealing phase, the dealer establishes a public key e and private key shares $sk_1 .. sk_R$, and verification keys $vk_1 .. vk_R$. The opponent gets the private key shares of the cheated players and the public key and verification keys. Following the dealing phase, the opponent passes signing demand to the honest players for document of his choice. Upon such a demand, a player results a signature share for the known document. The signature verification phase obtains a document, a signature and the public key, then verifies whether the signature is valid or not. The signature share verification phase obtains a document, a signature share on that document from players i, with pk, sk_x, vk_x, and verifies whether the signature share is valid or not.

IV. The Proposed Scheme

In this section, we describe the proposed scheme.

The dealer. The dealer must do the following:

1. Selects arbitrarily two primes p and q, such that $p = 2 * p^{'} + 1, q = 2 * q^{'} + 1$ with $p^{'}, q^{'}$ are also primes.
2. Finds the modulus $n = p * q$.
3. Selects the message $m = p^{'} * q^{'}$.
4. Selects the public key e as a prime $e > 1$.
5. The public key is (e, n).
6. Finds $d = e^{-1} \bmod m$.
7. Let $v_0 = d$.
8. Selects v_x arbitrarily from $(0,...,m-1)$ for $1 \leq x \leq r-1$.
9. The vector $v_0,...,v_{r-1}$ determine the polynomial $f(i) = \sum_{x=0}^{r-1} v_x * i^x$.
10. Finds $s_x = f(x) \bmod m$ for $1 \leq x \leq R$. (1)
11. This element s_x is a secret key share of player x. This indicate by D_n the subgroup of squares in Z_n^*.
12. Selects an arbitrary $u \in D_n$.
13. Finds $u_x = u^s \in D_n$ for $1 \leq x \leq R$.
14. These statements determine the verification keys, $vk = u$ and $vk_x = u_x$.

Remarks. We will ensure that all set computations are performed in D_n, and equivalent exponent arithmetic in Z_m.

This is suitable, because $m = p' * q'$ has no small prime factors. Because the dealer selects $u \in D_n$ arbitrarily, suppose that u creates D_n, because this occurs with all but small probability. Since of this, the number u_x entirely find out the result of $s_x \bmod m$. For each subgroup of r points in $(0,...,R)$, the result of $f(i) \bmod m$ at these points uniquely finds out the coefficients of $f(i) \bmod m$, and since the result of $f(i) \bmod m$ at every other point mod in $(0,...,R)$. This follows from the information that the equivalent Vandermonde vector is invertible mod m, because its determinant is co-prime to m. From this, it ensures that for each subset of $r-1$ points in $(1,...,R)$, the distributions of the result of $f(i) \bmod m$ at these points are standardized and equally independent. Suppose $a = R!$ for each subset s of k points in $(0,...,R)$ and for each $x \in (0,...,R) \setminus s$ and $y \in s$, we can describe:

$$H_{x,y}^{s} = a * \frac{\prod_{y' \in s \setminus (y)} (x - y')}{\prod_{y' \in s \setminus (y)} (y - y')} \quad (2)$$

These results are resulting from the standard Lagrange interpolation equation. They are obviously integers; hence the denominator divides $y!(R-y)!$ which in divides $R!$. It is also obvious that these results are easy to calculate. From the Lagrange interpolation equation, we have:

$$a * f(x) \equiv \sum_{y \in S} H_{x,y}^{S} f(y) \bmod m \quad (3)$$

Valid signature. We require a hash function h to elements of Z_n^*. If $i = h(m)$, thus the valid signature on m is $j \in Z_n^*$ where $j^e = i$. This is only a common RSA signature.

Generating signature share: In order to generate a signature share on a document m we should do the following.

1. Choose $i = h(m)$.
2. The signature share of player x is $i_x = i^{2*a*s_x}$ (4)

Correctness. The verification of correctness is just a proof of the discrete logarithm of i_x^2 to the base $i' = i^{4*a}$. (5)

However, we can simply adjust a well-known interactive scheme of Chaum and Pedersen [13]. We collapse the scheme, making it non-interactive, by employing a hash function to generate the challenge such that a random oracle model is required. We also have to handle the actuality that we are using a group D_n whose order is not known. So, this is unimportantly managed by just using adequately big integer. Suppose $L(n)$ is the bit-size of n. Assume that h' is the hash function, whose result is L_1 bit integer, such that L_1 is a security parameter. To build the verification of correctness player x select a random number $r \in (0,...2^{L(n)+2L_1} - 1)$, then finds:

- $u' = u^r$
- $i' = i'^r$
- $c = h(u, i', u_x, x_x^2, u', i')$
- $z = s_x * c + k$

The verification of correctness is (z, c).

Correctness. one verifies that $c = h'(u, i', u_x, x_x^2, u^z * u_x^{-c}, i'^z * x_x^{2*c})$. The cause for using i_x^2 instead of i_x is that because i_x is assumed to be a square, this is not simply checked. This means, we are certain to be using D_n, so we want to ensure soundness.

Combining shares. Assume that we have valid shares from a group s of players, such that $s = (x_1,...,x_r) \subset (1,...,R)$. Assume $i = h(m)$ and suppose that $i_{xy}^2 = i^{4*a_{s_{xy}}}$. Then to rearranged shares, we find $t = i_{x_1}^{2*H_{0,x_1}^{s}} .. i_{xy}^{2*H_{0,,xy}^{s}}$ Such that H is the integers described in (2). From (3), we hold $t^e = i^{e'}$, thus $e' = 4 * a^2$ (6) as $\gcd(e, e') = 1$, it is simple to find j where $j^e = i$, employing a standard method $j = t^q * i^b$ such that v and b are integers where $e' * v + e * b = 1$; that can be got from the extended Euclidean method on e', and e

V. Security Discussion

Theorem 1: the proposed scheme is a secure threshold signature protocol if the common RSA signature scheme is secure. We illustrate that to simulate the opponent vision, if the opponent requests for a signature share from the honest player. Assume $x_1,...,x_{r-1}$ is the set of cheated players. Consider $s_x \equiv f(x) \bmod m$ for all $1 \le x \le R$, and $d \equiv f(0) \bmod m$. To simulate the opponent vision, we just select the S_{xy} belonging to the group of cheated players randomly from the vector $(0,...,\lfloor n/4 \rfloor - 1)$.

We have by now discussed that the cheated player private key shares are arbitrary numbers in the vector $(0,...,m-1)$. We hold $n/4 - m = (p' + q')/2 + 1/4 = O(n^{1/2})$; and from this easy computation illustrates that the statistical distance between the regular distribution on $(0,...,\lfloor n/4 \rfloor - 1)$ and the regular distribution on $(0,...,m-1)$ is $O(n^{-1/2})$. When these S_{xy} members are selected, the values s_x for the honest players are also entirely fixed mod m, although cannot be simply

calculated. Though, provided i, j and $j^e = i$, we can simply find $i_x = i^{2*a*s_x}$ for the honest player x as:

$$i_x = j^{2(H^s_{x,0} + e(H^s_{x,x_1} * s_{x_1} + \ldots + H^s_{x,x_{r-1}} * s_{x_{r-1}}))}.$$

Such that $s = (o, x_1, \ldots, x_{r-1})$, this results from (3). Employing this method, we can create the vector $u, u_1, \ldots u_R$, and also create some share i_x of the signature, provided the common RSA signature. This case illustrates that we described the share i_x to be i^{2*a*s_x} and not i^{2*s_x}. This thought was employed by Feldman [14] in the situation of where another associated problem of provable secret sharing.

Proofs of correctness: entity can use the random oracle model for the hash value h' to obtain soundness and arithmetical zero-knowledge. This is very simple, but we drawing the information.

Now, we study soundness. We need to illustrate that the opponent cannot build, except with insignificant probability, the proof of correctness for an inaccurate share. Assume i and i_x is provided, and a valid proof of correctness (z, c). We hold $c = h'(u, i', u_x, i_x^2, u', i')$ such that:

- $i' = i^{4*a}$
- $u' = u^z * u_x^{-c}$
- $i' = i'^z * i_x^{-2*c}$

Right away, i', u_x, i_x^2, u', i' are simply observed in D_n, and we are supposing that u create D_n. So we hold:

$$i' = u^a$$
$$u_x = u^{s_x}$$
$$i_x^2 = u^B$$
$$u' = u^j$$
$$i' = u^g$$

For a number of integers v, B, j, g furthermore,

$$z - c * s_x \equiv j \bmod m$$
$$z * v - c * B \equiv g \bmod m$$

Multiplying the first formula by v and subtracting the second, we obtain: $c(B - s_x * v) \equiv v * j - g \bmod m$ (7)

So, the share is accurate when and only when $B \equiv s_j * v \bmod m$ (8)

When (8) fails to retain, therefore it should be unsuccessful to have $\bmod p'$ or $\bmod q'$, and thus (7) uniquely finds out $c \bmod$ one of these primes. Although in the random oracle model, the distribution of c is consistent and separate from the data to a hash value, and thus this still occurs with insignificant probability.

Now, we will study zero-knowledge. We can build a simulator that simulates the opponent vision without knowing the result of s_x. This observation contains the results of the random oracle at those situations where the opponent has queried the oracle, thus the simulator is in entire charge of the random oracle. When the opponent constructs a query to the random oracle, if an oracle has not been determine before at the provided point, the simulator describes it an arbitrary value, and in all cases return the result to the opponent. If an honest player is supposed to create a proof of correctness for a provided i, i_x, the simulator selects $c \in (0, \ldots 2^{L_1} - 1)$ and $z \in (0, \ldots 2^{L(n)+2*L_1} - 1)$ randomly, and for provided integers i, i_x determines the number of the random oracle $(u, i', u_x, x_x^2, u^z * u_x^{-c}, i'^z * i_x^{-2*c})$ to be c. With insignificant probability, the simulator has not described the random oracle at this point, and thus it is limitless to do so. The proof is (z, c). It is easy to check a distribution created by this simulator is statistically near to perfect.

From soundness, we obtain the strength of the threshold signature protocol. From zero-knowledge, we obtain the non-forgeability of the threshold signature protocol, supposing the common RSA signature scheme is secure, that is existentially non-forgeable anti-adaptive chosen message attack. Such approach is more correct in the random oracle model for h, this typed follows from the RSA-based provided random $i \in Z_n^*$, it is difficult to find j where $j^e = i$.

VI. Conclusion

In this paper, we illustrated the threshold signature scheme. We introduced a strong threshold signature scheme relied on a secret sharing scheme. The suggested signature scheme simplifies threshold RSA signature in which relied on Shamir secret sharing, and is an efficient. In addition, the method can be extended to further public key cryptography as the secret key is utilized in the exponent.

References

[1] Cachin C, Kursawe K, and Shoup V, "Random oracles in Constantinople: practical asynchronous Byzantine agreement using cryptography", Manuscript, 2000.

[2] Micali S and Sidney R, "A simple method for generating and sharing pseudo-random functions, with applications to Clipper-like key escrow systems", Advances in Cryptology, Crypto'95, pages 185-196, 1995.

[3] Desmedt Y and Frankel Y, "Threshold cryptosystems", Advances in Cryptology Crypto'89, pp, 307-315, 1989.

[4] Desmedt Y and Frankel Y, "Shared generation of authenticators and signatures", Advances in Cryptology, Crypto'91, pp. 457-569, 1991.

[5] Frankel Y and Desmedt Y, "Parallel reliable threshold multi-signature", Technical Report TR-92-04-02, University of Wisconsin, Milwaukee, 1992.

[6] Harn L, "Group-oriented (t; n) threshold digital signature scheme and digital multi-signature", IEE Proceeding Computer Digital, Tech., 141(5):307-313, 1994.

[7] Gennaro R, Jarecki S, Krawczyk H, and Rabin T, "Robust threshold DSS", "Advances in Cryptology, Eurocrypt'96, pp. 354-371, 1996.

[8] De Santis A, Desmedt Y, Frankel Y, and Yung M, "How to share a function securely", 26th Annual ACM Symposium on Theory of Computing, pp. 522-533, 1994.

[9] Rabin T, "A simplified approach to threshold and proactive RSA", Advances in Cryptology Crypto'98, 1998.

[10] Jun Ao and Guisheng Liao, "A Novel Non-interactive Verifiable Secret Sharing Scheme", Chunbo Ma, Communication Technology, ICCT'06, International Conference on 27-30 November 2006, pp. 1 – 4.

[11] Jin Li, Tsz Hon Yuen and Kwangjo Kim, "Practical Threshold Signatures without Random Oracles", Lecture Notes in Computer Science, 2007, Volume 4784, 2007, 198-207.

[12] Feng Wang, Yousheng Zhou, Yixian Yang and Yajian Zhou, "Comment on a Novel Non-interactive Verifiable Secret Sharing Scheme", Communication Software and Networks, ICCSN'10, Second International Conference on, 26-28 Feb. 2010, pp. 157-159

[13] Chaum D and Pedersen T, "Wallet databases with observers", Advances in Cryptology, Crypto'92, pp. 89-105, 1992.

[14] Feldman P, "A practical scheme for non-interactive verifiable secret sharing", 28th Annual Symposium on Foundations of Computer Science, pp. 427-437, 1987.

AUTHORS PROFILE

Sattar J Aboud is a Professor and advisor for Science and Technology at Iraqi Council of Representatives. He received his education from United Kingdom. Dr. Aboud has served his profession in many universities and he awarded the Quality Assurance Certificate of Philadelphia University, Faculty of Information Technology in 2002. Also, he awarded the Medal of Iraqi Council of Representatives for his conducting the first international conference of Iraqi Experts in 2008. His research interests include the areas of both symmetric and asymmetric cryptography, area of verification and validation, performance evaluation and e-payment schemes.

Mohammad Al-Fayoumi is a Professor at Umm Al-Qura University in Saudi Arabia. He received his education from Romania. Dr. Fayoumi has served his profession in many universities. His research interests include the areas of software engineering, verification and validation, information security and simulation.

Fault Tolerant Platform for Application Mobility across devices

T. N. Anitha [1]
Associate Professor, Dept of Computer Science & Engineering , SJCIT, Chickballapur-562101,Karnataka,India

Jayanth. A [1]
Project Engineer,Oracle India Private Limitted
Bhannerughatta Road, Near Diary circle, Koramangala, Bangalore, Karnataka,India

***Abstract*—In the mobile era, users started using Smartphone's, tablets and other handheld devices, The advances in telecom technologies like 3G accelerates the migration towards smart phones. But still battery power and frequent change of handsets is still a constraint. They burden on user had to manually synchronize their contacts, applications they use to the new phones. Also they loss whatever they are doing when the mobile get power down. In this paper, we propose a solution to the problem discussed with a new fault tolerant platform which can provide application mobility across the devices.**

Keywords- Fault tolerance; Middle ware; Midstore Manager.

I. INTRODUCTION

Application mobility refers to the idea application can run across multiple devices seamlessly even if it fails on one device say to battery down or users move to some other device. The application does not stop and it continues doing its work across user's movement to devices. We choose smart phones as the device and laptops as the devices to consider for application mobility. Application can be anything from word editing, preparing the power point presentations etc.

We explain the case for application mobility with some scenarios. Say a user is a browsing a web page in his smartphone, his battery power is low, can he work with he work with the same page in browser in laptop automatically, or he has another phone and want to work with same page in the handheld. Say the user is watching a you tube video, he is at 2 min another 3 min video is pending, he wants to watch this video on his tablet from the 2min or his phone switches off and he powers again, can he still watch the video from 2min automatically. Currently there is no solution available in market for these problems which motivated our work. In this paper, we address these challenges and design a software solution these problems.

II. RELATED WORK

Fault Tolerance solutions usually are designed for production servers like air traffic control, distributed disaster system, railways reservation system, internet banking where a single fault may lead to huge loss of money and even human lives. Replication based technique is one of the popular fault tolerance techniques [1]. A replica means multiple copies. Replication is a process of maintaining different copies of a data item or object. In replication techniques, request from client is forwarded to one of replica among a set of replicas. This technique is used for request that do not modify state of service. Replication adds redundancy in system. In this way failure of some nodes will not result in failure in system and thus fault tolerance is achieved.

Checkpoint with rollback-recovery is a well-known technique. Checkpoint is an operation which stores the current state of computation in stable storage. Checkpoints are established during the normal execution of a program periodically. This information is saved on a stable storage so that it can be used in case of node failures. The information includes the process state, its environment, the value of registers, etc. When an error is detected, the process is roll backed to the last saved state [2].

Although replication method is widely used as a fault tolerance technique but number of backups is a main drawback. Number of backups increases drastically as coverage against number of faults increases. As the number of backup increases management of these backups is very costly. Fusion based techniques overcome this problem. It is emerging as a popular technique to handle multiple faults. Basically it is an alternate idea for fault tolerance that requires fewer backup machines than replication based approaches. In fusion based fault tolerance a technique, back up machines is used which cross product of original as fusions is corresponding to the given set of machines [3]. Overhead in fusion based techniques is very high during recovery from faults. Hence this technique is acceptable if probability of fault is low.

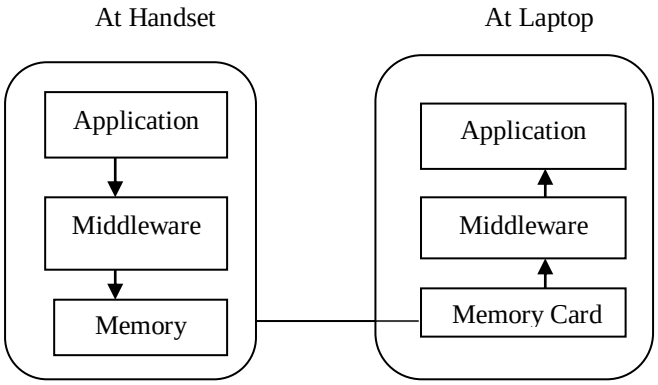
All these fault tolerance solutions address only for the server systems. We cannot use these solutions directly for our problem; all these solutions are designed for applications running in homogenous platform and the platform on which application run are same across devices.

But in our problem we need to work across devices with different platform. We need to provide application mobility for application running in Android platform to the Application running in Symbian platform or to windows on a laptop. So the fault tolerance solution for application mobility becomes even more complex.

III. PROPOSED SOLUTION

The proposed solution consists of designing a middleware. The middle ware will provide fault tolerant application mobility. The middleware provides API's for application to synchronize essential information.

The middleware can write the synchronization information into a memory card and it can also read this synchronization information and start the applications. We could have used cloud storage for synchronization of information, but relaying on cloud will sometimes become a problem like network not available. Using memory card has its advantages. The read and write operations for synchronization becomes very much faster. The memory card can be used to stat the application seamlessly on another laptop or other handheld easily, just insert the memory card and start the middleware.



So proposed system now standardizes the protocols and the standards used for Application Mobility.

A. Middleware

There are two approaches for the design of middleware. In first case, the application has to manually call the API's to set the synchronization information. In the next case, middle ware is dynamic and abstract the device API's.

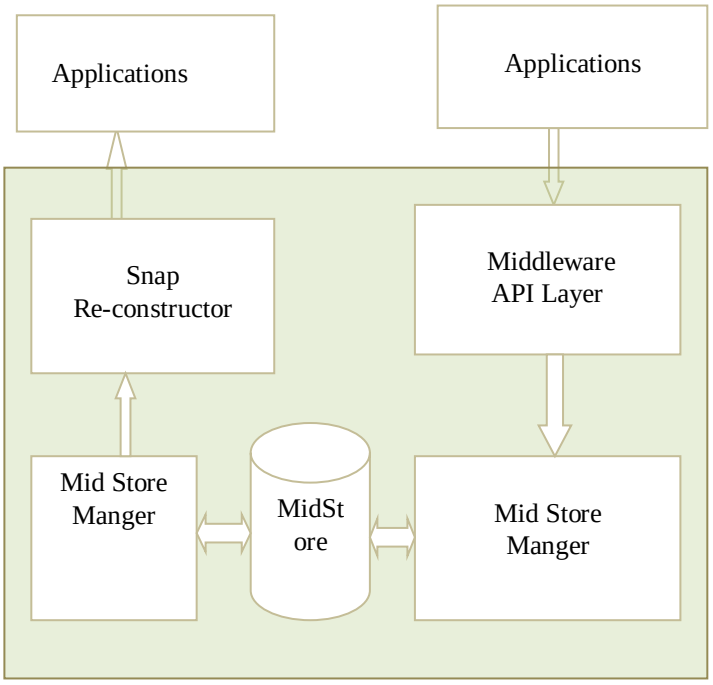
In this work , we limit our scope to application has to manually call the middle's API.

Middleware API's should take a minimal amount of time for providing the fault tolerance without affecting the application performance. Poor design of middleware will result in lower application performance and will seriously impact user experience. Middle ware records the session activity in a XML format in the Memory Card.

Since the Memory Card is also used by user for his contents, we should provide secure and incorruptible way for our XML. This can be provided by creating a folder in the Memory card for the purpose of middleware alone and using a virtual file system. These way middleware operations are assured a separate workspace in the Card. This virtual file system is referred as MidStore from here on.

Applications uses the middleware API's to record the current application session activity into the MidStore. At any point of time, the information in the middleware provides the snapshot of mobile user's current activity. The information is MidStore has this snapshot. Our job of Application mobility in case of fault tolerance is very easy, if we can reconstruct this snap on any other device or platform. To make this easy, first step is standardizing the format of recording the snapshot. We use XML standard for this purpose. XML is well standardized way to share data across cross platforms.

Middleware must also provide a reconstruct job as service which can invoked on the other platform to reconstruct the snapshot. Based on the discussion above, so far we formulate the architecture of middleware as



The core of our proposed solution is in the Middleware App Layer, Mid Store Manager, MidStore and Snap Re constructor.

B. Middleware App Layer

This layer provides the API's for the application to store the session activity in the MidStore. Say the Application is browser on the mobile handset. User has entered a URL and got the webpage. This activity must be recorded into the midstore.

Say the Middleware provides the API like below.

RecordActivity (Application, DescritorParam);

The Browser Application will call RecordActivity with the parameters filled as

Application: IE

DescriptorParam: www.yahoo.com

Once this API call is made, middle API Layer delegates the API to the MidStore manager.

We care not addressing any specific implementation of API's in this paper. But as a general guideline, the API should give the following information about user activity to the Middleware platform. It should provide the activity is related to which application and the additional descriptor of the activity.

If the user is browsing a webpage with internet explorer, the application is IE and the descriptor is URL for browsing. If the user is playing a song with media player, the application is WMP and the descriptor is the song name.

C. MidStore Manager

As discussed previously the memory card have a virtual file system called as MidStore. In order to manage the read and write to the MidStore we need the MidStore manager.

MidStore Manger handles the api call from the APP Layer and records this activity in the MidStore. It is upto the specific implementation of Mid Store Manager , to record all the activity in a single XML file or make separate XML file for each application. The case of making separate XML for each application has a particular advantage during reconstruct process. This will be addressed in the Snap re constructor.

We only suggest a way for the XML format, but it is upto the specific implementers to have their own way

```
<midstore>
  <application>
   < instance = "IE"/>
  < descriptor url =www.yahoo.com />
  </application>
</midstore>
```

Suppose the User have closed the browsing application , the browser should call an API

RemoveInstance(Application , Descriptor)

With Application value as IE.

When this api is called, the midstore manager must remove the application entries for IE.

D. SnapReConstructor

SnapReConstructor is an important module in our design. It reads the midstore and replicates the user activity on any target platform.
Suppose the mid store has the following

```
<midstore>
  <application>
   < instance = "IE"/>
  < descriptor url =www.yahoo.com />
  </application>
</midstore>
```

And reconstuctor is running on desktop system with windows. Then by reading this snap the reconstuctor should be able to open the IE browser with url as www.yahoo.com. SnapReConstructor can also have additional logic called as ApplicationAliter.

ApplicationAliter can help partial replication in target platform. Say IE is not in the target platform, but it's found out that Mozilla is installed in the target. Since both IE and Mozilla are web browsers, SnapReConstructor can start the Mozilla with url as www.yahoo.com. So this guarantees partial restoration.

SnapReConstructor must always try to achieve maximum restoration without user intervention. Instead of SnapReConstructor to be provided as an application and user always click to start , it can be made auto start whenever the device is introduced into target platform. But we leave this to implementers choice.

For any new application to be supported by the middleware, we need to extend the middleware API layer for providing parameters and SnapReConstructor must be extended to start the application or must ApplicationAliter must be extended for doing partial reconstruction.

Instead of application reading each application tag in the MidStore , if it can get all instances for that specific application, it can start the application at one start with all instances. We find this very useful in certain targets for some applications. In windows, stating a IE can be done with multiple url , so one ie instance comes up with different tabs for each url.

Not all application available in mobile platform is available for desktop platform. Also not all applications are available across different mobile platform. But popular applications for consumer user and the enterprise application are moving towards availability in all platforms. Also with the use of ApplicationAliter we can always find alternative for the applications.

E. Application of proposed Platform

In this section we detail the possible usage scenarios for our project.

A user adds contacts to his phone. He adds as many contacts to his phone. Since mobile phones are cheaper and new model comes to market every day, and in rapid mobile use in countries like India , people often change phone , so every time they change they have sync with contacts , important messages , calendar events etc , but with our platform in place , the user can just change the handset and connect to his computer, the platform will immediately get the application information like messages , calendar events etc applications continues in the new handset.

F. Platform against Battery Backup

Modern smart phones even claim a battery backup of 10 days or so, but for busy business users, they see that battery does not last for even 2 days. Many times business user has to carry another phone. He has to continue his operation using the new handset, but from the same point of continuity.

Currently there is no solution for this problem and the transitory phase for business user from handset to handset is a very bad experience, but with our platform the transitory is smooth for the user, with the application snapshot same as he worked previously.

Since our platform is not so complex and uses less resources it can even run with low end platforms.

IV. CONCLUSION

In this paper work, we have provided a solution for application fault tolerance and mobility across different platforms. Any Implementation can use the solution to realize it on different platforms. The Application mobility will be of great advantage to the enterprise and consumer applications. The application has to change a bit to save the user activity into the MidStore. This will be a disadvantage for the existing applications. In Android platform , with the concept of Broadcast receiver, any application can watch for certain events , so browsing a url , playing a song etc are all events , so a watcher application cab be written to watch for the events and record user activity to the MidStore. We are looking for the same kind of solutions on other platforms too. Once we are able to find the generic watcher solution, the Watcher

component can also be added to the middleware. This will be a big advantage for providing mobility for existing applications.

REFERENCES

[1] Bhargava B. and Lian S. R., "Independent Checkpointing and Concurrent Rollback for Recovery in Distributed Systems-An Optimistic Approach," Proceedings of 17th IEEE Symposium on Reliable Distributed Systems, pp. 3-12, 1988.

[2] Cao G. and Singhal M., "On coordinated checkpointing in Distributed Systems", IEEE Transactions on Parallel and Distributed Systems, vol. 9, no.12, pp. 1213-1225, Dec 1998.

[3] V. Agarwal, Fault Tolerance in Distributed Systems, I. Institute of Technology Kanpur, www.cse.iitk.ac.in/report-repository, 2004. ,

[4] "XML Media Types, RFC 3023". IETF. 2001-01. pp. 9–11. http://tools.ietf.org/html/rfc3023#section-3.2. Retrieved 2010-01-04.^ "XML Media Types, RFC 3023".

[5] Adnan Agbaria, William H. Sanders, " Distributed Snapshots for Mobile Computing Systems", Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications](Percom'04), pp. 1-10, 2004.

[6] Parveen Kumar, Lalit Kumar, R K Chauhan, "A low overhead Non-intrusive Hybrid Synchronous checkpointing protocol for mobile systems", Journal of Multidisciplinary Engineering Technologies,Vol.1, No. 1, pp 40-50, 2005.

[7] Parveen Kumar, Lalit Kumar, R K Chauhan, "Synchronous Checkpointing Protocols for Mobile Distributed Systems: A Comparative Study", International Journal of information and computing science, Volume 8, No.2, 2005, pp 14-21

[8] "XML Serialization in the .NET Framework". Msdn.microsoft.com. http://msdn.microsoft.com/en-us/library/ms950721.aspx. Retrieved 2009-07-31

[9] V.K Garg,. "Implementing fault-tolerant services using fused state machines," Tech-nical Report ECE-PDS-2010-001, Parallel and Distributed Systems Laboratory,ECE Dept. University of Texas at Austin (2010).

[10] Extensible Markup Language (XML) 1.1 (Second Edition)".W3.org. http://www.w3.org/TR/xml11/#charsets. Retrieved 2010-08-22.IETF. 2001-01. pp. 7–9. http://tools.ietf.org/html/rfc3023#section-3.1. Retrieved 2010-01-04.

[11] M. Murata, D. Kohn, and C. Lilley (2009-09-24). "Internet Drafts: XML Media Types". IETF. http://tools.ietf.org/html/draft-murata-kohn-lilley-xml-03. Retrieved 2010-06-10.

[12] "XML 1.0 Specification". W3.org. http://www.w3.org/TR/REC-xml. Retrieved 2010-08-22.

[13] M. Wiesmann, F. Pedone, A. Schiper, B. Kemme, G. Alonso," Understanding Replication in Databases and Distributed Systems," Research supported by EPFLETHZ DRAGON project and OFES).

[14] Checkpoint-based Fault-tolerant Infrastructure for Virtualized Service Providers.

[15] A Review of Checkpointing Fault Tolerance Techniques in Distributed Mobile Systems.

Viable Modifications to Improve Handover Latency in MIPv6

Mr.Purnendu Shekhar Pandey
Department of ICT,
Gautam Buddha University,
Gr.Noida, Gautam Buddha Nagar, Uttar Pradesh India

Dr.Neelendra Badal
Department of Electrical Engineering,
Kamla Nehru Institute of Technology, Sultanpur,
Uttar Pradesh, India

***Abstract*— Various Handover techniques and modifications for Handover in MIPv6 come into light during past few years. Still the problem remains, such as quality of services, better resource utilization during Handover and Handover latency. This paper focuses on such problems within various Handover techniques and proposes some modifications to reduce the Handover latency. This also improves the quality of services related with Handover in MIPv6. Experimental results presented in this paper shows that the Handover latency in MIPv6 will be reduced by applying these proposed modifications. This paper is organized in following sections: section I gives the introduction, Section II presents the Basic operations for MIPv6 Handover, Section III focuses on the related work with background, Section IV proposes modifications and Section V concludes the paper while underlining future prospects in this domain.**

Keywords- Handover; CoA; Home Agent; Foreign Agent; MIPv4; MIPv6.

I. INTRODUCTION

Various Handover techniques for MIPv6 is applied so that the user enjoys continuous internet connectivity and avoids rebooting their application as they move from one subnet to another subnet.

Internet connectivity is based on certain protocols such as Transmission Control Protocol (TCP) and Internet Protocol (IP). Such protocols require a unique IP Address for identifying the physical location of the Mobile Node (MN) for setting up the connection but the mobility of the MN is only achieved if its IP address keeps on changing from one subnet to another. To overcome this problem, Mobile IP was developed which removes the problem by providing MN with two types of addresses, i.e. first, a Home Address (HA) that does not change as the node moves and second, a Care-of-Address (CoA) that keeps on changing as the MN moves from one subnet to other subnet [4].

Mobility support was first incorporated using Internet Protocol Version 4(IPv4) bringing forth MIPv4,later on Internet Engineering Task Force (IETF) developed another Mobile IP that is MIPv6.The major factor that led to switching from MIPv4 to MIPv6 are Route Optimization in MIPv6 and Triangular Routing problem that existed in MIPv4.

Until now, there are approaches to solve the problem of Handover such as L3- Driven Fast Handover which not only uses network layer (L3-layer) information but also uses the link-layer (L2-Layer) information for better Handover process and Resource Efficient CoA Provisioning which makes use of the various caches such as active proxy cache and active garbage cache for better performance in terms of Handover and quality of services.

The basic operations related to Handover are presented in the next section.

II. MIPV6 BASIC OPERATIONS: A QUICK LOOK

MIPv6 protocol was developed which allows MN to be communicated and reachable while moving around in IPv6 internet. MIPv6 performs mobility with the help of three addresses such as HA (static address), CoA and Link Layer Address (Prefix).

Packets are to be transmitted to the MN using HA without actually bothering about MN current point of attachment to the internet. Only HA is not sufficient at all, the CoA generated by the Foreign Agent (FA) is also required. After knowing CoA, a correspondent node or HA can send the desired packets to the MN. So, these two addresses are sufficient enough for the proper flow of the packets. HA is a router on a MN's home network which maintains current location information for the mobile. Now, the question arises as to what for the Link Layer address is used. In brief, it is used for the better Handover. Handover or Handoff refers to the process of transferring an ongoing call or data session from one subnet connected to the next subnet where the mobile node is going to get attached. While, Correspondent Node (CN) is a peer node with which a MN is communicating. It may be either mobile or stationary.

When a MN moves from HA to FA, the actual communication starts between MN and CN after performing the following processes as mentioned also in figure 1.

A. Agent Solicitation

It refers to messages that are sent by the MNs which is looking for a Router to carry out its previous communication activity with its HA or Correspondent node [2].

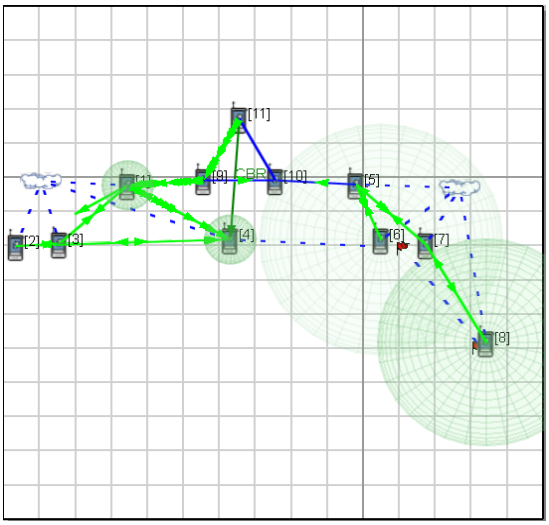


Figure 1: Qualnet Simulated example of Process Involved in Handover

B. Router Discovery

In this process, the Access Router (AR) multicasts Router Advertisement message on a loosely periodic basis and MN senses these messages to determine whether it is in the same AR or it has switched to other Access Routers domain.

C. Address Configuration.

As shown in Fig.1 MN configures a new global IP address called Care-of-Address by the help of prefix received which is present in the Router Advertisement message.

D. Movement Detection.

As shown in Fig.1 the Router Advertisement message contains a prefix that is generated by all the access routers but these prefixes are different for different routers respectively. This allows the MN to check for the change in prefix and, as soon as, it detects the change in prefix, it is able to decide that it has now changed its domain [2].

E. Mobile IPv6 Registration

As shown in Fig.1 After following all these previous processes, the MN comes to know that it has left its previous domain and has entered into domain of new access router. Now, MN has got the responsibility to get its New IP Address i.e. CoA registered with its HA and the Correspondent Node [5].

To make the registration more authentic and appropriate following processes are followed:

a) a) Home Registration

The MN sends a Binding Update Message to the HA after notifying it about the New Care-of-Address (CoA) and, in response, the HA sends a Binding Acknowledgement message.

b) b) Correspondent Registration

As shown in Fig.1 In Correspondent Registration, the Binding Update message are send to the Correspondent Node and, in response, Correspondent Node sends the Binding Acknowledgement message.

III. Related Work and Background

To improve the Handover process various techniques are described in this section to improve Handover process and quality of services. The various techniques are as follows:-

A. L3 (Network Layer) Driven Fast Handover Approach

The Handover is the main concern in MIPv6.The bottleneck of Handover is due to the fact that is that network layer use only the network layer information to detect whether the Handover had taken place or not. To solve this problem several fast Handover approaches came in to lime light which started using link layer information (L2) to speed the process of Handover [12].

There are two sub-processes related to the main Handover approach:

Link-Layer Handover (L2):-In this Handover, the MN comes to know about change of AR as MN detects link layer address change to which it is linked with.

Network layer Handover (L3)

L3 Handover consists of two phases:

Preparation Phase:-In this phase a CoA for mobile is generated as well as Duplicate Address Detection (DAD) protocol is executed.

Signaling Phase:-In this phase the CoA is registered with its HA.

The evolution of L3- Driven Handover was not at all abrupt. First of all Normal Handover Sequence was followed then after some time. Later on Handover Sequence was developed Using L2 information. It was followed for some time and then L3- Driven Handover was developed which was, indeed, more efficient than the other two previous approaches [7].

a) Normal Handover Sequence

First of all L2 Handover will take place, whose functioning is not at all known to network layer. After a while the MN would receive a Router Advertisement and only then the L3 Handover would start its sub phases: preparation phase and signaling phase will occur. However the figure clearly establishes that there is delay of few seconds before L3 Handover starts [6].

b) Handover sequence Using L2 Information

As soon as the L2 Handover is finished, the link layer notifies the network layer of the end of L2 Handover. After getting this notification, the network layer sends the Router Solicitation message to the AR and, consequently, AR starts sending router advertisement message [7].

c) L3- Driven Fast Handover using L2 Information

As shown in the Ref [7] Fig.2 the link-layer senses the link quality and it notifies the network-layer. As soon as the link is down below the threshold the network- layer correspondingly executes the preparation phase immediately.

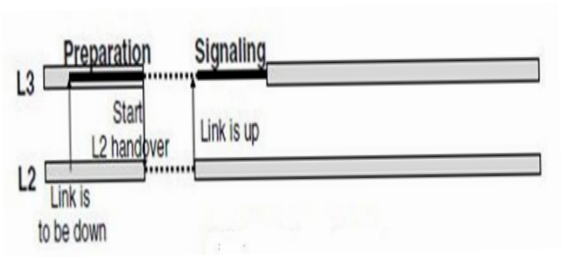


Figure 2: L3-Driven Handover

When the link layer finishes the L2 Handover, it informs the network layer for executing its signaling phase. It is important to know that the bottleneck of L3-Driven approach lies in being subservient to device dependent information. It is manifested through the radio wave strength by which to link layer can judge exact signal strength and notify the network layer that the link is down.

B. Resource Efficient Care-of-Address Provisioning

Duplicate Address Detection (DAD) protocol is followed so that each MN is using a unique address CoA i.e. no other node existed within the same domain subnet using the same global IP address (CoA) but it was found that DAD was a time consuming process ,which might interrupt the seamless Handover. in e.g. RFC 2462 DAD algorithm took more than 1000 millisecond to complete the DAD process which was not at all viable[3].

After that a protocol called aDAD (advance DAD) was developed. The major advantage of aDAD is that it reserves unique New Care–of-Address in advance, as a result, it almost eliminated the latency needed for the Address Configuration and Confirmation .Another major advantage of aDAD is that it follows a concept called "Piggybacking" in accordance with which the MN sends a message called Router Solicitation to the AR and then the AR reverts the same message giving reply in the form of Router Advertisement [4].The major drawback of this protocol is that it uses excessively a lot of network resources such as bandwidth to generate new CoA and verify its uniqueness as aDAD generates New CoA and checks their uniqueness one by one [11].

Third protocol is called Agent-based-DAD (XDAD).It's the most efficient approach for providing New Care-of –Address ,the advantage of this protocol is that it readily decreases the latency during Handover. Another advantage is that it also reserves New CoA in an optimized and effective way. It generates New CoA and stores them in a cache [10].

XDAD uses two types of cache:-Active Proxy Cache and Active Garbage Cache. Active Proxy Cache contains the newly generated CoA and the Active Proxy Garbage contains the CoA which MN relinquishes as soon as it leaves the subnet.

Actual Processes Involved In XDAD are:-

After receiving a solicitation message AR tries to reuse the CoA from the Active Proxy Garbage. If it doesn't get it from there, it uses store Active Proxy Cache address subsequently [6]. Access routers then check the uniqueness of the generated CoA. If it is unique, it sends back the CoA within the same message as a Router Advertisement message through Piggybacking process. Otherwise AR drops the CoA and regenerates it [8].

As soon as the CoA is assigned to a MN, CoA is deleted from the cache.

Therefore it is observed while reviewing various paper that there exist following challenges such as for L3 driven approach use of devices to gauge whether the link is down or up and for Resource efficient approach the problem is use of various caches that directly affect the quality of service. Various modifications for challenges are proposed in subsequent sections.

IV. PROPOSED MODIFICATION

There are two modifications proposed in this sections these are described as follows:-

A. Modification 1: In Domain Of L3-Driven Fast Handover Approach

In the previously defined L3-Driven Fast Handover approach, there is need to frequently check whether link is down or not, and if the link is down that means that the present value of signal strength is below than threshold value. It is important to note here that L3–Driven Handover requires some of the devices dependent information. So this dependence on the device for constant measuring and monitoring the signal strength is not at all required as it will frequently create problems such as wrong measurement of signal strength in a real dynamic environment.

To remove this device dependent problem, the internal entities such as MN and AR should decide about the weak signal strength (Link). To know whether the link is down or not MN sends a packet to the AR and starts a timer, the router will revert the same packet back to the MN. This process is done when MN is very much connected to the access router. The running timer will calculate the average time taken by the packet to traverse to and fro i.e. from MN to AR and from AR to the MN. Standard measurement of average time will be calculated by sending the packets at least 5 times. Then the average of separate to and fro dispatching will be calculated and compared if these values remain same in all 5 remittances. Only then it will be set as standard time within the subnet.

Now the MN will keep on moving and sending this packet. After some time, it will happen that packet would start taking more time to reach MN, because when MN moves away beyond the reach of present Access Router, packets will not be able to reach the MN within set standard limit of the time.

Now to enhance the efficiency of proposed modified version of MIPv6, it is also seen that if the to and fro dispatch time is more than standard time and shows increasing trend of delay for 3 subsequent packets, then the device is so customized that it understands the feeble signal strength immediately and starts multicasting solicitation message to the adjacent router. for e.g. in Fig.3.the to and fro dispatch of three consecutive dispatch are 2ms,7ms,10ms as a result it will again watch the another set of 3 to and fro dispatch and if it shows the increasing trend in these values the Handover will take place.

If the to and fro dispatch time of 3 subsequent packet for this router is less than the standard time, the MN is not going to immediately perform the L2 Handover. Whereas in Fig.4 the to and fro dispatch of three consecutive dispatch are 2ms, 10ms, 3ms. So Handover will not take place. This process finally eliminates the need of device-based-signal-based-monitoring system and also overcome the delay on account of it.

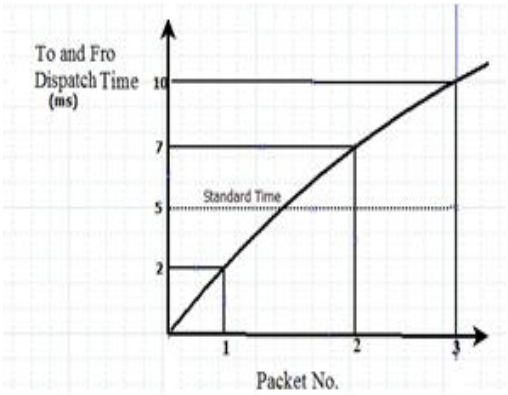


Figure.3 Handover is required as To and Fro Dispatch time keeps on increasing with the time as compared to Standard Measurement

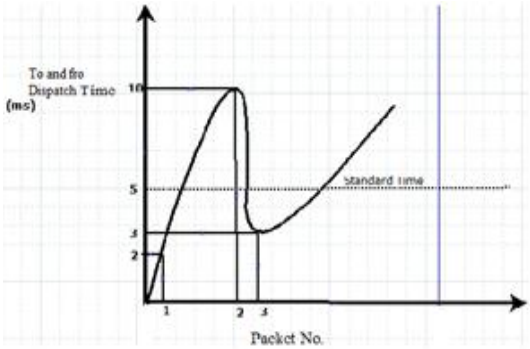


Figure.4 No Handover is required as To and Fro Dispatch time keeps on fluctuating with the increase in time as compared to Standard Measurement

B. Modification 2: In Domain of Resource Efficient CoA Provisioning

In the previously defined Resource-Efficient CoA Provisioning two caches are used:

a) Active Proxy cache.
b) Active Proxy Garbage.

These two caches are being managed by the Access Routers which will definitely increase to its overhead. To subside the overhead, it is recommended to reduce the no. of caches to one. Now, instead of using two caches, scope is reduced to one cache and named this cache as Working Cache. This Working Cache will store and allocate new unique CoA to the MN.

The allocated CoA will be removed from the Working Cache and as soon as the MN leaves the subnet, again that CoA is reallocated to the Working Cache for reuse purpose. The CoA is generated by the help of an algorithm and stored in the Working Cache.

The algorithm will always generate an unique CoA that will remove the need of running DAD protocol and multicasting Neighbor Solicitation messages as every CoA is unique. If the need of running DAD protocol and multicasting Neighbor Solicitation is removed, it will definitely reduce the latency.

Now the question arises that if the same algorithm is used by all the Access Routers to create unique CoA then what will differentiate the CoA if MN moves from one network to another and it will again get the same CoA in another Access Router. To resolve this problem use of the Link-Layer Address Prefix is suggested as it is very much unique for a subnet. The Prefix will be attached to the CoA and that's what will make it Unique within the subnet and outside the subnet.

V. CONCLUSION AND FUTURE SCOPE

This paper focuses on L3 Driven fast handover technique and Resource efficient CoA provisioning techniques and proposes some viable comprehensive modification to enhance the quality of services and resolve the problems of Handover in MIPv6. Experimental results show that the challenges during Handover can be drastically improved by applying the proposed comprehensive modification. Hopefully, if the proposed radical changes are incorporated in the existing MIPv6, it will not only bring about large scale viability in the system but also make it cost-effective, robust, and dynamic all the more.

The work may be further extended in the domain of care-of-address provisioning and fast Handover using the techniques as proposed in this paper.

REFERENCES

[1] Bi-Lynn Ong, Suhaidi Hassan, "Interworking of Protocols in IPv6 Mobility Management" IEEE International Conference on Telecommunications and Malaysia International Conference on Communications, May 2007.

[2] Christian Vogt "A Comprehensive and Efficient Handoff Procedure for IPv6 Mobility Support", Proceedings of the 2006 International Symposium on a World of Wireless, Mobile and Multimedia Network (WoWMoM'o0),IEEE, September 2009.

[3] D. Johnson, C. Perkins, and J. Arkko, "Mobility support in IPv6", RFC 6275, July 2011.

[4] Deng Ya-ping, Wu Ying-qiu, "Research on HMIPv6 Handover Latency of Improved DAD Policy", IEEE, February 2010.

[5] Gaogang XIE, Ji CHEN, Hongxia ZHENG, Jianhua YANG, "Handover Latency of MIPv6 Implementation in Linux" ,IEEE GLOBECOM 2007 proceedings, July 2007.

[6] Longjiang Li, Yuming Mao, Yonggang Li, "Resource-efficient Care-of Address Provisioning for Seamless IPv6 Mobility Support.", IEEE, July 2008.

[7] Kazutaka GOGO, Rie SHIBUI, Fumio TERAOKA " An L3-Driven Fast Handover Mechanism in IPv6 Mobility" Proceedings of the International Symposium on Applications and the Internet Workshops (SAINTW 09), IEEE, January 2009.

[8] P. Sangheon and C. Yanghee, "Performance analysis of fast handover in Mobile IPv6 networks," Lecture Notes in Computer Science (LNCS), vol.2775, pp.679-691, Springer-Verlag, 2003

[9] Reza Malekian ," The Study of Handover in Mobile IPv6 Networks" , IEEE International Conference on Telecommunications and Malaysia International Conference on Communication, 2008

[10] R. Koodli , "Fast Handovers for Mobile IPv6," RFC 4068,2005

[11] S. Menezes, "An Efficient Handover Scheme Based on Fast Mobile IPv6", IEEE 802.21Media Independent Handoff Working Group , 2007

[12] Seung Wook Moon ," Reducing Handover Delay in Mobile IPv6 by cooperating with Layer 2 and Layer 3 Handovers", IEEE 802.21 Media Independent Handoff Working Group, 2007

AUTHORS PROFILE

Mr.Purnendu Shekhar Pandey is a M.Tech Student at Gautam Buddha University. The author has done his B.Tech from Noida Institute of Engineering and Technology, Uttar Pradesh Technical University, Uttar Pradesh. The author has published and presented various papers in national and international conferences.

Dr.Neelendra Badal is an Asst. Prof. in the Department of Computer Science & Engineering at KNIT,Sultanpur (U.P), INDIA. He received B.E. (1997) from Bundelkhand Institute of Technology (BIET), Jhansi in Computer Science & Engineering, M.E. (2001) in Communication, Control and Networking from Madhav Institute of Technology and Science (MITS), Gwalior and PhD (2009) in Computer Science & Engineering from Motilal Nehru National Institute of Technology (MNNIT), Allahabad. He is Chartered Engineer (CE) from Institution of Engineers (IE), India. He is a Life Member of IE, IETE, ISTE and CSI-India. He has published about 30 papers in International/National Journals, conferences and seminars. His research interests are evinced at Distributed System, Parallel Processing, GIS, Data Warehouse & Data mining, Software engineering and Networking.

Different Protocols for High Speed Networks

Dr.Srinivasa Rao Angajala
Professor, Mekapati Rajamohan Reddy Institute of Technology & Science,
Udayagiri – 524 226, A.P., India.

***Abstract*—New challenges arise with the presence of various types of physical links, such as wireless networks, high speed and satellite in today's ever-changing network. It is clear that the TCP throughput deteriorates in high-speed networks with large bandwidth-delay product, and new congestion control algorithms have been proposed to address such deterioration. Traditional TCP protocols treat all packet loss as a sign of congestion. Their inability to recognize non-congestion related packet loss has significant effects on the communication efficiency. The proposed protocols such as TCP Adaptive Westwood, Scalable TCP, HS-TCP, BIC-TCP, FAST-TCP and H-TCP all have some improvement in functionality over the traditional TCP protocols. This survey gives a summarization of all the protocols for high speed networks.**

Keywords- TCP protocols; high speed networks; TCP congestion.

I. INTRODUCTION

The Transmission Control Protocol (TCP) congestion control algorithm is successful in making the functioning of internet efficiently. However, it shows very poor performance over the networks with high bandwidth - delay product paths. The problem starts from the fact that the standard AIMD (Additive increase and multiplicative decrease) congestion control algorithm increases the congestion window too slowly. This would lead to probably long file transfer times. With the present algorithms, latencies of only a few tens of milliseconds are quire sufficient to create bandwidth-delay products that yield poor throughput performance [1]. Therefore innovative schemes were designed to provide a solution to certain issues. TCP designed mainly for wired networks is a connection oriented transport protocol that provides reliable data communications. The reason for the performance degradation is that the congestion control mechanism in TCP cannot distinguish between the packet loss caused by wireless link error and that caused by network congestion, thus, reacting to the loss by reducing its congestion window (cwnd). Therefore, these in appropriate reductions of the cwnd lead to unnecessary throughput degradation for TCP applications [2].

A solution to the problem has been given by many authors. The main aim is to increase the rate at which cwnd is increased and thereby shortens the congestion epoch duration and also it should be suitable to standard TCP paths with low bandwidth – delay product (BDP). The previous research on these lines includes the HS-TCP proposal [8], the scalable TCP [7] and the FAST – TCP [11] and many more recent proposals include BIC-TCP [8] and H-TCP [10]. The rest of the document is organized as follows. Section 2 offers an overview of the basic design of TCP protocol and its mechanisms. Section 3 discusses various protocols for high speed networks provided in the previous research papers.

Section 4 describes the experimental results from the related work and Section 5 concludes the review of the study of various TCP protocols for high speed networks.

II. BASIC DESING OF TCP PROTOCOLS

TCP is a transport layer protocol that provides a reliable and in-order delivery of data between two hosts. It is a defensive protocol highly sensitive to network congestion.TCP issues an acknowledgement packet (ACK) as a response to a successfully delivered packet to ensure a reliable communication. The standard TCP congestion control mechanism is based upon a sliding window, which defines the number of packets injected in the network. The congestion window value is updated after the reception of an ACK and upon the detection of a packet loss. The reception of an ACK is interpreted by TCP as a signal of available bandwidth, so that the congestion window value, W, may be increased. The increase is fast at the beginning of a new connection (the so-called slow-start phase), where the TCP sender enlarges W by one segment for each received ACK. In this way, the congestion window grows exponentially on a Round Trip Time (RTT) basis. Once a given threshold value is crossed, the connection enters the congestion avoidance phase, where the TCP sender gently increases its transmission rate in a linear fashion over RTT (the congestion window is increased by 1 W upon an ACK reception).

The decrease phase is triggered by the detection of a packet loss. A packet loss can be detected either by a timeout expiration or by the reception of 3 duplicate ACKs (correspondently, the congestion window is halved). Since the timeout granularity is fairly large, the first case is interpreted by TCP as a signal of severe network congestion so that it reduces its transmission rate to the minimum. In the other case, since some packets have been correctly received after the lost one(s), the congestion is assumed to be a transient phenomenon It is easy to verify that this mechanism is pretty inefficient in the presence of large W values (or, equivalently, large BDP values). Indeed, let us consider an error occurring in the presence of a congestion window value of W. Since W grows approximately by 1 for RTT, the time it takes to go back from W/2 to W equals $Trec = (W \cdot RTT) / 2$. This will be called as the time to recover from a loss. The fact that Trec depends linearly on W suggests that the TCP's AIMD mechanism does not scale well with BDP. This is one of the main reasons that have motivated the study of congestion control mechanisms able to perform well in

a large BDP scenario. The packet loss is detected by (1) a timeout and (2) duplicate acknowledgement (ACK). A timeout occurs when the TCP sender does not receive any acknowledgement from the receiver even after a prescribed time. When timeout occurs, TCP treats the situation as network congestion and performs slow start. In the second case, when the TCP sender receives duplicate ACK, it identifies receiver received out-of order packets. TCP enters fast retransmit and fast recovery algorithm.

III. VARIANTS OF TCP

A. TCP Adaptive Westwood

TCP Adaptive Westwood (TCP-AW) is a combination of the best features from both TCP Westwood ABSE (TCP-ABSE) [4] and TCP Adaptive Reno (TCP-AReno) [5]. TCP-ABSE's best feature is its eligible rate estimation (ERE) mechanism, which helps predict imminent congestion. Essentially, the congestion window is adjusted according to the network congestion rate, and is not based on the traditional delay model. TCP-AReno's best feature is its use of packet loss interval time to adjust the congestion window, instead of using bandwidth estimation.

This approach improves RTT-fairness even when accurate bandwidth estimation is not available. For the high-speed scenario, TCP-AW[6] remarkably improved the aggregate throughput of the control group. This is due to the delay-based nature of the protocol. The protocol feels the presence of incoming TCP standard Reno flows, and then reacts accordingly to grab the extra bandwidth whenever the bottleneck is less loaded. TCP Adaptive Westwood shows good throughput in High Speed networks and performs safely. TCP-AW obtains a substantial improvement in coexistence due to its embedded loss discriminator component.

B. HS-TCP

HS-TCP was introduced to achieve high throughput in high bandwidth-delay product links without requiring unrealistically low packet loss rates. The HS-TCP modifies the standard TCP's Additive increase and multiplicative decrease (AIMD) algorithm to improve its loss discovery time. Such alteration would only be effective when higher congestion windows are encountered. This implies that if the congestion window is smaller than a given threshold, it makes use of the Standard (TCP Tahoe) AIMD algorithm, otherwise the High Speed algorithm is used [3]. The Standard AIMD algorithm upon receipt of ACK and in the event of congestion, the window is respectively given as:

$$w = w + 1/w$$
$$w = 0.5*w$$

Parameters a and b for the increase and the decrease of the AIMD algorithm are fixed at 1 and 0.5 respectively. The modified HS-TCP algorithm is as follows: upon the receipt of acknowledgement,

$$w = w + a(w)/ w$$

And when congestion is encountered, the window (w) is,

$$w = w - b(w)w$$

The increase and decrease parameters thus vary depending on the current value of the congestion window.

C. Scalable TCP

STCP seeks to improve the loss recovery time of Standard TCP; this idea mirrors that of the HS-TCP [3]. For standard TCP and HS-TCP connection, the packet loss recovery times increase (or reduce) in a proportion as the connection's window size and round trip time (RTT) does. In a Scalable TCP connection, packet loss recovery times are proportional to connection's RTT only. For the STCP [7], the slow start phase of the standard TCP algorithm is not modified, but its congestion avoidance phase is modifies thus: for every acknowledgement received in a RTT, the congestion window (cwnd) is

cwnd = cwnd + 1/cwnd (for Std TCP)

cwnd = cwnd + 0.01 (for STCP) -------(1)

and when congestion is encountered in a given RTT,

cwnd = cwnd - cwnd * 0.5 (for Std TCP)

cwnd = cwnd – cwnd * 0.125 (for STCP) ----(2)

Similarly the evolution of STCP's cwnd is similar to that of the HS-TCP; its threshold window size and the modified algorithm in equation (1) and (2) are used only when the size of the congestion window exceeds threshold window size. The values of 0.01and 0.125 are suggested for the increase and the decrease parameters. STCP default value for the threshold window size is given as 16 segments [7].

D. BIC – TCP

BIC-TCP [8] employs a binary search algorithm to update its congestion window.Briefly a variable w1 is maintained which holds a value halfway between the values of cwnd just before and Just after the last loss event. The window update rule seeks to rapidly increase its window beyond a specified distance Smax from w1, and update cwnd more slowly when its value is close tow1.

Multiplicative backoff of cwnd is used on detecting a packet loss with a backoff factor β of 0.8. It also implements an algorithm whereby upon low utilization detection, it increases its window more aggressively. This is controlled with two factors namely, low utility and utility check. In order to maintain backwards compatibility, it uses the standard TCP update parameters when cwnd is below the threshold.

E. FAST – TCP

FAST – TCP [9] is a delay based algorithm. It also includes rate packing. Rate pacing is a functional change and is thus it can be viewed as a part of congestion control algorithm. The FAST TCP flows typically converge quickly initially, flows may later diverge again to create significant and sustained unfairness. The main drawback of this is where the threshold is somewhat higher, owing to the standing queue created by the delay-based congestion control action used here.

F. H – TCP

H-TCP [10] uses the elapsed time Δ since the last congestion event, rather than cwnd, to indicate path bandwidth-delay product and the AIMD increase parameter is

varied as a function of Δ. The AIMD increase parameter is also scaled with path round – trip time to mitigate unfairness between competing flows with different round-trip times. The AIMD decrease factor is adjusted to improve link utilization based on an estimate of the queue provisioning on a path.

IV. EXPERIMENTAL RESULTS FROM THE RELATED WORK

The performance analysis relevant to the present study is derived from many papers. In [7], Kelly presents an experimental comparison of the aggregate throughput performance of scalable-TCP and standard TCP. In [11], aggregate throughput measurements are presented for FAST-TCP and TCP-Reno. In all of these studies, measurements focus on aggregate throughput i.e., link utilization. Here efficiency as a function of queue size is not considered, nor fairness, friendliness, responsiveness and convergence times.

In [9], throughput and cwnd time histories of FAST-TCP, HS-TCP, Scalable-TCP, and TCP-Reno are presented for a lab scale experiment test bed. Aggregate throughput, throughput Fairness and a number of other measures are presented. However, results are confined solely to an 800-Mb/s bottleneck link with a 2000-packet buffer. The impact of link rate, RTT, queue size, and level of Web traffic on fairness and responsiveness are not considered nor is the impact of queue size on efficiency. In [8], NS simulation results are presented comparing the performance of HS-TCP, Scalable-TCP, BIC-TCP, and standard TCP.

V. CONCLUSION

This article concludes by presenting a comprehensive survey of current research on running TCP in high speed networks from various experimental results evaluating the performance of Scalable-TCP, HS-TCP, BIC-TCP, FAST TCP and H-TCP. All the protocols studied are all successful in improving the link utilization in a relatively static environment with long-lived flows. And many of them showed poor responsiveness to changing network conditions. Though there are various schemas and mechanisms proposed, there is no single mechanism that can overcome the unreliable nature of network in a reliable way.

Each and every mechanism has its own advantages and disadvantages. In short, any mechanism will be effective based on the factors that are to be taken into consideration. To conclude this area is not completely explored to its maximum and still lot more research can be done towards establishing a basis for the development of new protocols.

REFERENCES

[1] M.Allman, “TCP congestion control with appropriate byte counting’, IETF RFC 3465, Feb 2003.

[2] Lakshman, T. V. and Madhow, U.: The performance of TCP/IP for networks with high bandwidth- delay products and random loss. IEEE/ACM Transactions on Networking, Vol. 5, 336–350, 1997.

[3] Sally Floyd, RFC 3649, “Highspeed TCP for large congestion window”, 2003.

[4] R. Wang, M. Valla, M. Sanadidi, and M.Gerla, “Using adaptive rate estimation to provide enhanced and robust transport over heterogeneous networks,” IEEE ICNP, Nov 2002.

[5] Cesar Marcondes, Jerrid Matthews, Robert Chen, “ A cross comparison of advanced TCP protocols in High speed and satellite environments”,IEEE, pp 179-185, 2008.

[6] T. Kelly, “Scalable TCP: Improving Performances in High speed Wide Area networks,” Computer Commn., Rev., vol.33, no.2, pp. 83-91, April 2003.

[7] C.Jin,D.X.Wei,S.H.Low,G.Buhrmaster, J. Bunn, D. H. Choe, R. L. A. Cottrell, J. C. Doyle, W.Feng, O. Martin, H. Newmann, F. Paganini, S. Ravot and S. Singh, “FAST TCP : From theory to experiments”, IEEE Network, vol. 19, no. 1, pp.4-11, 2005.

[8] L.Xu, K.Harfoush and I.Rhee, “Binary increase obstacle control for fast long-distance networks”, in Proc. IEEE INFOCOM, HongKong, 2004.

[9] D.J.Leith and R.N.Shorten, “H-TCP protocol for high speed long distance networks,”presented at thend Workshop Protocols Fast Long Distance Networks, Argonne, Canada, 2004.

AUTHORS PROFILE

Dr. ANGAJALA SRINIVASA RAO was born on 1964 and received Ph.D from University of Allahabad in 2008 and M.S(CSE) from Donetsk State Technical University, Ukraine in 1992. At present he is working as a Principal / Professor in CSE Dept. in Mekapati Rajamohan Reddy Institute of Technology&Science**,** Udayagiri- 524226, Sri Potti Sreramulu Nellore Dist., Andhra Pradesh. He wrote 5 books and published by Vikas Publications, New Delhi for various Universities in India. He is a member of IEEE, CSI, ISTE, IACSIT, CSTA, IWA AND IAENG and also attended no. of National and International Conferences.

Wideband Wireless Access Systems Interference Robustness: Its Effect on Quality of Video Streaming

[1]Aderemi A. Atayero, [2]Oleg I. Sheluhin
[1,4]Department of Electrical and Information Engineering
Covenant University
Ota, Nigeria

[3]Yuri A. Ivanov, [4]Julet O. Iruemi
[2,3]Department of Information Security
Moscow Tech. Univ. of Communication and Informatics
Moscow, Russia

***Abstract*—A necessary requirement incumbent on any information communication system and/or network is the capacity to transmit information with a predefined degree of accuracy in the presence of inevitable interference. The transmission of audio and video streaming services over different conduits (wireless access systems, Internet, etc.) is becoming ever more popular. As should be expected, this widespread increase is accompanied by the attendant new and difficult task of maintaining the quality of service of streaming video. The use of very accurate coding techniques for transmissions over wireless networks alone cannot guarantee a complete eradication of distortions characteristic of the video signal. A software-hardware composite system has been developed for investigating the effect of single bit error and bit packet errors in wideband wireless access systems on the quality of H.264/AVC standard bursty video streams. Numerical results of the modeling and analysis of the effect of interference robustness on quality of video streaming are presented and discussed.**

Keywords-codec; H.264/AVC; polynomial approximation coding; signal-to-noise ratio; video streaming.

I. INTRODUCTION

One of the most important Quality of Service (QoS) parameters for wireless networks is the probability of occurrence of bit and packet errors measured by the Bit Error Rate (BER) and Packet Error Rates (PER) respectively. Neither single packet losses nor single bit errors can provide a comprehensive imitation modeling of fading channels. In digital systems, errors often occur in packets as a result of transmission conditions. Specifically, signals are attenuated during transmission and this consequently leads to grouping (packetization) of errors. A group of erroneous packets is essentially a sequence of packets that are either lost in transit or received with error after transmission over a communication channel within a given period of time. Burst Error Length (BEL) is defined as the number of erroneous packets included in a given group of errors.

II. METHOD

The H.264/AVC standard is a compendium of innovations and improvements on prior video coding technologies vis-à-vis enhancement of coding efficiency and effective usage over a wide gamut of networks and applications [1]. For a complete analysis of the impact of errors on resultant signal quality, we investigate the influence of the conduit's (i.e. wireless transmission medium's) robustness on the perceivable quality of streaming video standard H.264/AVC using the developed hardware and software complex (HSC) [2], [3].

Objective and subjective indicators of video quality were obtained using methods described in [3]. For qualitative assessment, it is imperative to have the video file data before transmission over the network (on the transmitting end), and after reception from the network (at the receiving end).

Data required for the qualitative assessment at the transmitting end are:

a)the original unencoded video in YUV format,

b)the encoded video in MPEG-4,

c)transmission start time and

d)type of each packet sent to the network.

The following data are required for qualitative assessment at the receive end:

a)time of reception of each packet from the network,

b)type of each packet received from the network,

c)the encoded video (possibly distorted) in MPEG-4 format, and

d)the decoded video in YUV format for display.

We performed data evaluation by comparing the transmitted and received files.

Structure of the Hardware-Software Complex (HSC)

Data processing is carried out in the three phases described below:

First phase: the time taken in sending and receiving each packet on both sides as well as the packet type are analyzed. This results in a record of the type of frame and the time elapsed between transmitted and received packets. The distorted video file at the receive end is restored using the originally encoded video file and information about lost packets. Subsequently, the video is decoded for playback to the viewer.

Assessment of video quality is done at this stage. Video quality indicators always require a comparison of the received (possibly distorted) video frame and the corresponding source frame. In the case of a total loss of frame in transit, the necessary frame synchronization before and after transmission over the network becomes impossible.

Second phase: In this phase of data processing, the problem of quality assessment is resolved based on the analysis of information about frame loses. Substituting the last relayed frame for the lost frame restores frame synchronization. This methodology allows for subsequent frame-wise assessment of video quality.

Third phase: At this stage, the assessment of the quality of decoded video is achieved by means of both the restored and source video files.

Fig. 1 shows a block diagram of the HSC for assessing the quality of streaming video. The schematic diagram reflects the interaction between modules in the transmission of the digital video from a source through the network connection to the viewer. The HSC modules interact with the network by using traces containing all the necessary data listed above. Thus, for proper functioning, the HSC requires two traces, the source video and the decoder. The data network can be considered simply as a two-port black-box that introduces delay, packet loss, and possibly packet rearrangement. The network was simulated based on the aforementioned assumptions [4] in the NS2 environment. A detailed description of the functional modules of the HSC is given in [3].

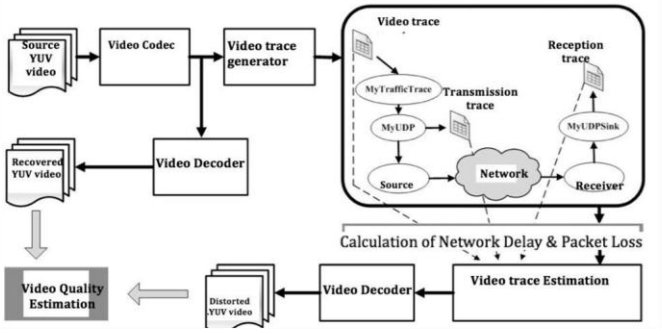


Figure 1. Block diagram of HSC.

III. Presentation of Data and Simulation Preparation

Video encoding begins with color space conversion from RGB to YUV also known as Y, Cr, Cb i.e. one luma and two chroma components [5]. It is common knowledge that there is a significant correlation of color components in any typical image of the RGB format, which makes it an obviously redundant format in terms of compression. The standard television uses a different representation of images, which also employs three components of the signal, but these components are uncorrelated (i.e. void of inter-componential redundancy). R, G and B components are converted to luminance Y component and two color difference components U and V of the YUV format. Since most information is stored in the luminance component, little information is lost if a thinning of the U and V components is done.

Standard test videos in the YUV format may be used as initial test video sequences. However, these videos have limited playback time and hence do not allow for the assessment of change in quality under prolonged video broadcast. Similarly, a vast amount of experimental data cannot be obtained from them. It is for these reasons that we recorded our own 30-minute video in YUV format (send.YUV) with a resolution of 640x480 pixels and frame rate of 25fps using special software.

The first step is to encode the source video to H.264 format (video stream file). This is done by the video codec (a device used for encoding and decoding video signals). Video codecs are usually characterized by a) channel throughput, b) decoded video distortion rate, c) startup latency, d) end-to-end delay, computational complexity, and e) memory capacity. A good codec is one capable of providing the necessary trade-off vis-à-vis these characteristics [5].

In the next step, the encoded video stream is packaged in an MP4-container for onward transport over the network using the User Datagram Protocol (UDP). The result of encoding the original video is an MP4-file. Since it is necessary to evaluate the quality of video transmitted over the network, the need arises to create a spare decoded YUV file from the newly created MPEG layer-4 file, which serves as the control in evaluating the quality of video transmitted over the network, excluding the impact of the codec. It is thus possible to estimate the influence of a wireless network on the received visual video quality, while excluding encoding and decoding losses.

For simulation purposes, it is necessary to create a video trace file that contains the following information: frame number, frame type, frame size, and the number of segments in which the frame is divided into packets. This video trace serves as the input to the simulator network, where the sending and reception of video data occurs. As a result of video transmission over the network, it is necessary to obtain transmission trace files and reception trace files, which contain the following packet data: the transmission/reception time, a unique identifier and trace file size. These two traces are used to determine lost packets in the network. In the end, we obtain files of the sent and received packets containing detailed information about the time of sending from the transmitter and the time of reception by the receiver.

IV. Modeling and Simulation of Transmission over a Wireless Network

The HSC allows for the simulation of the main types of errors encountered when transmitting video data over wireless networks. The two types of simulation required are as listed below:

A. Bit error simulation

Simulated transmission over a wireless channel model with Additive White Gaussian Noise (AWGN) is conducted. In the process of simulation, certain bits in the sequence are distorted (i.e. inverted) with a given probability. The probability value used is defined by the Bit Error Rate (BER).

B. Packet error simulation.

The UDP packets can be manually deleted from the received trace file. This allows for the observation of codec functionality and analysis of change in visual quality in cases of packet loss. At the same time, both the received and undistorted files can be obtained during transmission over an "ideal" channel with unlimited bandwidth and no delay, with subsequent removal of some packets.

V. CALCULATION OF LOSSES AND ESTIMATION OF OBTAINED VIDEO QUALITY

TABLE 2. RELATIONSHIP BETWEEN QUALITY INDICATORS AND BER

PSNR [dB]	MOS [%]	BER	ITU Quality Scale	Picture Degradation
> 37	81–100	$< 1x10^{-4}$	5 EXCELLENT	NOTICEABLE
31–37	61–80	$1x10^{-4} – 4x10^{-4}$	4 GOOD	NOTICEABLE, BUT NOT IRRITATING
25–31	41–60	$4x10^{-4} – 8x10^{-4}$	3 SATISFACTORY	SLIGHTLY IRRITATING
20–35	21–40	$8x10^{-4} – 1x10^{-3}$	2 POOR	IRRITATING
< 20	0–20	$> 1x10^{-3}$	1-VERY POOR	VERY IRRITATING

Calculation of losses given the availability of unique id of the package is quite easily achieved. With the aid of the video trace, each packet is assigned a type. Each package of the assigned type that is not included in the received trace is deemed lost. Loss of frame is calculated for any (and all) frame(s) with a lost packet. If the first packet in a frame is lost, then the whole frame is considered lost since the video decoder cannot decode a frame, which is missing the first part. Assessment of received traces is done by the module for trace assessment (see Fig. 1). The recovered file must be decoded in YUV format.

There are two major methods of estimating the quality of digital video, namely, the subjective and objective methods:

Subjective quality assessment is always based on viewer impression. It is extremely costly, very time consuming and requires specialized equipment. Traditionally, subjective video quality is determined by expert assessment and calculation of the average Mean opinion Score (MOS), which is assigned a value from 1 to 5 (ITU scale) [6], [7], where 1 and 5 represent worst and best received video quality respectively.

Objective video quality assessment is usually done by measuring the average luminance peak Signal-to-Noise Ratio (PSNR). The PSNR is a traditional metric, which allows for the comparison of any two images [8]. The PSNR module evaluates the objective quality of received video stream in polynomial approximation coding (PAC). The end result is the values of PSNR calculated for the original and distorted image (as shown in Fig. 2). MOS values are calculated from the PSNR indicator.

TABLE 1. CHARACTERISTICS OF ENCODED VIDEO

Format	MPEG-4 Part14 (MP4)
Codec	H.264
Bit rate	Constant @ 1150 kbps
Frame frequency	25 fps
Resolution	640 x 480 pixels
GOP type	IBBPBBPBB

VI. RESULTS, ANLYSIS AND DISCUSSION

In order to study the effect of transmission errors on the resulting video quality, the transmission of a 30-minute video over a wireless network with random packet errors in the channel was simulated. Characteristics of sequences used are listed in Table I.

For modeling purposes, the encoded video stream was split into RTP/UDP-packets using the hardware-software tool reported in [3]. Bit error simulation for transmission over a wireless channel was done using an AWGN error generator contained within the PAC structure. Simulation of packet errors during transmission over a wireless channel was done by deleting packets from the received trace file [3]. This allowed us to explore and analyze the change in visual quality during loss of packets. The received and undistorted trace files were obtained for transmission over an "ideal" channel with unlimited bandwidth and no delay in using the NS2 software environment [4], followed by a random removal of packets, according to PER and BER parameters. Quality assessment was carried out using PSNR and MOS indicators, calculated by using hardware and software tools [4]. The standard deviation of the quality of the average PSNR values was calculated using equation (1) [9].

$$S'_{PSNR} = \sqrt{\frac{1}{N-1}\sum_{n=0}^{N-1}(PSNR_n - \overline{PSNR'})^2} \qquad (1)$$

C. Effect of bit error

Fig. 2 shows the effect of BER on the quality of video streaming.

Analysis of the results of streaming video over the simulated wireless network with different values of BER revealed the following:

i) Simulating a wireless channel using AWGN model, and additive, bit errors with a value of BER $\leq 3x10^{-5}$ does not affect the quality of the video. However, when BER $\geq 4x10^{-3}$ packet loss in the network reaches its maximum value of $\geq$ 99,9%.

ii) Objectively, excellent quality of video transmission over a channel can be guaranteed for all bit error probabilities less than $1x10^{-4}$, good quality is in the range of $1x10^{-4}$ to $4x10^{-4}$, satisfactory quality is in the range of $4x10^{-4}$ to $8x10^{-4}$, poor quality is in the range of $8x10^{-4}$ to $1x10^{-3}$, while very bad quality is for any BER $> 1x10^{-3}$.

iii) The histograms of the distribution of PSNR values during simulation and broadcast over a real network in general are of a twin-peak form. One of the peaks characterizes the PSNR value of error-free video stream (the decoder is able to correct bit errors when they are relatively few in the frame). The second peak characterizes PSNR degradation due to the large number of corrupted video frames in fading moments (the decoder is unable to fix large numbers of bit errors). As the number of errors increases, this maximum increases commensurately with a decrease in the second. During transmission, depending on error

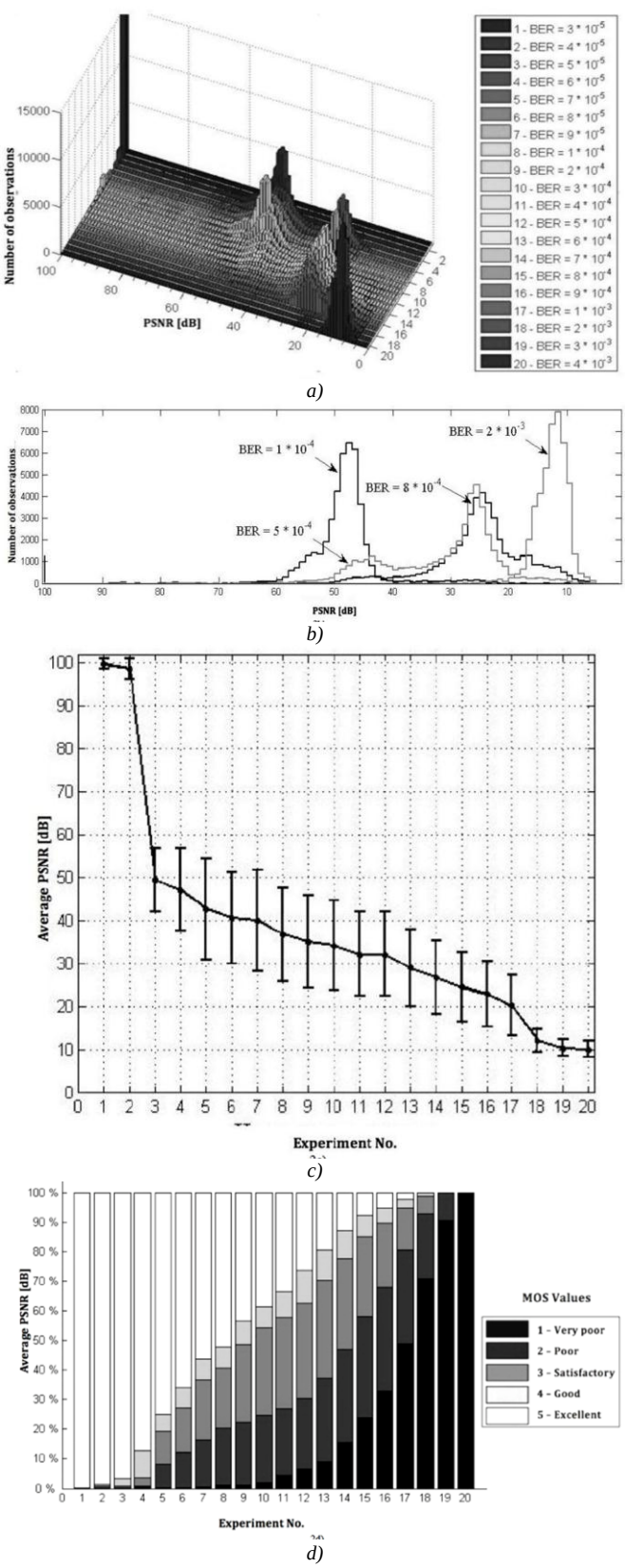


Figure 3. Values of video sequence quality indicators for different values of wireless channel BER: a) –PSNR value distribution histogram; b) – PSNR value distribution histogram for certain values of BER; c) – Quality deviation from average PSNR value; d) – quality gradation for MOS values.

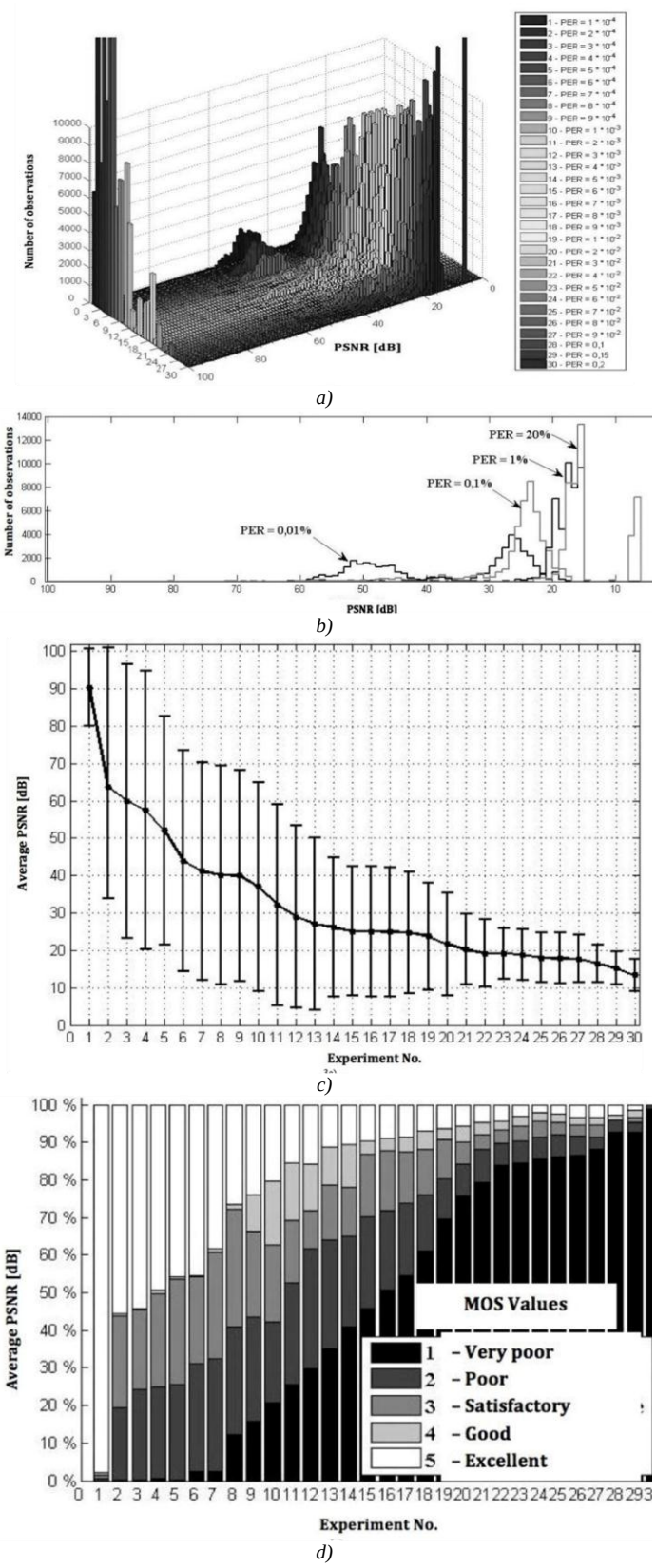


Figure 2. Values of video sequence quality indicators for different values of wireless channel PER: a) –PSNR value distribution histogram; b) – PSNR value distribution histogram for certain values of PER; c) – Quality deviation from average PSNR value; d) – quality gradation for MOS values.

when errors in the communication channel are negligible, the PSNR distribution has only one maximum.

Empirical values of BER transitions from an acceptable quality to the poor, according to the relationship between PSNR and MOS [6], are presented in Table II.

However, the AWGN model does not allow for adequately simulation of a fading channel. Typically, errors are often long term, since high probability of bit loss occurs in specific periods of transmission, e.g. during poor propagation. Attenuation of the transmitted signal results in packetizing (grouping) of errors. Another cause of error grouping can be physical defects of, and failures inherent in the information storage system. When using VLC, bit error occurrence results in group errors or packetization of errors.

B. Effect of packet error

Fig. 3 shows the effect of PER indicator on streaming video quality.

The range of values of PER, within which the resulting quality is maximal (i.e. almost equal to the original) and minimal are indicated. It is shown that with PER ≤ 1x10-4 error does not affect the resultant video quality and can be easily eliminated with decoders and existing methods of error correction. A further change in the quality has a stepwise nature and decreases with increasing PER.

Empirical values of PER transitions from an acceptable quality to the poor, according to the relationship between PSNR and MOS, are presented in Table III.

TABLE 3. RELATIONSHIP BETWEEN QUALITY INDICATORS AND PER

PSNR [dB]	MOS [%]	PER	ITU Quality Scale	Picture Degradation
> 37	81–100	$< 1x10^{-4}$	5 EXCELLENT	NOTICEABLE
31–37	61–80	$1x10^{-3}$– $3x10^{-3}$	4 GOOD	NOTICEABLE, BUT NOT IRRITATING
25–31	41–60	$3x10^{-3}$ – $1x10^{-2}$	3 SATISFACTORY	SLIGHTLY IRRITATING
20–35	21–40	$1x10^{-2}$ – $5x10^{-2}$	2 POOR	IRRITATING
< 20	0–20	$> 5x10^{-2}$	1VERY POOR	VERY IRRITATING

Analyzing the results of streaming video over a simulated wireless network with a given probability of packet loss, we safely conclude that:

a) A PER value of ≤ 1x10-4 in simulation of a wireless network does not affect the video quality. When PER ≤ 1x10-3 impact of errors on video quality is not noticeable and does not irritate during viewing experience. When PER ≥ 0.1, packet loss in the network has the worst effect on visual quality.
b) Objectively, excellent quality of video transmission over a channel can be guaranteed for all packet error probabilities less than 1x10-3, good quality is in the

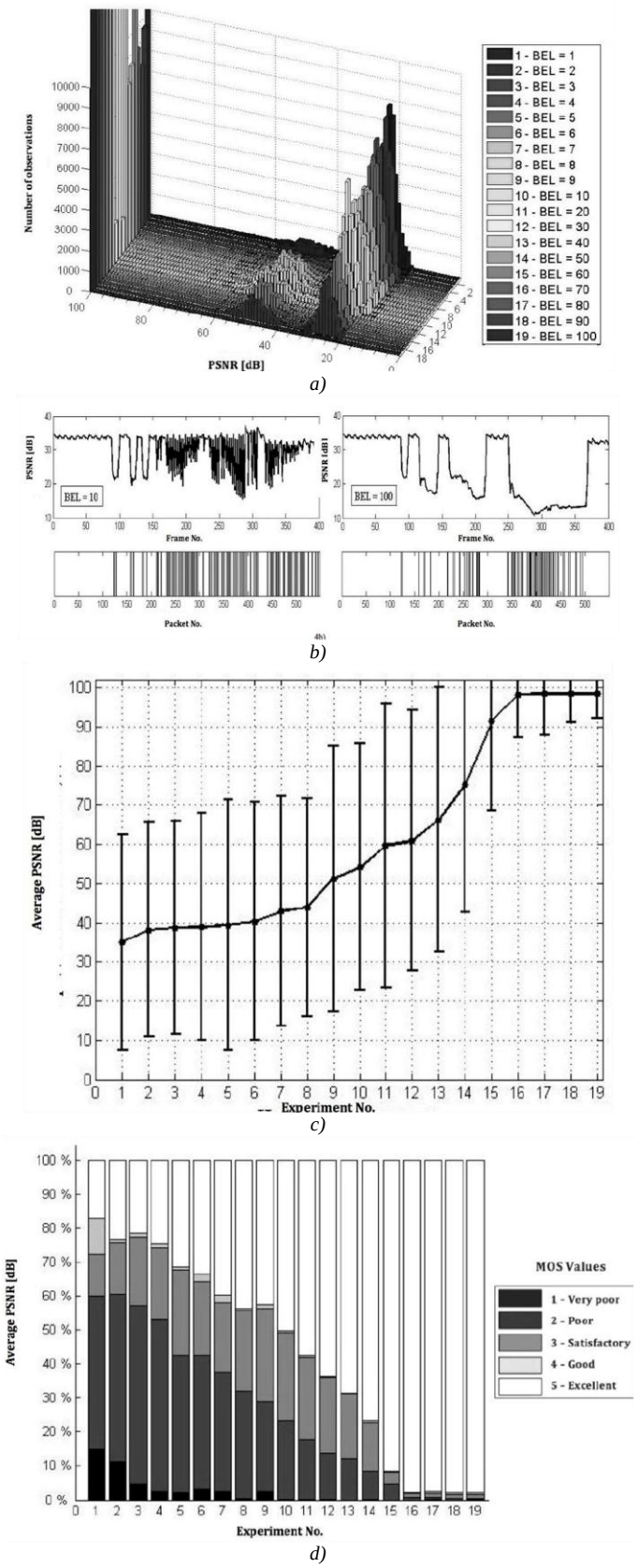


Figure 4. Values of video sequence quality indicators for PER=1x10-3 and varying values of wireless channel BEL: a) –PSNR value distribution histogram; b) – PSNR value and RTP/UDP packet distribution (black spaces correspond to lost packets) for certain values of BEL; c) – Quality deviation from average PSNR value; d) – quality gradation for MOS values.

range of 1x10-3 to 3x10-3, satisfactory quality is in the range of 3x10-3 to 1x10-2, poor quality is in the range of 1x10-2 to 5x10-2, while very bad quality is for any PER > 5x10-2.

Histograms of the distribution of values of PSNR when PER ≤ 6x10-4, in general, have a bimodal shape. One of the peaks characterizes the value of PSNR of video stream distorted due to packet loss. The second maximum characterizes deterioration in the PSNR of dependent frames. As the number of errors increases, one of the peaks increases due to a decrease in the other.

C. Effect of length of error groups

To study the effect of the length of error groups on resultant quality, the simulation of a 30-minute video transfer over a wireless network for the values of PER of $1x10^{-3}$ to $5x10^{-2}$ is repeated, since a visual change in video quality is observed at this range. The simulation of groups of error packets during transmission over a wireless channel was done by means of random deletion of packet groups from the receive trace file with a given BEL. For this particular example BEL=100 implies that the total random number of consecutively deleted packets does not exceed 100. The total sum of erroneous (deleted) packets in the video sequence for the whole experiment given PER=const. remained the same, irrespective of the value of BEL. Fig. 4 shows the effect of BEL on the quality of streaming video for PER=$1x10^{-3}$.

Analyzing the results of streaming video over the simulated wireless network with a given grouping of erroneous packets, we can draw the following conclusions:

v) For PER ≤$1x10^{-3}$ the effect of single packet errors on quality is insignificant and does not irritate the viewing experience.

vi) Histograms of the distribution of values of PSNR have two maxima. One of the peaks characterizes the value of PSNR of video frames distorted due to the loss of packets. The second maximum characterizes the deterioration of PSNR of dependent frames. With increasing quantities BEL is one of the peaks decreases as the number of dependent frames are also reduced, whereas the second peak remains unchanged. This is explained by the fact that the single scattered throughout the video sequence error number of distorted frames is large due to error propagation to dependent frames. An increase in the BEL value leads to a decrease in one of the maxima, since the number of dependent frames also decreases, while the second maximum remains the same. This is due to the fact that under singular errors spread across the whole video sequence, the number of distorted frames is large because of the distribution of errors on dependent frames.

vii) Increasing the length of the error groups leads to an increase in the average quality of the video sequence.

viii) Effect of error groups on the quality is more powerful because of the local concentration of errors. However, the average quality of the video sequence increases with increase in the length of the grouping for a given value of probability of occurrence of packet errors.

For BEL ≥ 60 the average quality is almost identical to the original video.

D. Relationship between PER and BEL

The average quality of the experimental video sequence for different values of PER and BEL is shown in Fig. 5.

In assessing the impact of erroneous packets received on quality, it is necessary to analyze not only the likelihood of occurrence of errors, but also their structure and length of their grouping. Additionally, the following conclusions can be drawn:

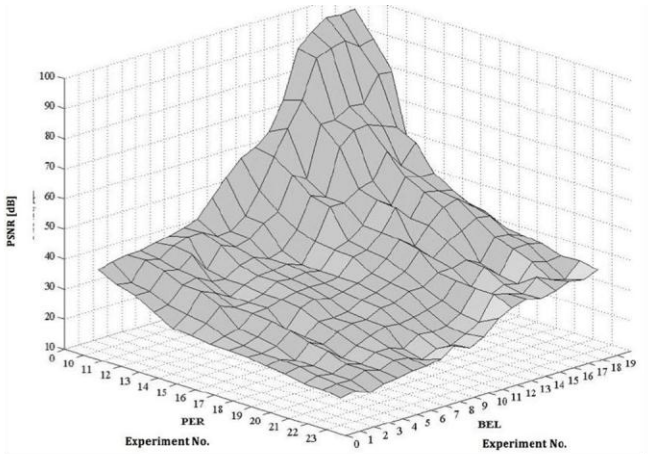


Figure 5. Estimate of video sequence quality indicator for different values of wireless channel PER and BEL.

i) Increasing the length of error groupings leads to an increase in the average quality of the video sequence. This is due to the deterioration of a small section of video, where error groups are concentrated, whereas in the case of single bit errors deterioration in the quality of video may be observed across the whole sequence;

ii) When the length of erroneous packets is BEL ≤ 6 the change in quality is minor and identical to the influence of single packet errors (BEL=1);

iii) When BEL ≥ 60 the average quality is almost identical to the original (PSNR <90 dB). It is logical to assume that the value of BEL in the longer video sequences, with the same average quality may have a higher value;

iv) The highest dynamics of change in PSNR=60 dB is observed in two cases: a) for a fixed PER = $1x10^{-3}$ and the variable values of BEL; and b) at BEL ≥ 80 and the varying values of the PER. In other cases, the dynamics is not essential and minimal in the absence of clustering of errors (BEL = 1)

v) With increasing PER, the effect of BEL on quality decreases due to increase in denseness of single errors.

vi) Analysis of the results of PER and BEL shows that for effective assessment of the impact of transmission errors on resultant quality it is necessary to analyze not only the likelihood of errors, but also their structure and length of their grouping. The most realistic and

accurate method of modeling statistical errors in communication channels is the use of probability data obtained from real networks.

At BER values $\leq 3x10^{-5}$ bit errors do not affect the quality of the received video and are easily eliminated by well-known methods of error correction implemented in WiMAX. When BER $\geq 4x10^{-3}$ packet loss in the network reaches its maximum value and leads to an unacceptable quality of the received video. Ensuring objectively *excellent* quality of video sequence over a channel can be done for probabilities of bit error rate less than $1x10^{-4}$; *good* quality in the range of $1x10^{-4} – 4x10^{-4}$; *satisfactory* quality in the range of $4x10^{-4} – 8x10^{-4}$; poor quality in the range of $8x10^{-4} – 1x10^{-3}$ and *very bad* at BER $\geq 1x10^{-3}$.

The use of H.264/AVC video in wireless access systems with VLC codec of variable length leads to a disruption of the synchronization of decoded video sequences and the occurrence of additional grouping of errors, whose impact on the quality for video decoding is much stronger than that of the bit error, since it leads to loss of large segments of the information. It is shown that the quality of the video affects not only the probability of error, but also the structure and length of errors. Analysis of individual errors showed that at PER $\leq 1x10^{-4}$ packet errors do not affect the quality of the received video and are easily eliminated by well-known methods of error correction deployed in wireless networks. When PER $\leq 1x10^{-3}$, the effect of errors on quality is not noticeable and does not irritate the viewing experience. For values of PER ≥ 0.1 packet loss in the network leads to an unacceptable quality of the received video. Ensuring objectively *excellent* quality of video sequence over a channel can be done for probabilities of bit error rate less than $1x10^{-3}$; *good* quality in the range of $1x10^{-3} – 3x10^{-3}$; *satisfactory* quality in the range of $3x10^{-3} – 1x10^{-2}$; *poor* quality in the range of $1x10^{-2} – 5x10^{-2}$ and *very bad* at BER $\geq 5x10^{-2}$.

To assess the impact on quality of video playback under error grouping conditions of error groups BEL, the use of a regular (deterministic) model is proposed. It is shown that the effect of errors on the average quality is stronger due to local concentration of errors. The average quality of the video sequence at the same time increases with increase in the length of the grouping for a given value of probability of occurrence of packet errors. For groupings of length BEL ≥ 60, average quality is almost identical to that of the source video. With increasing PER, the effect of BEL on quality decreases due to increase in the *denseness* of single errors. Increase in the BEL leads to an increase in the average quality of the video sequence irrespective of the PER value. The highest dynamics of change in PSNR is observed for fixed PER $=1x10^{-3}$ and the variable values of BEL; at BEL ≥ 80 as well as for the changing values of the PER. In other cases, the dynamics is not significant and is minimal in the absence of clustering of errors. To assess the quality of video under packetization of errors under real conditions, it is necessary to investigate the actual distribution of packetization of errors in the communication channel.

VII. Conclusion

We have presented in this paper the results of investigating the effect of single bit error and bit packet errors on the quality of H.264/AVC standard bursty video streams in wideband wireless access systems, using a software-hardware composite system that was developed specifically for this purpose. From analyses of simulation results, we safely conclude as discussed in sections VI A through D for BER, PER, BEL, and relationship between PER and BEL respectively.

References

[1] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, “Overview of the H.264/AVC Video Coding Standard”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 13, No. 7, pp. 560-576, July 2003.

[2] O. I. Sheluhin, Y. A. Ivanov, “Assessment of the quality of streaming video in telecommunication networks using software-hardware methods,” Electro-technical and Information Complexes and Systems, vol. 5, No.4, pp. 48-56, 2009.

[3] Y.A. Ivanov, V.S. Pryanikov, “Imitation Modeling of Wireless Networks using Hardware-Software Complex for the Assessment of Streaming Video Quality,” Chuvash University Digest, vol.1, No.1, pp.35-48, 2010.

[4] NS-2 documentation, available at: http://www.isi.edu/nsnam/ns/ns-documentation.html, accessed 29.06.2010.

[5] G. J. Sullivan, T. Wiegand, “Video Compression-From Concepts to the H.264/AVC Standard”, Proceedings of the IEEE, Vol. 93, No. 1, pp.18-31, Jan. 2005.

[6] ITU P.800: Methods for subjective determination of transmission quality, available at: http://www.itu.int/rec/T-REC-P.800-199608-I/en.

[7] Atayero A.A., “Estimation of the Quality of Digitally Transmitted Analogue Signals over Corporate VSAT Networks”, Ph.D Thesis (unpublished), Moscow, Jan. 2000.

[8] J. Ostermann, et al., “Video coding with H.264/AVC: Tools, Performance, and Complexity”, IEEE Circuits and Systems Magazine, pp. 7 – 28, Q1. 2004.

[9] J. J. Lemmon, “Wireless link statistical bit error model,” NTIA Report. 02-394, U.S. Department of Commerce, June 2002.

Authors Profile

Aderemi A. Atayero graduated from the Moscow Institute of Technology (MIT) with a B.Sc. Degree in Radio Engineering and M.Sc. Degree in Satellite Communication Systems in 1992 and 1994 respectively. He earned a Ph.D in Communications/Signal Processing from Moscow State Technical University of Civil Aviation, Russia in 2000. He is a two-time Head, Department of electrical and Information Engineering, Covenant University, Nigeria. He was the coordinator of the School of Engineering of the same University.

Dr. Atayero is a member of a number of professional associations including: the Institute of Electrical and Electronic Engineers, IEEE, the International Association of Engineers, IAENG, among others. He is a registered engineer with the Council for the Regulation of Engineering in Nigeria, COREN, as well as a professional member of the International Who’s Who Historical Society (IWWHS). He is widely published in International peer-reviewed scientific journals, proceedings, and edited books. He is on the editorial board of a number of highly reputed technical and scientific publications. He is a recipient of the ‘2009/10 Ford Foundation Teaching Innovation Award’. His current research interests are in Radio and Telecommunication Systems and Devices; Signal Processing and Converged Multi-service Networks.

Oleg I. Sheluhin was born in Moscow, Russia in 1952. He obtained an M.Sc. Degree in Radio Engineering1974 from the Moscow Institute of Transport Engineers (MITE). He later enrolled at Lomonosov State University (Moscow) and graduated in 1979 with a Second M.Sc. in Mathematics. He received a PhD at MITE in 1979 in Radio Engineering and earned a D.Sc. Degree in Telecommunication Systems and Devices from Kharkov Aviation Institute in 1990. The title of his PhD thesis was ‘Investigation of interfering factors influence on the structure and activity of noise short-range radar’.

He is currently Head, Department of Information Security, Moscow Technical University of Communication and Informatics, Russia. He was the Head, Radio Engineering and Radio Systems Department of Moscow State Technical University of Service (MSTUS).

Prof. Sheluhin is a member of the International Academy of Sciences of Higher Educational Institutions. He has published over 15 scientific books and textbooks for universities and has more than 250 scientific papers. He is the Chief Editor of the scientific journal Electrical and Informational Complexes and Systems and a member of Editorial Boards of various scientific journals. In 2004 the Russian President awarded him the honorary title 'Honored Scientific Worker of the Russian Federation'.

Yury A. Ivanov was born in Moscow, Russia in 1985. He obtained an M.Sc. degree in Systems, network and devices in telecommunications from Chuvash State University in 2007. He obtained a Ph.D in Telecommunication Networks and Systems in 2011 from Moscow State University of Communication and Informatics. His dissertation topic was "The impact of errors in channels of broadband wireless access systems on the quality of streaming H.264/AVC video". Dr. Ivanov has published over 35 scientific papers and his current research interests include Radio and Telecommunications Systems and Devices: transmission of multimedia data across telecommunication networks, assessment of the quality of video sequences.

Juliet O. Iruemi was born in Kaduna, Nigeria in 1984. She obtained a B.Eng. degree in Information and Communication Technology from Covenant University in 2008. She is currently on her M.Eng. Programme in Information and Communication Technology in Covenant University. Her thesis topic is "Hybrid WLAN Access Point (AP) based on Software Defined Radio (SDR)". Her current research interests include Radio and Telecommunication Systems: Wireless access transmission over broadband network.

Survey on Impact of Software Metrics on Software Quality

Mrinal Singh Rawat[1]
Department of Computer Science
MGM's COET,
Noida, India

Arpita Mittal[2]
Department of Computer Science
IIMT,
Merrut, India

Sanjay Kumar Dubey[3]
Department of Computer Science
Amity University,
Noida, India

***Abstract*—Software metrics provide a quantitative basis for planning and predicting software development processes. Therefore the quality of software can be controlled and improved easily. Quality in fact aids higher productivity, which has brought software metrics to the forefront. This research paper focuses on different views on software quality. Moreover, many metrics and models have been developed; promoted and utilized resulting in remarkable successes. This paper examines the realm of software engineering to see why software metrics are needed and also reviews their contribution to software quality and reliability. Results can be improved further as we acquire additional experience with variety of software metrics. These experiences can yield tremendous benefits and betterment in quality and reliability.**

Keywords- Software metrics; Software quality; Software reliability; Lines of code; Function points; object oriented metrics.

I. INTRODUCTION

Software metrics are valuable entity in the entire software life cycle. They provide measurement for the software development, including software requirement documents, designs, programs and tests. Rapid developments of large scaled software have evolved complexity that makes the quality difficult to control. The successful execution of the control over software quality requires software metrics. The concepts of software metrics are coherent, understandable and well established, and many metrics related to the product quality have been developed and used.

It is essential to introduce definition of software metrics. Software metrics provides measurement of the software product and the process of software production. In this paper, the software product should be seen as an abstract object that begins from an initial statement of requirement to a finished software product, including source and object code and the several forms of documentation exhibited during the various stages of its development.

Good metrics should enable the development of models that are efficient of predicting process or product spectrum. Thus, optimal metrics should be: [1]

- Simple, precisely definable—so that it is clear how the metric can be evaluated;
- Objective, to the greatest extent possible;
- Easily obtainable (i.e., at reasonable cost);
- Valid—the metric should measure what it is intended to measure; and
- Robust—relatively insensitive to (intuitively) insignificant changes in the process or product.

II. OVERVIEW OF SOFTWARE METRICS

A. Classification of Software Metrics

There are three types of software metrics: process metrics, project metrics and product metrics. [3]

1) Process Metrics:

Process metrics highlights the process of software development. It mainly aims at process duration, cost incurred and type of methodology used. Process metrics can be used to augment software development and maintenance. Examples include the efficacy of defect removal during development, the patterning of testing defect arrival, and the response time of the fix process.

2) Project Metrics:

Project metrics are used to monitor project situation and status. Project metrics preclude the problems or potential risks by calibrating the project and help to optimize the software development plan. Project metrics describe the project characteristics and execution. Examples include the number of software developers, the staffing pattern over the life cycle of the software, cost, schedule, and productivity. [4]

3) Product Metrics:

Product metrics describe the attributes of the software product at any phase of its development. Product metrics may measure the size of the program, complexity of the software design, performance, portability, maintainability, and product scale. Product metrics are used to presume and invent the quality of the product. Product metrics are used to measure the medium or the final product.

We can find more efficient ways of improving software project, product and process management.

B. Mathematical Analysis

A metric has a very explicit meaning in mathematical analysis .It is a rule used to determine distance between two points. More formally, a metric is a function 'd' defined on pairs of objects p and q such that d (p, q) expresses the distance between p and q. Such metrics must satisfy certain properties: [11]

d (p,p) = 0 for all p : that is, the distance from point p to itself is zero;

d (p, q) = m (q, p) for all p and q: that is, the distance from p to q is similar to the distance from q to p;

d (p, r) ≤ d (p, q)+d (q, r) for all p, q and r: that is, the distance from p to r is no larger than the distance measured by stopping through an intermediate point.

A prediction system comprise of a mathematical model along with a set of prediction processes for determining unknown parameters and depicting the results. The model should not be complicated for use. Suppose we want to predict the number of pages, P that will print out as a source code program, so that we can bring sufficient paper or calculate the time the program will take for printing. We can use a simple model,

$$P = x/a \quad (1)$$

Where x is a variable, acts as a measure i.e. length of source code program in LOC (line of code), and 'a' is a constant that represents the average number of lines per page. There are number of models to determine effort estimation; from analogy based estimation to parametric models. A generic model can be used to estimate effort predication.

$$E = aS^b \quad (2)$$

Where a and b are constants. E is effort in person-months. S is the size of source code in Line of code.

III. Importance of Software Quality

In recent times the importance of software quality has come to light when random errors on a say a telephone bill, or on a bank statement were randomly attributed to a bug in the "computer code" or using the ignorant adage of "the computer does things" without making an effort to undermine the cause of the problem or even separating it by hardware or software. The problem arises when "computer errors" creep into highly critical aspects of our lives involving situations where a small error can lead to a cataclysmic chain of events. Bearing all this in mind, the importance of enforcing software quality in computer practices has become highly important. Seeing the penetration of computer code into everyday objects like washing machines, automobiles, refrigerators, toys and even things like the mars rover, any system be it a large one or a small system running embedded IC technology, ensuring the highest levels of software quality is paramount.

However, that brings us to the next logical question, how do we assess the quality of something intangible like software quality? This is a highly subjective question whose answer will vary according to the situation. For example, a small word processing error in a student's assignment will not be a huge issue. But a slight code error in a space shuttle's guidance computer might be mission critical and endanger human lives.

Hence in terms of software quality, it is imperative that we understand that it's impossible to have a boilerplate definition or meaning of software quality. The definition will differ according to factors like quality of products and business. Also crucial is the proper setting of goals as well as proactive monitoring of quality factors and goals making sure that that the goals set are resolved and completed in the given timelines and specifics. Views on Software Quality

Software quality, as stated earlier, depends on a number of factors. Also as theorized by David & Garwin, quality is a complex as well as multifaceted concept, which can be viewed according to different points of view as follows

1) User View

The user viewpoint of software quality tends to be a lot more concrete and can be highly subjective depending upon the user. This view evaluates the software product against the user's needs. In certain types of software products like reliability performance modelling and operational products, the user is monitored according to how they use the product.

2) Manufacturing View

This viewpoint looks at the production aspect of the software product. It basically stresses on enforcing building the product without any defects and getting it right the first time rather than subsequently making a defective product and spending valuable project time and more importantly costs ironing out the defects or bugs at a later stage. Being process based, this viewpoint focuses on conformity to the process, which will eventually lead to a better product.

Models such as ISO 9001 as well as the Capability Maturity Model do encompass this viewpoint that stress on following the process as opposed to going by specification. However, that being said, the theory that following the best and high quality manufacturing process will automatically lead to a better product cannot be inferred. The critic's viewpoint is that following an optimized and high quality product manufacturing method can also lead to the standardization of a product making it more of a commodity rather than a standout product.

That being said, there have been a lot of industry example where the philosophy of "doing it right" the first time been profitable. Also both the models CMM as well as the ISO, indirectly do imply by following the principle of "Documenting what you do and doing what you say" helps in improving the product quality.

3) Product View

The product viewpoint looks at the internal features as well as the characteristics of the product. The idea behind this viewpoint is that in case a product is sound in terms of the features and functionality it offers, and then it will also be favourable when viewed from a user viewpoint in terms of software quality. The idea is that controlling the internal product quality indicators will influence positively the external product behaviour (user quality) There are models trying to link both the views of software quality but more work is needed is this area.

4) Value based view

The value-based view becomes important when there are lots of contrasting views, which are held by different departments in an organization. For example, the marketing department generally take a user view and the technical department will generally take a product-based view. Though initially these contrasting viewpoints help to develop a 360-

degree product with the different viewpoints complementing each other, the later stages of the software product development might have issues

The issues arise when there might be a set of change proposed to a certain view that can end up throwing a conflict in the other view. For example, say the marketing department (user view) want changes to the user interface that are not technically feasible (product view).

This is where a value-based view comes into play helping resolve such conflicts so that the software product is not delayed indefinitely. The value-based viewpoint looks the conflict with a cost to benefit angle. It help in resolving such issues by looking at the issue in relation to terms like costing, constraints, resources, time. Using this viewpoint, it's possible to resolve interview conflicts helping to keep the software product on track and within initial cost and timeline estimates.

IV. Case Study on Software Quality

The Boeing 777 project – Boeing with its 777 airplane project was a giant leap forward in the direction of Software quality and is compelling case and point in the importance in reinforcing strong software quality management. With almost 2.5 million lines of code written for the new jetliner's state of the art avionics and other on board software, it was super critical to ensure best software quality practices and implementation. Complications like an extensive network of third party suppliers who would supply crucial components for the 777 made it a large challenge to ensure that deadlines are met without a compromise on software quality as a whole. [15]

Measures taken by Boeing – interestingly, at the beginning of the 777 project, since there was extensive vendor fragmentation, each vendor was using different measures and metrics to keep in track of software quality and measure the status of the work. As a result, this soon snowballed into a situation where due to non-standard practices being followed, it was extremely hard to understand the progress of the project as a whole. Therefore, around the 777 project's midway point, Boeing implemented measures, which called for uniformity in reporting as well as monitoring all variables related to the project status and software quality. A uniform use of metrics like came into effect which made the suppliers report around the simple metrics like test definition, resource utilization, test execution as well as detailed plans for the software coding and design.

As a result, since the reporting was uniform as well as the enforcement of these metrics was universal for the of Boeing's vendors, each vendor was now reporting on a bi weekly basis as which now contained information about completed code, testing as well as design. This not only lowered the effort on Boeing's part in consolidating the fragmented data (as was happening previously) but also allowed Boeing to adjust its own plans in sync with the vendor's estimates and hence keep the project on schedule.

Key Takeaways – Boeing realized early enough of the importance of enforcing a uniform set of metrics. Also vital learnings from Boeing's experience is that done properly, enforcing software quality in a project ensures that program risk points can be identified early which would allow a reasonable time to apply corrective measures without delaying a project indefinitely. Additional key points are the implementation of metrics allowed each project point to be having a check and balance so that the project flows smoothly without any major roadblocks. A good consequence of the metrics implementation was the streamlining and the regularity of communications between Boeing and its vendors, which was touted as being of equal importance to the metrics as well. Clear goals, milestones and constant monitoring of the key metrics around software design coding and testing made sure the 777 project was a success.

V. Comparison of Software Metrics- Strengths and Weaknesses

The software industry does not have standard metric and measurement practices. Most of the software metric has multiple definitions and ambiguous rules for counting. There are also important subject issues that do not have specific metrics, such as quantifying the volume or quality levels of databases, web sites and data warehouses. There is a lack of strong empirical data on software costs, schedules, effort, quality, and other tangible elements, which results in metric problems. [12]

A. Source Code Metrics

"Source lines of code" or SLOC was the first metric developed for quantifying the outcome of a software project. The divergent "lines of code" or LOC has similar meaning and is also widely acceptable. "Lines of code" could be defined either:

- A physical line of code.
- A logical line of code.

Physical lines of code are sets of coded instructions terminated by hitting the enter key of a keyboard. Physical lines of code and logical lines of code are almost identical for some languages, but for some languages there can be considerable differences. Generally, the difference between physical lines of code and logical lines of code is often excluded from the software metrics literature.

Strengths of physical lines of code (LOC) are:

- It is easy to measure.
- There is a scope for automation of counting.
- It is used in a verity of software project estimation tools.

Weaknesses of physical LOC are:

- It may include significant "dead code."
- It may include white spaces and comments.
- This metric is vague for software reuse.
- It does not function for a few "visual" languages.
- Direct conversion to function points is erroneous.
- It is inconsistent for direct conversion to logical statements.

Strengths of the logical LOC are:

- It omits dead code, blanks, and comments.
- Mathematical conversion of logical statements into function point metrics is possible.
- Logical LOC are used in many software project estimation tools.

Weaknesses of logical LOC are:

- It can be difficult to measure.
- These are not comprehensively automated.
- These are ambiguous for a number of "visual" languages.
- This metric is vague for software reuse.
- Direct conversion to the physical LOC metric may be erroneous.

"Measuring software productivity by lines of code is like measuring progress on an airplane by how much it weighs." – Bill Gates.

It is prudent to focus more on building expertise on Function Point Analysis and use it effectively.

B. Function Point Metrics

The function point analysis to measure software application is enumerated from analysis of the requirements and logical design of the application. Function Point count can be applied to Development projects, Enhancement projects, and existing applications as well. [13] There are five key elements of Function Point Analysis, which capture the functionality of the application. These are:

- External Inputs (EIs),
- External Outputs (EOs)
- External Inquiries (EQs)
- Internal Logical Files (ILFs) and External
- Interface Files (EIFs).

First three elements are of Transactional Function Types and last two are of Data Function Types. Function Point Analysis consists of performing the following steps:

- Determine the type of Function Point count
- Determine the application boundary
- Identify and rate transactional function types to calculate their contribution to the Unadjusted Function Point count (UFP)
- Identify and rate the data function types to calculate their contribution to the UFP
- Determine the Value Adjustment Factor (VAF) by using General System Characteristics (GSCs)

Finally, calculate the adjusted Function Point count

When we examine the patterns of strengths and weaknesses of function point metrics, we observe that for economic studies and for studies that include non-coding work such as specifications, function points are clearly superior to lines of code metrics. [12]

Strengths of function point metrics are:

- It stays stable regardless of programming languages used.
- It can compute non-coding activities such as documentation.
- It can measure non-coding defects in requirements and design.
- These are useful for software reuse analysis.
- Function points are used for object-oriented economic studies.
- These are supported by a lot of software cost estimating tools.
- Mathematical conversion of function points into logical code statements is very easy.

Weaknesses of function point metrics are:

- Function Point counting requires good deal of experience.
- Function point counting can be protracted and pricey.
- Function point counting automation is of indefinite accuracy.
- Function point counts are unreliable for those projects that are below 15 function points in size.
- Function point variant have no conversion rules to IFPUG function points.

C. Object-Oriented Metrics

In today's software development environment, Object-oriented analysis and design concepts are well known. Object-Oriented Analysis and Design of software provide many advantages such as reusability, decomposition of problem into easily understandable object and the aiding of future modifications. Object-oriented software development requires a diverse approach from more traditional functional decomposition and dataflow development methods. But the OOAD software development life cycle is not easier than the typical procedural approach. Therefore, it is necessary to provide dependable guidelines that one may follow to help ensure good OO programming practices and write reliable code. Object-Oriented programming metrics is an aspect to be considered. Metrics should be a set of standards against which one can measure the effectiveness of Object-Oriented Analysis techniques in the design of a system. [2]

Strengths of OO metrics are: [12]

- The OO metrics are psychologically attractive within the OO community.

- The OO metrics come out to be able to differentiate simple from complex OO projects.
- Weaknesses of OO metrics are:
- The OO metrics do not support studies outside of the OO paradigm.
- The OO metrics have not yet been applied to testing.
- The OO metrics have not yet been applied to maintenance.
- The OO metrics have no conversion rules to lines of code metrics.
- The OO metrics have no conversion rules to function point metrics.
- The OO metrics lack automation.
- The OO metrics are difficult to enumerate.
- Software project estimation tools do not support the OO metrics.

OO metrics are not linked to all other known software metrics. There are no conversion rules between the OO metrics and any other metrics, so it is complicated to perform alongside comparisons between OO projects and conservative projects using the currently available OO metrics.

VI. Future Scope

Looking at rising demand for the implementation and successful case studies of software quality, it is safe to conclude that in the coming years, software metric's importance will increase multifold as industry leaders like embrace newer and more stringent approaches to monitoring, improving as well as delivering better software quality in products as well as processes. A number of metrics are proposed and exercised for measuring the quality of a system before implementation. Future research directions include improvement in existing metrics based on the nature and magnitude of the problem statement. There is a scope for various tools to support software project development reducing time, effort and cost of the project in consistent manner.

VII. Summary and Conclusion

With the rapid advancement in software industries, software metrics have also developed fast. Software metrics become the basis of the software management and crucial to the accomplishment of software development. It can be anticipated that by using software metrics the overall rate of progress in software productivity and software quality will improve. If relative changes in productivity and quality can be determined and studied over time, then focus can be put upon an organization's strengths and weaknesses. Although people appreciate the significance of software metrics, the metrics field still needs to mature. Each of the key software metrics candidates has broken into many competing alternatives, often following national restrictions. There is no adequate international standard for any of the extensively used software metrics. Absence of firm theoretic background and the assurance of methods, software metrics are still young in comparison of other software theories.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lips Software Metrics SEI Curriculum Module, SEI-CM-12-1.1, December 1988

[2] "Software Quality Metrics for Object Oriented System Environments", June 1995, National Aeronautics and Space Administration, Goddard Space Flight Center, Greenbelt Maryland

[3] Tu Honglei1, Sun Wei1, Zhang Yanan1, "The Research on Software Metrics and Software Complexity Metrics", International Forum on Computer Science-Technology and Applications, 2009.

[4] http://www.pearsonhighered.com/samplechapter/0201729156.pdf.

[5] http://www.wohlin.eu/Articles/ICGSE11.pdf.

[6] Barbara Kitchenham, Shari Lawrence," Quality: The Elusive Target", IEEE Software-Vol. 13, No. 1: JANUARY 1996, pp. 12-21.

[7] Henrike Barkmann, Rudiger Lincke and Welf L owe, "Quantitative Evaluation of Software Quality Metrics in Open-Source Projects".

[8] Dindin Wahyudin, Alexander Schatten, Dietmar Winkler, A Min Tjoa, Stefan Biff,"Defect Prediction using Combined Product and Project Metrics ", March 2008.

[9] Nachiappan Nagappan, Brendan Murphy, and Victor Basili, "The Influence of Organizational Structure On Software Quality: An Empirical Case Study", January 2008.

[10] David I. Heimann,"Implementing Software Metrics at a Telecommunications Company" – A Case Study,2oo4

[11] Rakesh.L , Dr.Manoranjan Kumar Singh ,and Dr.Gunaseelan Devaraj ,"Software Metrics: Some degree of Software Measuremen and Analysis ",(IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 2, 2010

[12] Capers Jones, Chief Scientist Emeritus, "Strengths and Weaknesses of Software Metrics", Version 5, March 22, 2006.

[13] Kurmanadham V.V.G.B. Gollapudi, "Function Points or Lines of Code? – An Insight"

[14] Arti Chhikarai, R.S.Chhillar , "Impact of Aspect Orientation on Object Oriented Software Metrics". Vol. 2 No. 3 Jun-Jul 2011, Indian Journal of Computer Science and Engineering (IJCSE)

[15] Lytz, R., Software Metrics for the Boeing 777: A Case Study, Software Quality Journal, 4, 1-13 (1995)

Authors Profile

[1]Ms. Mrinal Singh Rawat is Assitant Professor in the Department of Computer Science and Engineering in MGM's COET, Noida, UP, INDIA. Her Research activities are based on Software Engineering and Reliability Engineering.She is pursuing her M.Tech in Computer Science and Engineering from Amity University.

[2]Ms. Arpita Mittal is working as Assistant Professor in Department of Computer Science at IIMT Merrut, UP, INDIA. Her Research activities are based on Software Engineering and Software Tesing. She is pursuing her M.Tech in Computer Science and Engineering from Amity University.

[3]Mr. Sanjay Kumar Dubey is working as Assistant Professor in Department of Computer Science and Engineering in Amity University Noida, UP, INDIA. His Research area includes Software Engineering and Usability Engineering.He is pursuing his Ph.D in Computer Science and Engineering from Amity Unversity.

A Cost-Effective Approach to the Design and Implementation of Microcontroller-based Universal Process Control Trainer

[1]Udeze Chidiebele. C, [3] Uzedeh Godwin,
[1,3] R & D Department, Electronics Development Institute (FMST-NASENI), Awka, Nigeria.

[2]Prof. H. C Inyiama,[4] Dr C. C. Okezie,
[2,4] Electronics and Computer Engineering Department, Nnamdi Azikiwe University, Awka, Nigeria.

***Abstract*—This paper presents a novel approach to the design and implementation of a low-cost universal digital process control trainer. The need to equip undergraduates studying Electronic Engineering and other related courses in higher institutions with the fundamental knowledge of digital process control practical was the main objective of the work. Microcontroller-based design and implementation was the approach used where only one AT59C81 with few flip-flops were used for the whole eight processes covered by the trainer thereby justifying its low-cost and versatility.**

Keywords- process control; control algorithm; Algorithmic State machine (ASM) chart; State Transition Table (STT); Fully-expanded STT; Control software.

I. INTRODUCTION

The basic objective in process control is to regulate the value of some quantity which means to maintain that quantity at some desired value regardless of external influences. The desired value is called the reference value or set point [2]. Inyiama H. C and Okezie C.C (2007) stated that a process control is typically a sequential logic system whose control algorithm can be represented in the form of a flow chart called an algorithmic State machine (ASM) chart [3] or in the form of State Transition Diagram (STD) [4]. The ASM chart is a diagrammatic description of the output function and the next-state function of a state machine to implement an algorithm and becomes part of the design documentation when completed. The symbols covered are the STATE BOX, THE DECISION BOX, THE CONDITIONAL OUTPUT BOX and ASM BLOCK.

According to Clare C.R [1973], the ASM chart has three basic elements: the state, the qualifier and the conditional output [3]. The need to equip undergraduates studying Electronic Engineering and other related courses in higher institutions with the fundamental knowledge of digital process control practical was the main objective of this work. The design approach used was the use of Algorithmic State machine (ASM) charts State Transition Diagrams (STD), State Transition Tables (STT), Fully Expanded STT, and Link Path Addressable ROM Structure. The significance of the work extends to the fact that it presented a standard approach to the design and implementation of the process control systems that can be applied to any process controls in the industries and it also bridges the gap created in the lives of these students due to their lack of exposure to electronics practical.

The following processes were covered by the trainer:

- Temperature Level Control systems
- Automatic Liquid Dispenser System
- Automatic Water Pump control system
- Traffic Light Control System
- Upper Tank Water level Control systems
- Lower Tank Water level Control systems
- Automatic Gate control system
- Automatic Street Light Control system

II. METHODOLOGY & DESIGN ANALYSIS

The design of the digital process control system can be achieved through various methods which include: Gate-Oriented Design which involves equation-to-gate conversions, map simplifications, output function synthesis, and next-state function synthesis, flip flops, state assignment and hazards. When such discrete logic gates (such as AND, NAND, OR, NOR, EX-OR, INVERTERS etc) and memory flip flops are used, what is termed a RANDOM LOGIC system results. Since such a system involves mostly small scale integration components, the component count is usually high for a fairly complex circuit. This implies several interconnections and many potential sources of error. Modifications and maintenance are also difficult to achieve when either re-designing or fault-finding becomes necessary. A random logic system is therefore very inflexible.

Fortunately however, logic systems can be implemented in forms more structured than random logic. Such systems employ structured logic devices such as Multiplexers (or Data Selectors) (5), and Read-Only-Memories (ROMs). A multiplexer-based controller requires as many multiplexers as there are columns in the bit-pattern sequence to be generated and each of these multiplexers must have at least as many data input lines as there are rows in the bit-pattern sequence. Thus, if an 8-bit pattern is to be generated by means of multiplexers, a total of 64 data input lines would be involved. This is in addition to other control inputs and outputs necessary in such a system. The rapid proliferation of data input lines (and hence

potential sources of error) in multiplexer based complex sequential logic systems is its main disadvantage in such applications. Multiplexers being structured logic devices do, however have a considerable advantage over random logic in that errors in the design can be corrected simply by altering the logic levels applied to their data inputs. For compact, reliable and easily maintainable implementation of a complex sequential logic ROMs have an edge over multiplexers and are usually preferred.

The microcontroller-based implementation was chosen for this work due to its numerous advantage over the others which include its simplicity, and the fact that a simple control software can be developed which can be used without any modification in any control system, no matter how complex or how simple, and even when the numbers of input qualifiers, state code, size and number of output lines differ from one control system to the other, provided one input port is sufficient for Address inputs and one output port for the control pattern output. The system is made up of the hardware subsystem and software subsystem. The hardware subsystem is comprised of the input interface, the control systems and the output interface.

The Structure of the Microcontroller-based Digital Process Control is shown in Fig. 2 below. It comprise of the microcontroller with its input and output ports, the input interface subsystem connected to the input port of the microcontroller, the output interface subsystem connected to output port of the microcontroller. The input interface system comprise of all the sensors that will be used for a particular process. For example in the temperature level control system, LM 35 is the sensor, which senses the temperature of the water container. Also the output interface comprises all the transducers such as light emitting diodes, LCDs and so on. Note that buttons and keypads are part of the input subsystems since they are used for selecting the particular process to be controlled.

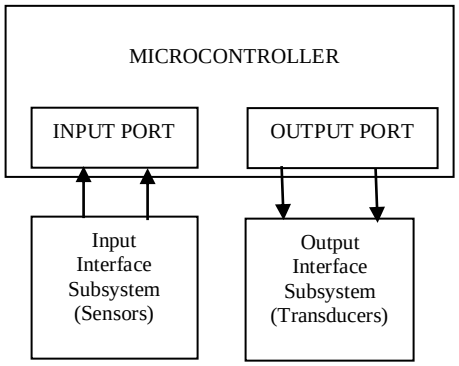


Figure 4. The Structure of the Microcontroller-based Digital Process Control.

III. THE GENERATION OF CONTROL BIT- PATTERN SEQUENCE

The procedure that is followed for the generation of the control bit-patterns include: the drawing of ASM chart for each of the processes, transforming the ASM charts into a state transition table, and expanding the state transition table fully so that the location address and the content address for the process are generated which is then burned into the flash drive of the microcontroller.

This procedure was used for all the processes to generate the control bit-patterns in Table 3 below but the design process that led to that was illustrated in this paper by one of the processes which is the temperature level control system of water or other liquids . The ASM chart of the temperature control system is shown in Fig. 2 below. These ASM chart is then transformed into a state transition table of Table1 and then to fully expanded state transition table of Table2.

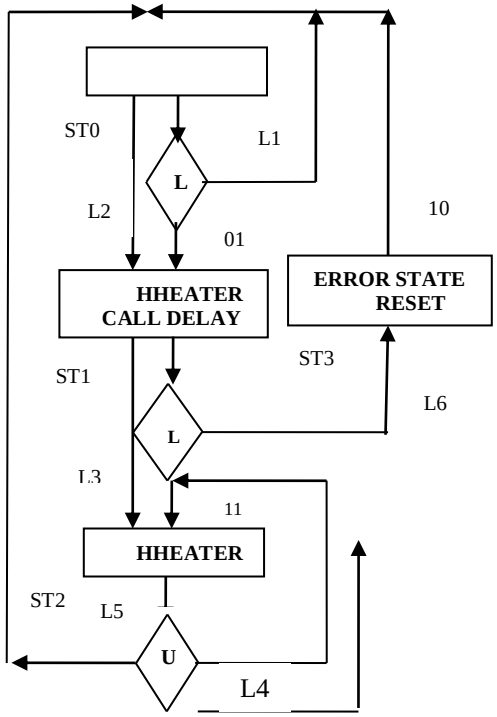


Figure 5. (i)The ASM chart of the temperature control system

2 (iii) STATE ASSIGNMENT 2 (ii) STATE MAP

STATE NAME	STATE CODE
ST0	0 0
ST1	0 1
ST2	1 1
ST3	1 0

B \ A	0	1
0	ST0	ST1
1	ST3	ST2

The labels or names inside some of the state boxes are the state outputs. The ASM charts of the Fig. 2 above have only one state output namely: HHEATER. The logic level of the output signal is high or active when the control system is in that state. The bit pattern at the top right end of the state box is its state code. The letters B, A above the state code signify that two flip flops B and A are used to represents the various states of the machine.

The state codes are the logical levels at the Q outputs of these two flip flops respectively. Each rectangular box in the

ASM chart is a state box. The word (e.g ST0) enclosed in a circle at the bottom left hand corner of a state box is the state name. Here ST0 stands for state 0, and similar interpretations apply to the other states, hence ST1 Means state 1, and so on. Each decision box has one entry path and two exit paths. The exact value of an input qualifier determines which exit path is followed out of a decision box. In the ASM chart of Fig. 2 Uth and Lth are the input qualifiers.

With the help of a K_map in which the state names are inserted serially in an adjacent cells (Fig. 2ii) and which is called a state map, the state codes are chosen such that only one bit changes level as one moves from one state of the control system to another. This is clearly brought out in the state assignment of Fig. 2iii. This method of state assignment facilitates complexity reduction when hardwired logic is the preferred technique of control system implementation and helps to prevent race hazards [3].

Every ASM chart has an equivalent tabular representation known as a State Transition Table (STT) [3]. An ASM chart can be fully described in terms of the link paths comprising it. A State machine such as is represented by an ASM chart attempts to change state (i.e. transits from the present state to the next) when a clock pulse occurs. A link path is a path followed from the present state back to itself or to another state, when the clock pulse arrives. When there is an input qualifier between the present state and another, the logic level of the qualifier determines the next state the machine goes to at the clock pulse. If there is no qualifier between the present state and the next, the machine must unconditionally transit from its present state to the adjacent state in the forward direction, when a clock pulse arrives.

The ASM chart of Fig. 2i has 7 link paths labeled L1 to L7. L1 is the link path from state 0 back to itself when the input qualifier is 0. Also L2 is the link path from state 0 to state 1 when the input qualifier Lth is 1. Similarly, L3 represents the transition from state ST1 to state 2 when the qualifier Lth is 0. L4 is the transition from state2 back to itself when the input qualifier Uth is 0 and L5 is the transition from state 2 to state 0 when Lth input qualifier Uth is 1. L6 is the transition from state 1 to state 3 which is the error state where the system stays until a key is pressed to return it to state 0 through link path L7.

In the STT of Table 1 a number of dashes appear under the columns headed by the input qualifiers Lth and Uth. A dash (_) implies that the input qualifier above that column is not relevant for the transition being made in the link path shown on the STT row where the dash appeared. That input qualifier may become relevant in some other link path transition and then its value would become 0 or 1, rather than a dash. A dash in the STT also means that the input qualifier that appears as column heading for that dash may be at logic 0 without affecting the control process at that material time.

A State Transition Table (STT) is said to fully-expanded when all the dashes on each row are given all possible combination of logic values, leading to new rows in the State Transition Table, one for each combination of the logic values for the dashes on that row. In this respect, a STT data row with one row when that (dashed) qualifier is given the logic value 0 and the other when that qualifier is given the logic level 1. Similarly a STT data row with 2 dashes expands into 4 STT rows. Assume the (dashed) qualifiers are represented by q1, q2. Then the first STT row in the expansion will give q1, q2 = 0, 0, the second would have q1, q2 = 0, 1, the third row would have q1, q2 = 1, 0, the fourth row would have q1, q2 = 1,1. Three dashes on an STT data row would in like manner lead to 8 rows in the fully expanded STT of Table 2 results.

IV. MICROCONTROLLER-BASED IMPLEMENTATION

The high capacity of ROMs relative to the number of unique bit-patterns in a fully expanded STT suggests the use of a single ROM to store the fully expanded bit-pattern of all the processes. The link-path addressable word structure is based on storing the next state and the output for each link path in the ASM charts. The next-state portion of the ROM word is called the LINK PART, while the output portion of the ROM word is called the INSTRUCTION PART. Each address is a function of the present state and qualifier inputs and called a link-path address. In general any process described by an ASM chart can be implemented with this structure, called a link-path addressable ROM.

Since a microcontroller-based implementation is used, the same link path addressable ROM patterns will be programmed into its ROM or the Erasable and Programmable ROM (i.e. EPROM) or Flash Drive (now available in newer microcontrollers). The ADDRESS inputs to the Address Decoder part of the ROM device would now be input via an input port of the microcontroller which is then used to locate the corresponding NEXT STATE and STATE OUTPUT.

When a clock pulse occurs, the NEXT STATE pattern becomes the present state pattern. This joins the input qualifiers to comprise the next Address input to be used by the microcontroller. A power-up one-shot applied to the D flip flops that feedback the Next State as the present state when a clock pulse occurs, initializes the system to state 0 at the start up[6]. Thereafter, the control system behaves as defined by the ASM chart, with the help of the control software running in the microcontroller. The user may use a simple switch to stop the control software run.

A. The Control Software

A very important universal concept that is a natural outcome of the use of a fully expanded STT in a link path addressable ROM structure, as part of a microcontroller-based implementation is that a simple software can be developed which can be used without any modification in any control system, no matter how complex or how simple, and even when the numbers of the input qualifiers, state code, size and number of the output lines differs from one control system to the other, provided one input port is sufficient for address inputs and one output port for control pattern output.

The flowchart for this universally applicable control software is shown in the Fig. 3 below; with a pseudo code that explains what it does above it.

The control cycle time is the time delay required for the system to settle and become ready for the next round of input-output. Its default value is zero seconds.

TABLE 1. THE STT FOR TEMPERATURE LEVEL CONTROL.

LINK PATHS	INPUT QUALIFIERS Lth	Uth	PRESENT STATE NAME	PRESENT STATE CODE B	A	NEXT STATE NAME	NEXT STATE CODE B'	A'	STATE OUTPUT HHEATER
L1	0	_	ST0	0	0	ST0	0	0	0
L2	1	_	ST0	0	0	ST1	0	1	0
L3	0	_	ST1	0	1	ST2	1	1	1
L4	_	0	ST2	1	1	ST2	1	1	1
L5	_	1	ST2	1	1	ST0	0	0	1
L6	1	_	ST1	0	1	ST3	1	0	1
L7	_	_	ST3	1	0	ST0	0	0	0

TABLE 2. FULLY EXPANDED STT TABLE FOR THE SYSTEM

LINK PATH	LOCATION ADDRESS (hex)	ADDRESS PATTERN Lth Uth B A	CONTENT PATTERN B' A' H	LOCATION CONTENT (hex)
L1	0 0 0 4	0 0 0 0 0 1 0 0	0 0 0 0 0 0	0 0 0 0
L2	0 8 0 C	1 0 0 0 1 1 0 0	0 1 0 0 1 0	0 2 0 0
L3	0 1 0 5	0 0 0 1 0 1 0 1	1 1 1 1 1 1	0 7 0 7
L4	0 3 0 B	0 0 1 1 1 0 1 1	1 1 1 1 1 1	0 7 0 7
L5	0 7 0 F	0 1 1 1 1 1 1 1	0 0 1 0 0 1	0 1 0 1
L6	0 9 0 D	1 0 0 1 1 1 0 1	1 0 1 1 0 1	0 5 0 5
L7	0 2 0 6 0 E 0 A	0 0 1 0 0 1 1 0 1 1 1 0 1 0 1 0	0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0

B. Pseudo Code

```
BEGIN
    Repeat
    Input Next Control Pattern Address from Input Port;
        Retrieve Next Control Pattern from Link-path
        Address Rom Structure;
        Output Next Control Pattern Address from Output Port;
        Delay for Control Cycle Time.
        Until HSTOP = 1
END
```

C. Flow Chart

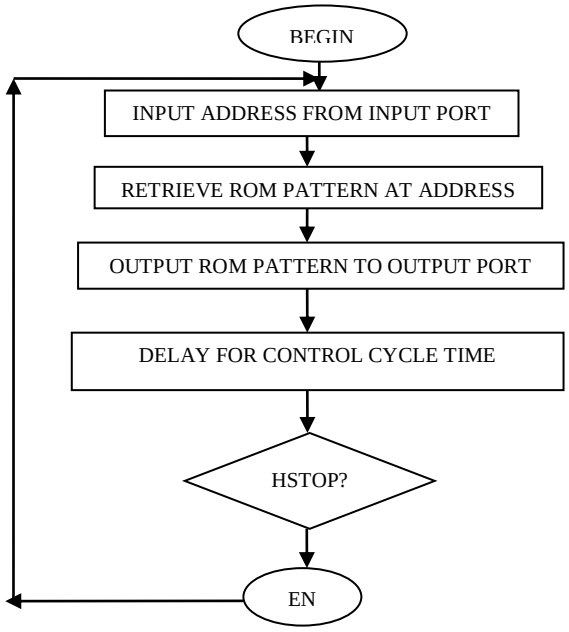


Figure 6. The Flowchart rep. of the control software

V. Results and Discussion

A software-based universal digital process control system was achieved with the program design method presented. Following the procedure for developing a fully-expanded state transition table, which begins with representing the process in an ASM chart, developing a state transition table for the process from the ASM chart and then translating the state transition table to a fully-expanded state transition table, the fully-expanded state transition table for the other processes was also developed and all of them were shown in Table 3. The location address is the input address of the processes, which is used to locate where in the ROM memory the location content is stored. The location address and the location content will be burned into the ROM permanently.

The difference in terms of programming effort, between a software-based controller intended for just one control system and that designed for several control systems (each with a different number of input qualifiers) is minimal. Fig. 4 illustrates a typical microcontroller-based digital process control system. With this trainer the student can perform several experiments on each of the processes as covered in the laboratory manual developed for the trainer which enables him gain the needed practical knowledge on digital process control. The keyboard (Fig. 4) is used for selecting the particular process desired to be controlled. It also facilitates the input of variables or control parameters which make software-based controllers very flexible, while the LCD display enables the microcontroller to transmit responses to user commands in addition to providing the current status of the controlled device where necessary.

TABLE 3. LOCATION ADDRESS & CONTENT FOR THE CONTROLLER ROM FROM THE FULLY-EXPANDED STT OF THE 8 PROCESSES.

Location address (hex)	Location content (hex)	Location address (hex)	Location content (hex)
000	00	400	00
004	00	404	00
008	02	408	02
00C	00	40C	00
001	07	401	07
005	07	405	07
003	07	403	07
00B	07	40B	07
007	01	407	01
00F	01	40F	01
009	05	409	05
00D	05	40D	05
002	00	402	00
006	00	406	00
00E	00	40E	00
00A	00	40A	00
104	02	500	02
108	02	504	02
10C	02	508	02
110	06	50C	02
114	06	510	06
118	06	514	06
11C	06	518	06
101	04	51C	06
105	04	501	04
110	04	505	04
115	04	510	04
109	0C	515	04
10D	0C	509	0C
119	0C	50D	0C
203	0D	603	0D
20B	0D	60B	0D
213	0D	613	0D
21B	0D	61B	0D
207	09	607	09
20F	09	60F	09
217	09	617	09
21F	09	61F	09
202	08	602	08
206	08	606	08
212	08	612	08
216	08	616	08
20A	00	60A	00
20E	00	60E	00
304	00	700	00
308	02	704	00
30C	00	708	02
301	07	70C	00
305	07	701	07
303	07	705	07
30B	07	703	07
307	01	70B	07
30F	01	707	01
309	05	70F	01
30D	05	709	05

TABLE 4. COST IMPLICATIONS OF THE PROJECT

	ITEM	QTY	Unit Cost (Naira)	Cost (Naira)
1	Sensors	7	400	2800
2	ADC0804	1	200	200
3	Microcontroller, AT89C51	1	350	350
4	4x20 Liquid crystal display	1	2500	2500
8	DC relays (12V & 6V)	2	70	140
12	10k ohm resistors	7	5	35
13	Transistor, BC337	5	5	25
14	Diodes, 1N4001	6	5	30
15	10kohm Variable resistor	2	20	40
16	10uf & 33pf capacitor	2	20	40
19	7805	1	40	40
21	Copper clad board	1	100	100
22	40 pin IC socket	1	40	40
23	Soldering Lead	1	150	150
25	Casing materials	-	1000	1000
	TOTAL			7490NGN

Table 4 shows the cost implication of the project. The total cost implication of the project is 7490 NGN (Nigerian Naira). When compared to the cost implication of the same project using other methods such gate-oriented designed method, multiplexer and ROM-based design method as shown in Table 5, it is cheaper thereby justifying the cost-effectiveness of the approach used in this work.

TABLE 5. COST COMPARISON WITH OTHER METHODLOGIES

Methodology	Cost (NGN)
Gate-oriented design method	12,350
Multiplexer-based design method	10,200
ROM-based design method	9,400
Microcontroller-based design method	7,490

VI. CONCLUSION

The use of a single microcontroller to control several processes, based on storing the fully expanded State Transition Tables of those processes in its ROM or flash drive makes possible the realization of a low-cost universal processes control trainer.

REFERENCES

[1] Inyiama H. C, Okezie C.C, Designing microcontroller-based universal process control systems, Volume 2 Number 2 (Electroscope), November 2007. Department of Electrical and Electronics Engineering, Nnamdi Azikiwe University Awka. Pp.11-26.

[2] Curtis D Johnson, Process Control Instrumentation Technology, 8th Edition, 2006, Prentice-Hall Inc. pp. 1-10.

[3] Clare C.R, Designing logic systems using state machines (U.K, MacGraw-Hill, 1973) pp 1-108.

[4] Roger L. Tokheim, Digital Electronics, Principles and Applications Fifth Edition,1999, Glencoe McGraw-Hill. pp269-276.

[5] Williams, G.E., "Digital Technology, Principles and Practice". Science. Research Associats Inc., (1974), pp. 96-105,202-245.

[6] Walter G. Jung IC timer cookbook (4300West 82nd St. Indianapolis 48258 USA, Howard W. Sams & Co. Inc. 1977) pp1-36.

[7] Michael J. Pont, Embedded C, 2002, Pearson Education Limited, pp 17-34.

[8] Bertram J. E: 'The concept of state in the the analysis of discrete time control system,' 1962 Joint Autom. Control Conf. New York University (June 27-29), Paper No. 11-1

[9] Inyiama H.C, Unpublished lecture Notes on Real-Time computing and control, Department of Electronic and Computer Engineering, Nnamdi Azikiwe University Awka 2008.

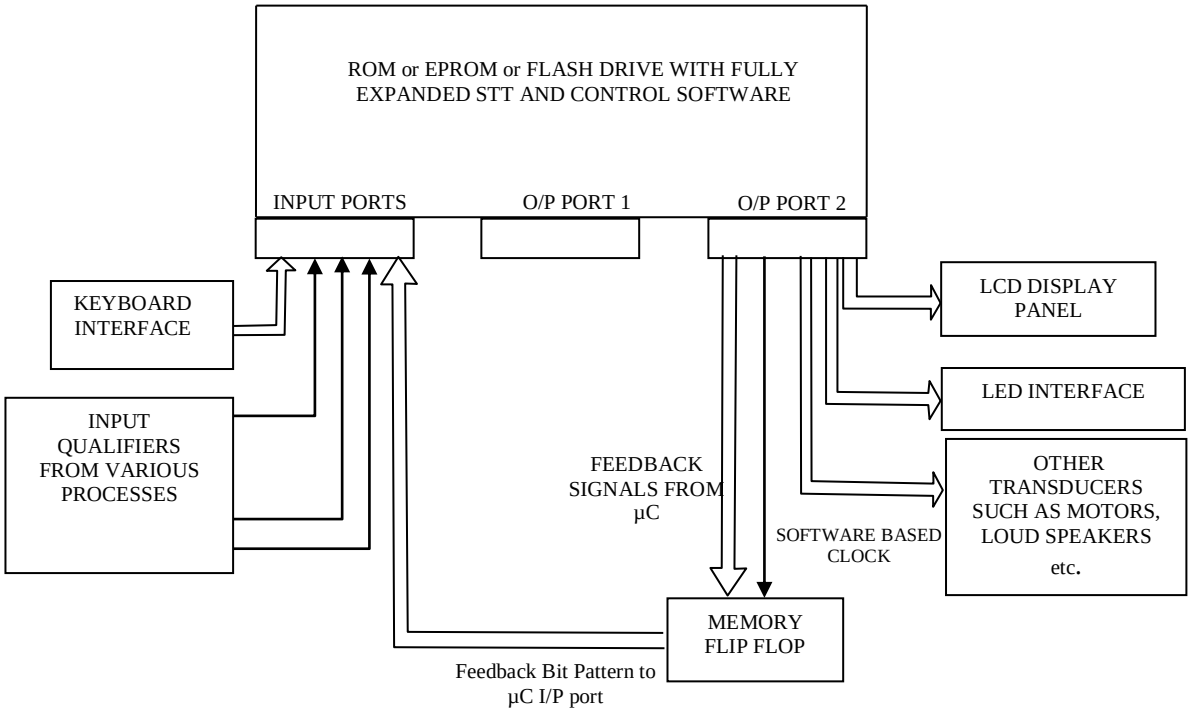


Figure 4. Microcontroller-based digital process control system

Self-regulating Message Throughput in Enterprise Messaging Servers – A Feedback Control Solution

Ravi Kumar G
HP, Research Scholar, JNTUH
Bangalore, India

C.Muthusamy
Yahoo
Bangalore, India

A.Vinaya Babu
JNTUH
Hyderabad, India

***Abstract*—Enterprise Messaging is a very popular message exchange concept in asynchronous distributed computing environments. The Enterprise Messaging Servers are heavily used in building business critical Enterprise applications such as Internet based Order processing systems, pricing distribution of B2B, geographically dispersed enterprise applications. It is always desirable that Messaging Servers exhibit high performance to meet the Service Level Agreements (SLAs). There are investigations in this area of managing the performance of the distributed computing systems in different ways such as the IT administrators configuring and tuning the Messaging Servers parameters, implement complex conditional programming to handle the workload dynamics. But in practice it is extremely difficult to handle such dynamics of changing workloads in order to meet the performance requirements. Additionally it is challenging to cater to the future resource requirements based on the future workloads. Though there have been attempts to self-regulate the performance of Enterprise Messaging Servers, there is a limited investigation done in exploring feedback control systems theory in managing the Messaging Servers performance. We propose an adaptive control based solution to not only manage the performance of the servers to meet SLAs but also to pro-actively self-regulate the performance such that the Messaging Servers are capable to meet the current and future workloads. We implemented and evaluated our solution and observed that the control theory based solution will improve the performance of Enterprise Messaging Servers significantly.**

Keywords-Feedback control; Message Oriented Middleware; Enterprise Messaging; Java Messaging Service; JMS Providers; Adaptive Control

I. INTRODUCTION

Enterprise Messaging also known as Message Oriented Middleware [1] is a popular asynchronous message exchange mechanism in heterogeneous distributed applications. It provides the applications in a distributed environment to send and receive messages, but still being loosely coupled. Loose coupling between enterprise class applications and legacy systems such as business workflow applications, databases, and data warehouses plays a significant role in Enterprise Application Integration (EAI) [2]. The Message based integration provides automation and simplifies the time consuming integration tasks like create, deploy and manage integration solutions. There are many such applications such as Business to Business (B2B) solutions, messaging across various entities within a business enterprise that are geographically separate where asynchronous messaging becomes a major building block [2]. Asynchronous Messaging is a backbone for many of the Event driven architectures due to the obvious advantages of asynchronous systems where the message client need not maintain the connection and session with the message receiver; no confirmation is required from the receiving application [2]. As we discussed Enterprise Messaging is an important element in the business critical environments, it is always important for the Enterprise Messaging Servers to exhibit high performance and availability. Typically there would be Service Level Agreements (SLAs) [3] defined between the business service providers and the consumers. Performance is an obvious Service Level Objective in such SLAs. Any violation of performance SLOs [4] will affect the business and reputation of the business enterprise. In this paper we want to discuss the performance regulation of Java based Enterprise Messaging Servers. There are different implementations of such Enterprise Messaging Servers. The Java based Messaging Service is called as Java Message Service [5], included in the specification for Java based Enterprise Environments called as JEE (called as J2EE previously) [6]. There are different vendors who implemented the JMS Specification and Java based Enterprise Messaging Servers are referred as JMS Providers. Hence forth in the document the Enterprise Messaging Servers are referred as JMS Providers [7].

Typically the performance of JMS Providers is measured by its message throughput, though CPU and Memory usage [8] are common metrics to measure the performance of any computing server. The message throughput will depend upon various factors such as the number of subscribers, message size, number of publishers, and number of JMS message brokers [9]. By tuning these different parameters the desired performance can be achieved on the JMS Providers. One of the mechanisms to improve the JMS provider's performance is by following some best practices such as setting non-durable messages, set the message time to live parameter appropriately, close message publishers and subscribers when they complete their jobs [10]. But these kinds of practices will not be able to address different kinds of JMS environments and applications limiting the performance improvement. The other mechanism is to provide the facilities to the administrators to configure [11] and fix the various parameters values which influence the JMS Provider performance. Due to the dynamics of messages flow and workload on the JMS Provider, it will be difficult for the administrator to tune these values accurately and

periodically. When there are sudden huge loads administrator may decide to provision additional resources which may be left unutilized [12] later when there are relatively lesser loads. This will eventually lead to either not addressing the performance needs or ineffective utilization of computing server assets. Another way to manage the performance is to include conditional programming within JMS Provider implementation to change the values of the parameters at runtime based on the workload and deviation from the expected performance. This method is though useful it is very complex because during design the workload dynamics have to be accurately estimated. During implementation the conditional programming is implemented which is very complex [13] as the conditions implemented may not be sufficient to meet the run time dynamics, any spikes in the workload. To summarize, though there are different mechanisms to adjust the JMS Providers parameters to regulate and improve the message throughput either it involves manual intervention, involve complex conditional programming implementation.

In order to handle such situations, we propose an adaptive control [14] based solution that regulates the message throughput according to the pre-defined reference using a feedback controller. Also, predict the future load on the JMS Provider, modify the control parameters accordingly. There may be a case where the future load predicted may demand additional servers; our controller will actuate a signal to provision additional resources.

In this paper we will first present a background on the choosing feedback control systems in Distributed computing systems, then a brief overview on the Java Messaging Service, followed by discussing the adaptive controller algorithm that we have implemented to regulate the performance of the JMS Provider.

II. BACKGROUND

We have discussed the importance of JMS Providers and importance of their performance in building business critical applications and services in enterprise level or at internet level. There are attempts to predict the performance of JMS Providers [15], or study and compare the performance of different vendors of JMS Providers [16]. Additionally there are some best practices [10] identified to improve the JMS Providers performance. Manual configuration is one of the most common approaches followed to tune the JMS Provider performance. There is a very limited investigation done in automatic regulation the JMS Providers performance. The message throughput of the JMS Providers depends upon the number of subscribers, publishers and number of brokers. Allowing more number of subscribers on a given JMS Provider may decline message throughput or having a less number of subscribers may leave the JMS Provider less utilized. The control system based solutions provide mechanism to automatically tune the maximum number of subscribers in an optimal operating range. In this paper we propose a control systems based solution for managing the performance by tuning the maximum number of subscribers that influence the message throughput.

Control systems theory has been in investigation to address these kinds of problems related to regulating the performance, in computing [17]. But the majority of focus is on Web Servers [18][19][20], Application Server performance regulation [21], in computer networks such as congestion control [22]. There are recent investigations to explore the applicability of control systems in other areas of Java based Cloud and Enterprise Environments [23], database driver cache hit ratio improvement [24], spring based software applications [25]. But in our study we have observed there is no investigation carried out in applying feedback control system theory in improving the performance of JMS based servers. We investigated to apply control systems theory in Enterprise Messaging server performance improvement and evaluated how the feedback controllers improve the JMS Providers performance significantly.

III. THEORITICAL AND PRACTICAL CONSIDERATIONS

The message throughput of the JMS Providers depends upon various factors such as publishers, subscribers, JMS brokers. The performance varies based on whether the messages are persisted are not. The JMS Providers exhibit higher performance when the messages are not persisted. In this paper the persistence factor is not considered and the performance is evaluated with proposed solution. The subscribers are identified as a significant independent variable influencing the message throughput. The number of publishers and the brokers will have a definite impact on the message throughput to cater huge publisher and subscriber volumes. When the subscribers and message throughput are depicted in mathematical model, the accuracy of the JMS Provider model depends upon the constant values chosen for that model. These constants can be determined by using different data set values of message throughput for varying subscribers. These values may not hold good for different workload conditions on the JMS Provider, but the best suitable constants can be chosen before running the experiments.

IV. ENTERPRISE MESSAGING PRIMITIVES

In this section we discuss a brief overview of the Enterprise messaging [26] also known as Message Oriented Middleware (MoM). The key concept behind MoM is the asynchronous messaging. It means that the sender is not required to wait for the message to be received or handled by the receiver. The Fig 1 shows high level diagram of MoM. The sender can forward the message and continue the processing. The asynchronous messages are treated as autonomic units. The message contains all the data and state needed by the business logic that processes it.

A. Enterprise Messaging Architectures

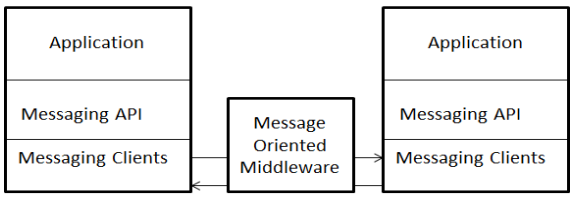


Figure 1. Message Oriented Middleware

1) Centralized Architectures :

In Centralized Architectures there will be a Message Server also called as a message router or message broker that is responsible of sending messages such that the message sender is decoupled from the message receiver. This enables the clients to be added and removed without impacting the system. In this model, the hub-and-spoke topology is used as shown in Fig 2 below:

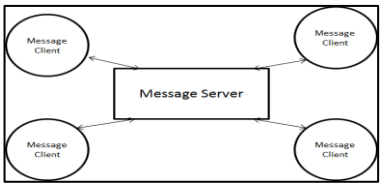


Figure 2. Centralized MoM Architecture

2) Decentralized Architectures

In Decentralized architectures, the IP multicast is used at the network level. There is no centralized server and some of the JMS functionality like persistence, transactions, security is embedded in the client application. The messaging routing is delegated to the network layer by using the IP multicast protocol as shown in Fig 3.

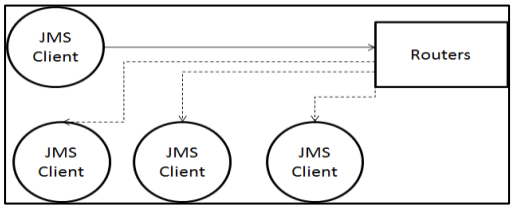


Figure 3. Decentralized MoM Architecture

B. Java Messaging Service [26]

The Java Message Service (JMS) is a specification that proposes programming API for Enterprise Messaging. JMS supports messaging as a first-class java distributed computing paradigm. There are many vendors who implemented the JMS specification and such implementations are called JMS Providers, which are nothing but Enterprise Message Servers based on Java.

1) JMS Messaging Models

The JMS provides two types of messaging models, point-to-point and publish-subscribe models. The intermediate element that enables the communication between the message producer and message consumer in JMS is called a broker. There are two types of JMS brokers as explained below.

a) Point-to-Point

The Fig 4 below shows point-to-point model in which the producer can send a message to only one consumer. In JMS Providers such JMS Brokers called Queues.

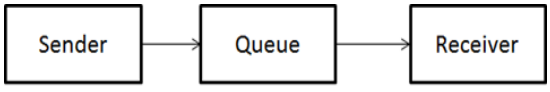


Figure 4. Point-to-Point Model

b) Publish-Subscribe

The Fig 5 below shows publish-subscribe model in which the producer can send a message to many consumers. In JMS Providers such JMS Brokers called Topics.

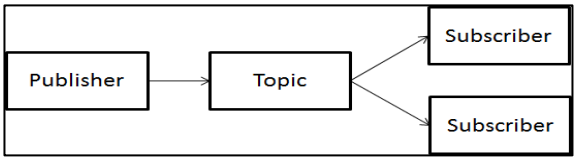


Figure 5. Publish-Subscribe Model

V. Adaptive Control

The message throughput (T) depends upon various factors such as the number of subscribers and publishers to the different brokers of the JMS Provider, The messages size, number of brokers running. In this paper we have considered how the number of subscribers of the JMS brokers affects the message throughput (T). Though there are other parameters that influence the JMS Provider message throughput, the maximum number of subscribers allowed on the server will affect significantly. The number of publishers are considered to be constant as 1 in our implementation. The Fig 6 is a Single Input Single Output (SISO) Adaptive control system [27] that shows how the message throughput is regulated using the controller and the Predictor.

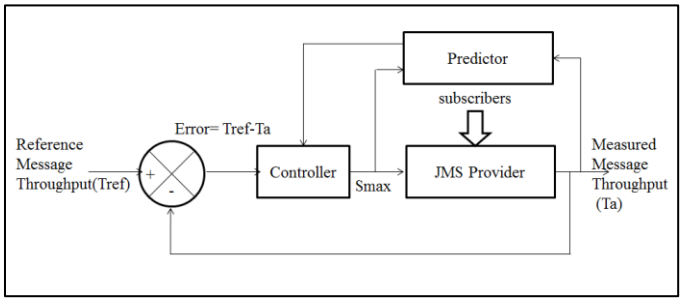


Figure 6. Adaptive Control of JMS Provider

We explain how the message throughput depends upon the number of subscribers of the JMS Provider.

The following equation (1) represents the Message Throughput and its relation with the number of subscribers.

$$T = bS \quad (1)$$

Where

T = Message Throughput of JMS Provider measured as number of messages per unit time

S = Total number of subscribers on the JMS Providers

b = proportional coefficient for the Subscribers

There is a feedback control loop that is implemented which is used to calculate the error signal of the actual Message throughput (Ta) and the Reference Message Throughput (Tref). The error signal is represented by the equation (2)

$$E = T_{ref} - T_a \quad (2)$$

The controller takes the error signal as one input and the other input signal to the controller is the predicted values of number of subscribers and message throughput. The predicted subscribers will help in estimating the possible future subscriber's volume. The predicted subscriber's value is used to predict the message throughput and the latter one is important to determine the future resource requirements. The resources may either be more JMS brokers or additional virtual machines [28] that can be scaled to cater the future load requirements on the JMS Providers. There are threshold values defined for the message throughput based on which the actuator signals are triggered either to add new virtual machines or brokers. The following sections explain the different parts of the solution in detail.

A. Modeling JMS Provider

The JMS Provider, whose message throughput needs to be controlled, has to be mathematically modeled first in order to apply the feedback control techniques. There are many ways to model the compute systems such as difference equations [29], ARMA models [30] that is based on Least Squares Parameter Estimation [31]. In our solution we have used the ARMA model to represent the JMS Provider. The Fig 7 below shows ARMA based modeling of the JMS Provider.

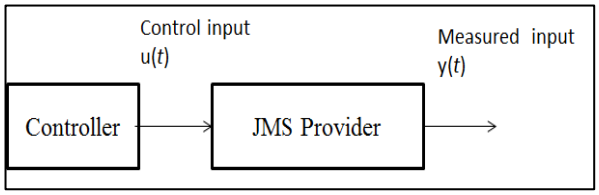


Figure 7. JMS Provider Model for Feedback control

1) Parameter Estimation

In the JMS Provider the message throughput is defined as a function

- The number of subscribers that the JMS Provider is supporting.
- Though the number of publishers and number of message brokers also influence the JMS Provider performance we considered the number of subscribers as the factor in our paper.

According to ARMA, in a Single Input Single Output model, for a given sample data set, the next sample of the output can be predicted using the current and previous inputs. The same is explained in the equation (3) below:

$$y(t+1) = ay(t) + bu(t) \quad (3)$$

Where

$y(t)$ = The current output
$u(t)$ = The current input
a = The model parameter to be estimated
b = the model parameter to be estimated
$y(t+1)$ = the output in the next step

The same ARMA model if is applied to model the JMS Provider, it is represented by the equation (4):

$$T(t+1) = aT(t) + bS_{max}(t) \quad (4)$$

Where

$T(t)$ = The current output of message Throughput
$S_{max}(t)$ = the current input of maximum number of Subscribers
a = the model parameter to be estimated
b = the model parameter to be estimated
$T(t+1)$ = the output in the next step

The ARMA model is used to estimate the model parameters 'a' and 'b'. The details of the experiments and the estimated values are explained in the section VI. "Implementation and Analysis". Based on our experiments the parameter 'a' is determined as 0.91 and 'b' is 0.12.

2) Input Operating Range

It is important to determine the operating range of the maximum number of Subscribers (Smax). The training data set is used again to determine the range of Smax that provides the desired Tref.

In order to achieve the desired value of the Tref, the maximum number of subscribers will have to be adjusted. This value of Smax again will change during runtime due to the stochastic nature of the load and the controller is useful to automatically adjust the Smax to meet the Tref. The details are explained in the Section VI.A "Implementation and Analysis – Modeling JMS Provider"

B. Adaptive Controller

The adaptive controller is designed and implemented to self-regulate the message throughput of the JMS Provider for a pre-defined threshold of message throughput.

We implemented the adaptive control algorithm such that any changes in the JMS Provider load can be well managed such that the desired message throughput (Tref) is achieved at any given point of time. The adaptive control has two different parts.

- Feedback Controller: The feedback controller is reactive in nature and tunes the controller gain based on the current measured message throughput, but cannot handle the future load on the JMS Provider. This runs a "sub-control loop" and at the end of each such loop the controller parameter is tuned such that the message throughput is in an allowed range of Tref
- Predictor: In order to handle the future dynamics of the loads on the JMS Provider, a predictor is used that predicts the Smax and Tref. Based on these predicted values the P-Controller Gain is tuned if predicted desired message throughput is lesser than a pre-defined error. We defined a "parent control loop" that runs periodically. In each parent-control loop the Smax and Tref are predicted. After each parent-control loop, the predicted value of message throughput is compared with the Tref. If the predicted value is less than Tref within a pre-defined deviation then controller tunes the Smax allowed, by adjusting the Controller gain (Kp) such that subsequent loads on the JMS Provider meet the Tref. We used the basic P-Controller [32] to tune

the value of Smax. If this deviation is greater than a pre-defined threshold then it demands additional resources, then the actuator triggers request to create a new Virtual Machine.

Now we explain the two different parts of the proposed Adaptive control solution, Feedback Controller and the Predictor.

1) Feedback P- Controller

The JMS Provider during its operation will have varying workloads that may affect its performance. In order to maintain and regulate the performance in terms of Tref, the maximum number of subscribers that can be allowed on the JMS Provider will have to be tuned. We implemented a P-Controller [32] to adjust the Smax during runtime such that JMS Provider exhibits desired performance. The lower number of subscribers will have the possibility of high Tref, but having a too low value of Smax keeps the JMS Provider under-utilized. In the section VI.A "Implementation and Analysis – Modeling JMS Provider" we have discussed optimal operating region of Smax based on our experiments. In order to keep the desired Tref, the P-Controller will tune the Smax in the operating region. But there may be cases where the actual measure Throughput (Ta) is much lower than Tref. In such scenarios, the control law will trigger a request to provision additional compute resource such as more compute power (e.g., Virtual Machine). The Fig 8 below shows the P-Controller to tune the Smax.

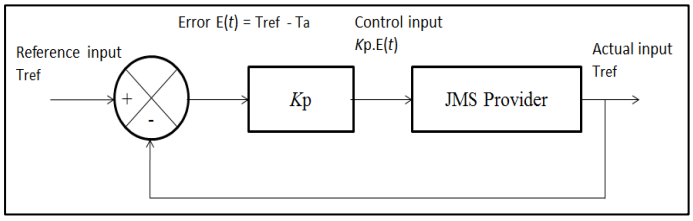


Figure 8. Feedback Control of JMS Provider

The output of the controller is represented by the equation (5) below

$$u(t) = K_p.E(t) \qquad (5)$$

Where

$u(t)$ = The controller output

K_p = Proportional Gain

The P-Controller Gain is represented in the equation (6) in z-Transform

$$K_p = E(z)/U(z) \qquad (6)$$

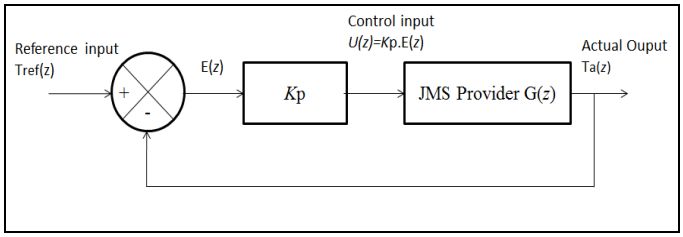


Figure 9. Feedback Control of JMS Provider in z-Transform

The Fig 9 shows the z-Transform [33] of the JMS Provider adaptive loop control.

The equation (5) represents the controller output is. The controller output, which is the new Smax becomes the control input to the JMS Provider. The reactive feedback control runs for every "sub-control loop".

The JMS Provider is represented by G (z) is a first order system as shown in the equation (7) below

$$G(z) = b/(z - a) \qquad (7)$$

2) Predictor

The Predictor is a component in our proposed solution that predicts the maximum number of subscribers for the future periods. The Time-Series Triple Exponential Smoothing [34] is used to predict the Smax that represents the possible future maximum number of subscribers that could be allowed on the JMS Provider based on the past history.

The smoothing technique has the ability to forecast up to 'm' periods ahead. It means that the maximum number of subscribers that can be supported after 'm' periods from the current time can be predicted and hence the corresponding Tref.

The Reference Message Throughput is predicted using the predicted Smax and the previous value of the reference message throughput. The equation (8) below shows how the Tref is predicted

$$T_{ref}(t + 1) = aT_{ref}(t) + bS_{max}(t + 1) \qquad (8)$$

In the Fig 6, we can notice that the Predictor accepts the measured throughput (Ta), current Smax and outputs the predicted Tref. (TrefPred) thus helps in tuning the Kp for the future period.

C. Controller Algorithm

In this section we explain the controller algorithm

The following are the pre-conditions and Initialization operations before the controller is executed

- The JMS Provider model parameters '*a*' and '*b*' are estimated
- The "parent-loop control" and "sub-control loop" is initialized
  - Sampling time of sub-control loop = '*m*'
  - Sampling time of parent-control loop = '*c*' times of '*n*'
- Determine the P-Controller Gain 'K_p'
- Initialize subscribers at the beginning = S_i
- new VM triggering actuating signal message throughput threshold = 'NT_h'
- Error Threshold range to tune the K_p = $E_{T,min}$, $E_{T,max}$
- Parent-control loop execute threshold for message throughput = P_T
  - During running the sub-control loop if the message throughput , when $T_a <= P_T$ then the parent-control loop is triggered

ALGORITHM

i. Start the JMS Provider
ii. Start loading the JMS Provider with initial number of subscribers as S_i
iii. for every '*m*' units of time run the sub-control loop as shown below
  a. measure the actual message throughput (T_a)
  b. If T_a is observed to be less than T_{ref} for more than 4 times, then trigger the parent-control loop (step iv.)
  c. Get T_a , compute the Error '*e*'
  i. If '*e*' is between $E_{r,min}$ and $E_{r,max}$ where $T_a < T_{ref}$, then adjust the P-Controller Gain 'K_p' to meet T_{ref}
  d. Repeat the steps from a. to d. for '*m*' times
iv. For every '*c* x *m*' units of time run the parent-control loop (i,e for every '*c*' sub-control loops)
  a. Define the prediction period '*p*' determines the number of parent control loops from the current parent control loop)
  b. Compute or predict S_{max} for '*p*' periods in advance $S_{max, p}$
  c. Compute or predict T_{ref} for '*p*' periods in advance $T_{ref, p}$
  d. Feed the predicted values to feedback controller
v. The controller will compare the $T_{ref, p}$ and the current message throughput T_a.
  a. If the $T_{ref, p}$ is more by NT_h than T_a, then trigger the actuator to provision new Virtual Machine
  b. If the $T_{ref, p}$ is less by $E_{r,max}$ than T_a, then tune the P-Controller Gain 'K_p'

VI. IMPLEMENATATION AND ANALYSIS

The adaptive control discussed is implemented in Java using an experiment data collected on Apache JMS Provider ActiveMQ [35] running on Ubuntu Linux 10.04 , i5 Intel 2 GHz CPU, 4 GB RAM, 1 TB Hard disk. A sample custom JMS application is run to generate the experiment data. A single JMS topic and a single publisher are used. The subscribers are increased which read different messages from the JMS Topic that are published. The data is collected on the explained experimental setup.

Then the proposed solution is run offline on the experimental data to examine the improvement in the message throughput, without running the proposed controller on the ActiveMQ server online.

The following are the different steps performed for implementing and evaluating the performance of the proposed solution.

- The model parameters (as in Equation (7)) are estimated with two different data sets. The parameters with least error are identified and used for the controller
- Operating range of maximum possible number of subscribers is determined for best possible message throughput, which is between 60 to 90 subscribers
- Based *on* the operating range, *the* P-controller Gain (K_p) is calculated as 2.67 and the Reference Message Throughput (T_{ref}) is determined as 220.
- The Feedback Controller and Predictor are implemented based on the values of P-controller Gain (K_p) and Reference Message Throughput (T_{ref}). The improvement in the message throughput using the proposed controller is evaluated in comparison with the actual message throughput.

We explain the implementation details of modeling the JMS Provider, the controller and discuss the results below.

B. Modeling JMS Provider

The model parameters 'a' and 'b' of the Equation (7) are estimated using the ARMA model where the actual message throughput is measured by linearly increasing the number of subscribers, and predicting the Message throughput. The error percentage is computed between the measured throughput and the predicted throughput. The experiments are run with two different data sets. The Table I shows the estimated model parameters for both the data sets with their error. We observe that the values a = 0.91 and b = 0.12 proved to have a lesser percentage of prediction error.

TABLE I. MODEL PARAMETER ESTIMATION

Data Sets	Model Parameter Estimation		
	a	***b***	***Percentage of Error***
Data Set 1	1	0.28	9.12
Data Set 2	0.91	0.12	8.33

Now using these constants the JMS Provider model in z-Transform is represented as the equation (9) below, using the model parameters estimated.

$$G(z) = 0.12/(z - 0.91) \quad (9)$$

The Fig 10 and Fig 11 show the parameter estimation with actual message throughput (Ta) and the predicted throughput (Tpred). The message throughput in these figures is number of messages per second. In Fig 10 the number of messages is plotted against the increasing number of subscribers. There is a saturation of message throughput after a certain number of subscribers.

C. Adaptive Control

The Fig 12 below shows the performance evaluation of the message throughput without Controller and with adaptive controller proposed in this paper. We observe that the message throughput using proposed Controller is better by about 25 % which is a significant improvement in message throughput over the throughput without controller. We can notice that there are spikes where there is a sudden increase of the number of subscribers. The actual message throughput has reduced suddenly in such cases, but using a P-Controller tuning along with the predictor, provided the adaptive control and has regulated the throughput to be in the operating range between 200 and 250.

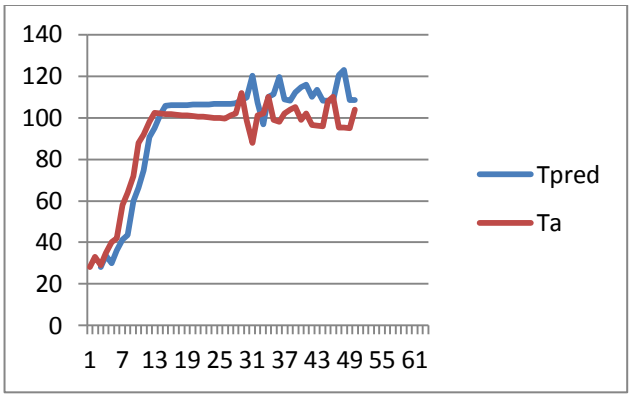


Figure 10. Model Parameter Estimation – Data set 1

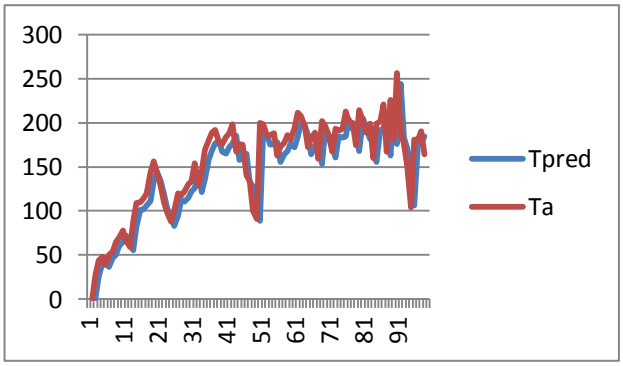


Figure 11. Model Parameter Estimation – Data Set 2

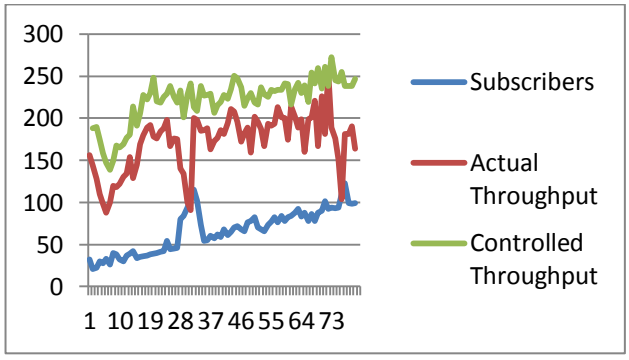


Figure 12. Performance Evaluation of Message Throughput using Adaptive Control

1) Feedback Controller

From the experimental data the value of P-Controller Gain Kp is determined as 2.67. This is computed by adjusting the value of Kp between 2 and 4 and the average Kp is computed. The reference message throughput is computed from the operating range average as 220. By tuning Kp the output of the controller is adjusted which is nothing but the tuning of Smax to obtain the desired reference message throughput. But our implementation has shown that the value of Smax is typically around 59 with a maximum value of 120. The optimal operating range of Smax is $59 \leq Smax \leq 90$. The Table II shows the operating range limits of Kp and Smax.

2) Predictor

We implemented the Triple Exponential Smoothing predictor using openforecast Java API [36]. The Table III shows the different values chosen for Predictor.

TABLE II. OPERATING RANGES

P-Controller Gain (K_p) Range		
K_p Range	S_{max} Range	T_{ref}
2.67, 3, 2, 2.4	59-90	220

TABLE III. PREDICTOR PARAMETER VALUES

Predictor Values		
Triple Exponential Smoothing Coefficients	$E_{r,max}$	*Forecast period (p)*
0.2, 0.6, 0.6	70	1

The Fig 13 below shows the predicted values of the Smax (Smaxp) and the Tref. (Trefpred). These values are predicted using Triple Exponential Smoothing with coefficients shown in the Table III. In our experiment the parent control loop is run once the Trefpred starts decreasing less than 150, which is less than the Tref by 70. From the Fig 13, the predicted Tref (TrefPred) is less than 150, the P-Controller gain Kp is tuned to a value of 4 such that message throughput is regulated without fluctuations. The Er,max is set to as 70 (220-150). The predictor adjusts the Kp once the Tref,pred is less by Er,max (70) than original Tref. In our experimental data we didn't simulate the condition of the Ta exceeding the threshold to trigger addition Virtual Machine requests.

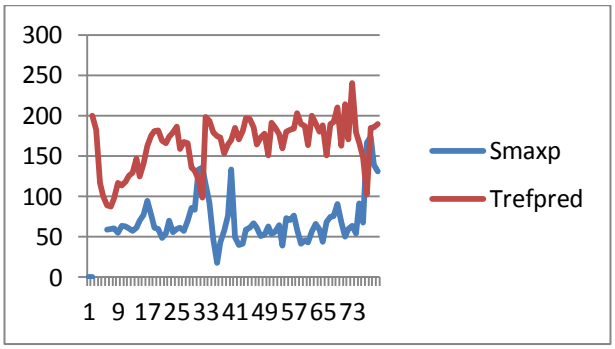


Figure 13. Prediction of Smax and Tref

VII. CONCLUSION

We observed using the P-Controller will have a distinct improvement in the message throughput of the Enterprise messaging servers. Our experiments are currently limited to using the P-Controller only which helps in reducing the rise time [37], but in order to obtain reduce the overshoot and settling time using the PI-Controller [17] is more helpful. Additionally, the parameter estimation is done on experimental data and only two data set samples are used. Our results are based on a simulation like environment as the P-Controller is not directly verified online on the JMS Provider. Our experiments are rather run on the data collected from the JMS Provider by running a sample application with one publisher and one JMS topic. We infer that applicability of adaptive control systems will have significant improvement on the performance of the Enterprise messaging servers in distributed computing systems.

VIII. FUTURE WORK

There is a scope of improvement of the solution explained in this paper. We intend to extend the experiments to adjust the model parameters during runtime such that model represents the behavior of the system to be controlled in a real time. Also, we want to examine the SASO [17] properties of our control system to determine the controller stability and accuracy. We also want to verify the solution on the ActiveMQ server with varying publishers and topics, not limiting to the subscribers only.

We suggest exploring a hybrid approach where techniques like fuzzy control [38] can be used in conjunction with the classic PI controller which can show better performance. The applicability of fuzzy control enables creating a knowledge base of rules and can be evaluated against using Triple Exponential Smoothing for predicting future message throughput. These rules can be helpful when the Enterprise Messaging servers are used in massively large distributed computing systems. We are also studying the different aspects of Data Mining which can be used to build novel prediction algorithms there by the adaptive control system is more robust.

REFERENCES

[1] Message Oriented Middleware (MoM): "http://en.wikipedia.org/wiki/Message-oriented_middleware"

[2] Matjaz B.Juric, S.Jeelani Basha, Rick Leander, Ramesh Nagappan, "Professional J2EE EAI", Shroff Publishers 2005

[3] Service Level Agreement , "http://en.wikipedia.org/wiki/Service-level_agreement"

[4] Service Level Objective, "http://en.wikipedia.org/wiki/Service_level_objectives"

[5] JMS Specification: "http://www.oracle.com/technetwork/java/javaee/tech/index.html"

[6] JEE Specification: "http://www.oracle.com/technetwork/java/javaee/tech/index.html"

[7] JMS Providers: "http://en.wikipedia.org/wiki/Java_Message_Service"

[8] A. Robertsson, B. Wittenmark, M. Kihl, and M. Andersson, "Design and evaluation of load control in web server systems", IEEE American Control Conference, 2004

[9] JMS Performance Benchmarks : "http://www.codeproject.com/KB/showcase/PerformanceBenchmarks.aspx"

[10] http://www.precisejava.com/javaperf/j2ee/JMS.htm#JMS111

[11] Bruce Snyder, Dejan Bosanac and Rob Davies, "ActiveMQ In Action", Dreamtech Press, 2011

[12] Pradeep Padala, Xiaoyun Zhu, Mustafa Uysal et al. Adaptive Control of Virtualized Resources in Utility Environments. In the proceedings of the EuroSys 2007

[13] Evgeny Dantsin, Thomas Eiter, Georg Gottlob, Andrei Voronkov, "Complexity and expressive power of logic programming", ACM Computing Surveys 2003

[14] Karl J.Astrom and Bjorn Wittenmark, "Adaptive Control", Pearson Education, 2009

[15] Yan Liu, Ian Gorton, "Performance Prediction of J2EE Applications using Messaging Protocols", Proceedings of 2005 Symposium on Component-based Software Engineering

[16] Michael Menth, Robert Henjes, Christian Zepfel, and Sebastian Gehrsitz, "Throughput Performance of Popular JMS Servers", ACM SIGMETRICS '06

[17] Joseph L. Hellerstein, Yixin Diao, Sujay Parekh, and Dawn Tilbury, "Feedback Control of Computing Systems", John Wiley 2004

[18] Ying Lu, Avneesh Saxena and Tarek E Abdelzaher, "Differentiated Caching Services; A Control-Theoretical Approach", IEEE International Conference on Distributed Sysytems, 2001

[19] Keqiang Wu, David J. Lilja, Haowei Bai, "The Applicability of Adaptive Control Theory to QoS Design: Limitations and Solutions", IEEE Parallel and Distributed Processing Symposium, 2005

[20] Ying Lu, Tarek Abdelzaher and Gang Tao, "Direct Adaptive Control of A Web Cache System", Proceedings of the American Control Conference, Denver, Colorado, 2003

[21] Giovanna Ferrari, Santosh Shrivastava,Paul Ezhilchelvan, "An Approach to Adaptive Performance Tuning of Application Servers", IEEE International Workshop on QoS in Application Servers, 2004

[22] Seungwan Ryu, Chulhyoe Cho,"PI-PD-controller for robust and adaptive queue management for supporting TCP congestion control", Simulation Symposium, 132 - 139 April 2004

[23] Ravi Kumar Gullapalli, Dr.Chelliah Muthusamy and Dr.A.Vinaya Babu , "Control Systems application in Java based Enterprise and Cloud Environments – A Survey" , IJACSA, Volume 2, No 8, August 2011

[24] Ravi Kumar Gullapalli, Dr.Chelliah Muthusamy, Dr.A.Vinaya Babu and Raj N. Marndi, "A FEEDBACK CONTROL SOLUTION IN IMPROVING DATABASE DRIVER CACHING", IJEST, Vol 3, No 7, July 2011

[25] Dr. Wolfgang Winter , "Applying control theory concepts in software applications", http://www.theserverside.com/feature/Applying-controltheory-concepts-in-software-applications

[26] Richard Monson-Haefel and David A.Chappell, "Java Message Service", O'Reilly 2001

[27] Single Input Single Output : "http://en.wikipedia.org/wiki/Single-input_single-output_system"

[28] Virtual Machines : "http://en.wikipedia.org/wiki/Virtual_machine"

[29] Erwin Kreyzig, "Advanced Engineering Mathematics", John Wiley and Sons

[30] ARMA: http://en.wikipedia.org/wiki/Autoregressive_moving_average_model

[31] MICHAEL L.JOHNSON and LINDSAY M.FAONT Parameter Estimation by Least Squares Error :" http://mljohnson.pharm.virginia.edu/pdfs/174.pdf"

[32] P-Controller, "http://en.wikipedia.org/wiki/Proportional_control"

[33] Z-Transform: Saed Vaseghi, "http://dea.brunel.ac.uk/cmsp/Home_Saeed_Vaseghi/Chapter04-Z-Transform.pdf

[34] Triple Exponential Smoothing: "http://itl.nist.gov/div898/handbook/pmc/section4/pmc435.htm"

[35] Apache ActiveMQ, "http://activemq.apache.org/"

[36] Openforecast API, "http://openforecast.sourceforge.net/docs/"

[37] Jinghua Zhong, "PID Controller : A Short Tutorial", Purdue University, 2006

[38] Jan Jantzen, "Design of Fuzzy Controllers", Tech report, 1988, Technical University of Denmark

AUTHORS PROFILE

Ravi Kumar G is working as a Technical Expert in Hewlett-Packard., Bangalore, India. He obtained his M.Tech in Computer Science from Birla Institute of Technology, Mesra,India. He is currently pursuing Ph.D from JNTU Hyderabad,AP, India.

Dr.Chelliah Muthusamy is Academic Relations Head at Yahoo, Bangalore,. He obtained his Ph.D from Georgia Tech and M.Sc(Engg) in Computer Science from Indian Institute of Science(IISc), Bangalore India

Dr.A.Vinaya Babu is a Professor of Computer Science working as Principal, JNTUH College of Engineering, JNTU Hyderabad, AP, India. He obtained his Ph.D and M.Tech in Computer Science from JNTU, Hydreabad.

Improved Face Recognition with Multilevel BTC using Kekre's LUV Color Space

H.B. Kekre,
Senior Professor
Computer Engineering Department
MPSTME, SVKM's NMIMS
Mumbai, India

Dr. Sudeep Thepade
Associate Professor
Computer Engineering Department
MPSTME,SVKM's NMIMS
Mumbai,India

Sanchit Khandelwal, Karan Dhamejani, Adnan Azmi,
B.tech Students
Computer Engineering Department
MPSTME, SVKM's NMIMS
Mumbai, India

***Abstract*—The theme of the work presented in the paper is Multilevel Block Truncation Coding based Face Recognition using the Kekre's LUV (K'LUV) color space. In [1], Multilevel Block Truncation Coding was applied on the RGB color space up to four levels for face recognition. The experimental results showed that Block Truncation Coding Level 4 (BTC-level 4) was better as compared to other BTC levels of RGB color space. Results displaying a similar pattern are realized when the K'LUV color is used. It is further observed that K'LUV color space gives improved results on all four levels.**

Keywords- Face recognition; BTC; RGB; K'LUV; Multilevel BTC; FAR; GAR.

I. INTRODUCTION

The term face recognition refers to identifying and verifying a face image. It is basically the process of classifying a face as 'known' or 'unknown', based on training set. While humans can easily identify faces, it is a challenging task for computer systems. The computer systems store the faces in such way that the important contents of the face image they store, can be used efficiently for recognizing the face.

There are many biometric systems such as finger prints, voice, iris, face and retina. Among these face recognition turns out to be the most effective system since it requires very less human interaction [21, 22]. Researchers from the field of biometrics, image processing, computer vision, pattern recognition system and neural network give a lot of importance to face recognition. It is the fastest growing biometric technology [18]. Some of the applications of face recognition include physical, security and computer access controls, law enforcement [12, 13], criminal list verification, surveillance at various places [15], forensic, authentication at airports[17], etc.

Many algorithms are used to make effective face recognition systems. Some of the algorithms include Principle Component Analysis (PCA) [2, 3, 4, 5], Linear Discriminant Analysis (LDA) [6, 7, 8], Independent Component Analysis (ICA) [9, 10, 11], Block Truncation Coding (BTC) [1, 15, 19, 22] etc.

The paper presents an approach to enhance the performance of BTC based face recognition using K'LUV color space.

Applying the technique described in [1], using K'LUV color, it is observed that K'LUV out performs RGB color space at each level of Multilevel BTC.

II. BLOCK TRUNCATION CODING

Block truncation coding (BTC) [1, 12, 13, 14] was developed in the early years of digital imaging more than 29 years. It was first developed in 1979 for greyscale image coding [14]. It is comparatively a simple image coding technique. BTC has played a vital role in the history of digital image coding in such a way that many advanced coding techniques have been developed, based on BTC.

III. MULTILEVEL BLOCK TRUNCATION CODING [1, 13, 20]

The feature vector in this algorithm is calculated by using Block Truncation Coding [12, 13 and 14]. In [1], BTC has been implemented up to four levels on RGB colour space for face recognition. The feature vector size at BTC-Level 1, BTC-Level 2, BTC-Level 3 and BTC-Level 4 was 6, 12, 24 and 48 respectively. In the same way BTC on K'LUV colour space is implemented up to four levels for face recognition.

IV. KEKRE'S LUV COLOR SPACE

It was obvious to extend BTC to multi-spectral images such as color images. Most color images are recorded in RGB space, which is perhaps the most well-known color space.

K'LUV color space [12] is a special case of Kekre transform. Where L gives luminance and U and V gives chromaticity values of color image. Positive value of U indicates prominence of red component in color image and negative value of V indicates prominence of green component.

Equation (1) gives the RGB to LUV conversion matrix which indicates the corresponding L, U and V components for an image from the R, G and B components.

$$\begin{bmatrix} L \\ U \\ V \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ -2 & 1 & 1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

The reverse conversion, that is from LUV color space to RGB color space is given in Equation (2).

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1 & -2 & 0 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} L/3 \\ U/6 \\ V/2 \end{bmatrix} \quad (2)$$

V. Proposed Method

For color space, for each BTC level; the feature vector for the query image and database set is by using Multilevel BTC.

In each level of BTC, the feature vector of the query image is compared with the feature vector of each image in the training set. The comparison (Similarity measurement) is done by Mean Square Error (MSE) given by equation 3.

$$\mathrm{MSE} = \frac{1}{\mathrm{MN}} \sum_{y=1}^{M} \sum_{x=1}^{N} [\mathrm{I}(x,y) - \mathrm{I}''''''(x,y)]^2 \quad (3)$$

Where,

I & I' are two feature vectors of size M*N which are being compared.

To assess the performance of the different BTC levels based face recognition techniques, False Acceptance Rate (FAR) and Genuine Acceptance Rate (GAR) are used.

VI. Implementation

A. Platform

The effectuation of the Multilevel BTC is done in MATLAB 2010. It was carried out on a computer using an Intel Core i5-2410M CPU (2.4 GHz).

B. Database

The experiments were performed on two face databases.

1) Face Database [16]

This database is created by Dr Libor consisting of 1000 images (each with 180 pixels by 200 pixels), corresponding to 100 persons in 10 poses each, including both males and females. All the images are captured against a dark or bright homogeneous background, little variation of illumination, different facial expressions and details. The subjects sit at fixed distance from the camera and are asked to speak, whilst a sequence of images is taken. The speech is used to introduce facial expression variation. The images were taken in a single session. The ten poses of Face database are shown in Figure 1.

2) Our Own Database [1, 20]

This database consists of 1600 face images of 160 people (92 males and 68 females).For each person 10 images are taken. The images in the database are captured under numerous illumination settings. The images are taken with a homogenous background with the subjects having different expressions. The images are of variable sizes, unlike the Face database. The ten poses of Our Own Database are shown in Figure 2.

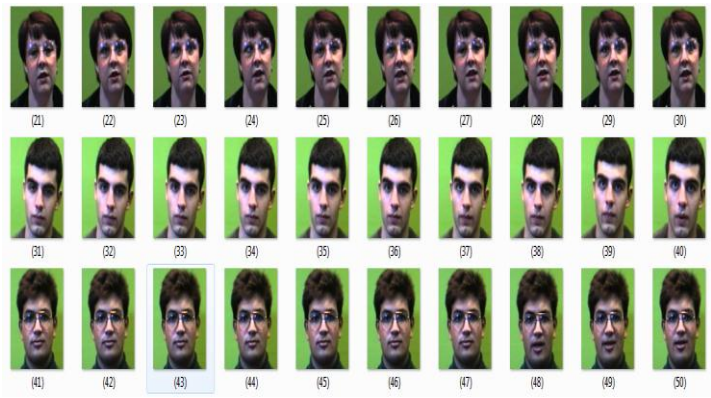


Figure 1. Sample images from Face database

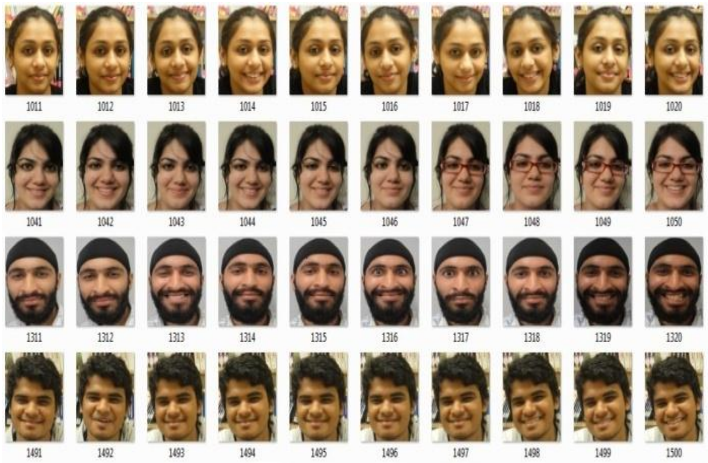


Figure 2. Sample images from Our Own Database

VII. Results And Discussions

False Acceptance Rate (FAR) and Genuine Acceptance Rate (GAR) are standard performance evaluation parameters of face recognition system.

The False acceptance rate (FAR) is the measure of the likelihood that the biometric security system will incorrectly accept an access attempt by an unauthorized user. A system's FAR typically is stated as the ratio of the number of false acceptances divided by the number of identification attempts.

FAR = (False Claims Accepted/Total Claims) X 100 (4)

The Genuine Acceptance Rate (GAR) is evaluated by subtracting the FAR values from 100.

GAR=100-FAR (in percentage) (5)

In all 10000 queries (10 images for each of 1000 persons) are fired on face database and 16000 queries (10 images for each of 1600 persons) are fired on our own database. For each query, FAR and GAR values are calculated for respective BTC level based face recognition technique. At the end the average FAR and GAR of all queries in respective face databases are considered for performance ranking of BTC levels based face recognition techniques.

FAR and GAR are calculated for both RGB color space and K'LUV color space.

A. Face Database

To analyze the performance of proposed algorithm, 10000 queries are tested on the database. The feature vectors of each image for all four BTC levels in RGB color space and K'LUV color space were calculated and then compared with the database. The FAR and GAR values are calculated by employing equations 4 and 5.

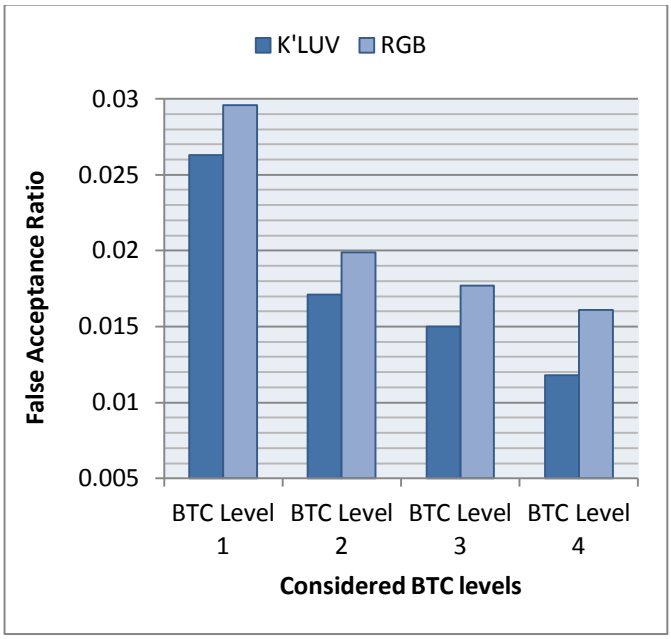


Figure 3. FAR values at different BTC levels of K'LUV and RGB color spaces for Face Database

Figure 3 gives the FAR values of the different BTC levels based face recognition techniques tested on face database for both RGB and K'LUV color spaces. Here it can be seen that the FAR values go on decreasing for each succeeding level of BTC of respective color spaces. This shows that the accuracy of face recognition increases with increasing level of BTC and hence BTC-level 4 gives the best result with the least FAR value in both the color spaces. Also the FAR values of K'LUV are lesser than the RGB as shown in the figure. Thus, it can be concluded that the implementation of BTC levels based face recognition techniques is better when applied in K'LUV color space.

Figure 4 gives the GAR values of the different BTC levels based face recognition techniques tested on face database for both RGB and K'LUV color spaces. Here it is observed that with each successive level of BTC the GAR values go on increasing in respective color spaces and hence a BTC-level 4 gives the best result with the highest value in both the color spaces. Also the GAR values of K'LUV are greater than the RGB as shown in the figure. Thus, it can be concluded that the implementation of BTC levels based face recognition techniques is better when applied in K'LUV color space. For optimal performance the FAR values must be less and accordingly the GAR values must be high for each successive levels of BTC. Thus, the performance of K'LUV color space for BTC levels based face recognition techniques is superior to the performance of RGB color space for Face database.

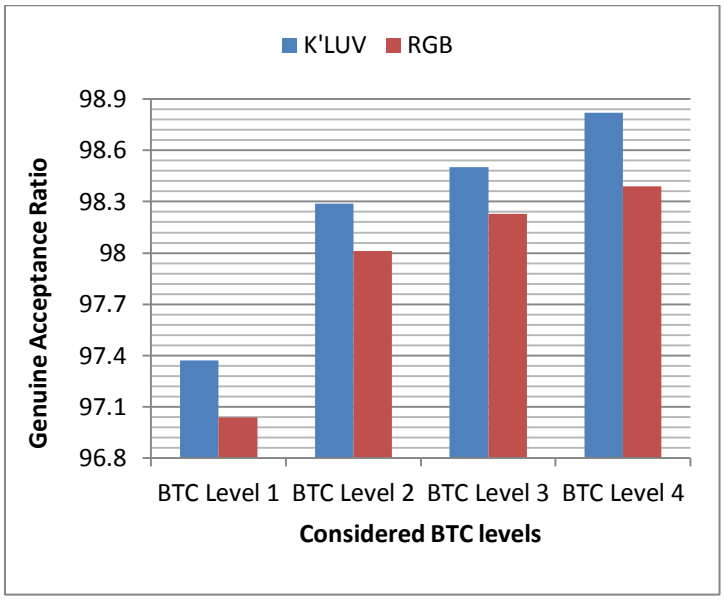


Figure 4. GAR values at different BTC levels of K'LUV and RGB color spaces for Face Database

B. Our Own Database

In all 16000 queries were tested on the database for analyzing the performance of the proposed BTC level based face recognition algorithm for both RGB color space and K'LUV color space. The experimental results of proposed face recognition techniques have shown that BTC level 4 gives the best performance in respective color spaces. The efficiency of the Multi-level BTC based face recognition increases with the increasing levels of BTC.

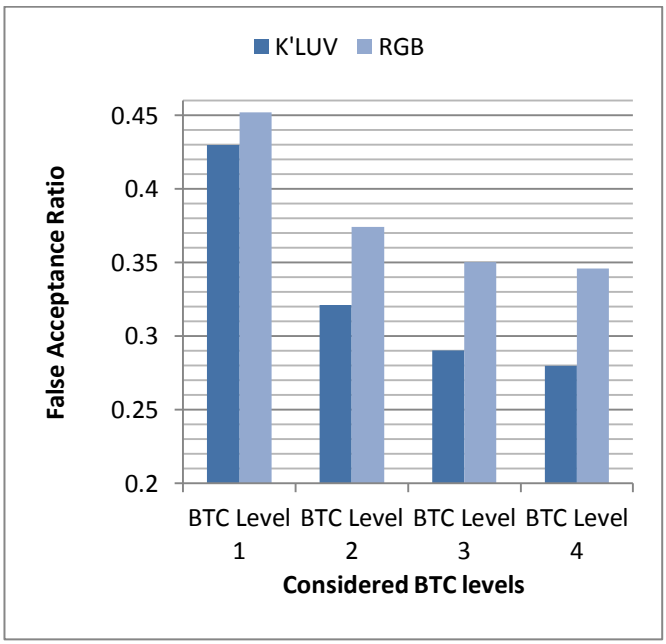


Figure 5. FAR values at different BTC levels of K'LUV and RGB color spaces for Our Own Database

Figure 5 gives the FAR values of the different BTC levels based face recognition techniques tested on Our Own Database for both RGB and K'LUV color spaces. The FAR values go on decreasing for each succeeding level of BTC of respective color spaces. Thus, when BTC based face recognition techniques is applied on Our Own Database, it gives a result similar to the Face Database; The BTC level 4 gives the best result for respective color spaces and K'LUV color space is better than RGB color space for implementing this proposed algorithm.

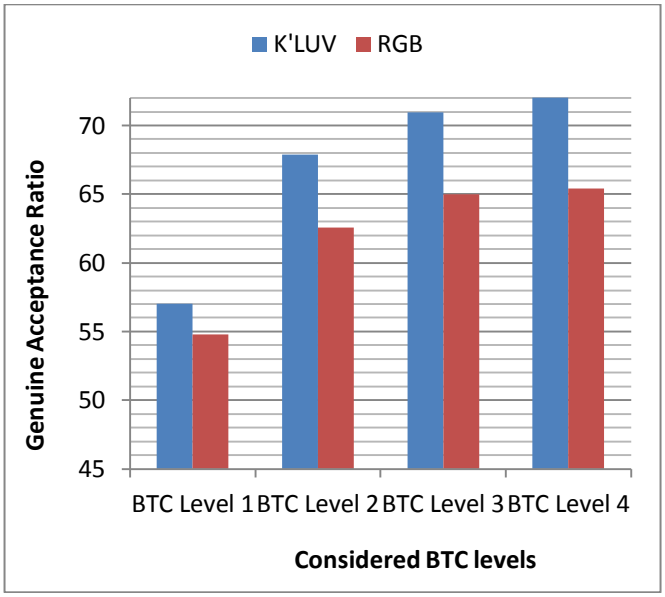


Figure 6. GAR values at different BTC levels of K'LUV and RGB color spaces for Our Own Database

Figure 6 gives the GAR values of the different BTC levels based face recognition techniques tested on Our Own Database for both RGB and K'LUV color spaces. It can be seen from the above figure that BTC-Level 4 has the highest GAR values and hence it is better than other BTC-Levels. Also the GAR values of K'LUV color space are greater than RGB color space at all the levels. Thus it can be concluded that the implementation of BTC levels based face recognition techniques is better when applied in K'LUV color space.

As seen from the performance of both the databases it can be concluded that the implementation of BTC based face recognition techniques on K'LUV color space is superior to RGB color space.

VIII. Conclusion

As the Multilevel BTC yields a greatly reduced feature space, this reduces the processing time required by the system. Thus, this system can be implemented is real time applications which generally require fast recognition time and have low computation power. The proposed face recognition system using Multilevel BTC has been tested using two face databases. For experimental analysis in all 10000 queries are fired on Face Database and 16000 queries on Our Own Database. The average FAR and GAR values of these queries clearly indicate that better performance is obtained when Multilevel Block Truncation Coding is employed using Kekre's LUV color space than RGB color space for face recognition.

References

[1] H.B.Kekre, Sudeep D. Thepade, Sanchit Khandelwal, Karan Dhamejani, Adnan Azmi, "Face Recognition using Multilevel Block Truncation Coding" International Journal of Computer Applications (IJCA) December 2011 Edition.

[2] Xiujuan Li, Jie Ma and Shutao Li 2007. A novel faces recognition method based on Principal Component Analysis and Kernel Partial Least. IEEE International Conference on Robotics and Biometrics, 2007. ROBIO 2007

[3] Shermin J "Illumination invariant face recognition using Discrete Cosine Transform and Principal Component Analysis" 2011 International Conference on Emerging Trends in Electrical and Computer Technology (ICETECT).

[4] Zhao Lihong , Guo Zikui "Face Recognition Method Based on Adaptively Weighted Block-Two Dimensional Principal Component Analysis"; 2011 Third International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN)

[5] Gomathi, E, Baskaran, K. "Recognition of Faces Using Improved Principal Component Analysis"; 2010 Second International Conference on Machine Learning and Computing (ICMLC)

[6] Haitao Zhao, Pong Chi Yuen" Incremental Linear Discriminant Analysis for Face Recognition", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics

[7] Tae-Kyun Kim; Kittler, J. "Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image" IEEE Transactions on Pattern Analysis and Machine Intelligence, March 2005

[8] James, E.A.K., Annadurai, S. "Implementation of incremental linear discriminant analysis using singular value decomposition for face recognition". First International Conference on Advanced Computing, 2009. ICAC 2009

[9] Zhao Lihong, Wang Ye, Teng Hongfeng; "Face recognition based on independent component analysis", 2011 Chinese Control and Decision Conference (CCDC)

[10] Yunxia Li, Changyuan Fan; "Face Recognition by Non negative Independent Component Analysis" Fifth International Conference on Natural Computation, 2009. ICNC'09'.

[11] Yanchuan Huang, Mingchu Li, Chuang Lin and Linlin Tian. "Gabor-Based Kernel Independent Component Analysis on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP).

[12] H.B.Kekre, Sudeep D. Thepade, Varun Lodha, Pooja Luthra, Ajoy Joseph, Chitrangada Nemani "Augmentation of Block Truncation Coding based Image Retrieval by using Even and Odd Images with Sundry Color Space" Int. Journal on Computer Science and Engg. Vol. 02, No. 08, 2010, 2535-2544

[13] H.B.Kekre, Sudeep D. Thepade, Shrikant P. Sanas Improved CBIR using Multileveled Block Truncation Coding International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2535-2544

[14] H.B.Kekre, Sudeep D. Thepade, "Boosting Block Truncation Coding using Kekre"s LUV Color Space for Image Retrieval", WASET International Journal of Electrical, Computer and System Engineering (IJECSE), Volume 2, Number 3, pp. 172-180, Summer 2008.

[15] H.B.Kekre, Sudeep D. Thepade, "Image Retrieval using Augmented Block Truncation Coding Techniques", ACM International Conference on Advances in Computing, Communication and Control (ICAC3-2009), pp. 384-390, 23-24 Jan 2009, Fr. Conceicao Rodrigous College of Engg., Mumbai

[16] Developed by Dr. Libor Spacek. Available Online at: http://cswww.essex.ac.uk/mv/otherprojects.html

[17] Mark D. Fairchild, "Color Appearance Models", 2nd Edition, Wiley-IS&T, Chichester, UK, 2005. ISBN 0-470-01216-1

[18] Rafael C. Gonzalez and Richard Eugene Woods "Digital Image Processing", 3rd edition, Prentice Hall, Upper Saddle River, NJ, 2008. ISBN 0-13-168728-X. pp. 407–413.S

[19] Dr.H.B.Kekre, Sudeep D. Thepade and Shrikant P. Sanas, "Improved CBIR using Multileveled Block Truncation Coding", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 07, 2010, 2471-2476

[20] Dr. H.B.Kekre , Sudeep D. Thepade and Akshay Maloo, "Face Recognition using Texture Feartures Extracted from Walshlet Pyramid ", Int. J. on Recent Trends in Engineering & Technology, Vol. 05, No. 01, Mar 2011.

[21] International Journal on Biometrics and Bioinformatics ,Volume 4 Issue 2 2010, http://www.slideshare.net/CSCJournals/international-journal-of-biometrics-and-bioinformaticsijbb-volume-4-issue-2.

[22] Dr. H.B.Kekre , Sudeep D. Thepade ,Varun Lodha, PoojaLuthra, Ajoy Joseph, Chitrangada Nemani,"Performance Comparison of Block Truncation Coding based Image Retrieval Techniques using Assorted Color Spaces", (IJCSIS) International Journal of Computer Science and Information Security,Vol. 8, No. 9,December 2010

AUTHORS PROFILE

Dr. H. B. Kekre has received B.E. (Hons.) in Telecomm Engineering from Jabalpur University in 1958, M.Tech. (Industrial Electronics) from IIT Bombay in 1960, M.S.Engg. (Electrical Engg.) from University of Ottawa in 1965 and Ph.D. (System Identification) from IIT Bombay in 1970 He has worked as Faculty of Electrical Engg. and then HOD Computer Science and Engg. at IIT Bombay. For 13 years he was working as a professor and head in the Department of Computer Engg. at Thadomal Shahani Engineering. College, Mumbai. Now he is Senior Professor at MPSTME, SVKM"s NMIMS University. He has guided 17 PhDs, more than 100 M.E./M.Tech and several B.E./B.tech projects. His areas of interest are Digital Signal processing, Image Processing and Computer Networking. He has more than 350 papers in National / International Conferences and Journals to his credit. He was Senior Member of IEEE. Presently He is Fellow of IETE and Life Member of ISTE Recently ten students working under his guidance have received best paper awards and two have been conferred Ph.D. degree of SVKM'sNMIMS University. Currently 10 research scholars are pursuing Ph.D. program under his guidance.

Dr. Sudeep D. Thepade has Received B.E.(Computer) degree from North Maharashtra University with Distinction in 2003, M.E. in Computer Engineering from University of Mumbai in 2008 with Distinction, Ph.D. from SVKM"s NMIMS (Deemed to be University) in July 2011, Mumbai. He has more than 08 years of experience in teaching and industry. He was Lecturer in Dept. of Information Technology at Thadomal Shahani Engineering College, Bandra (W), Mumbai, for nearly 04 years. Currently working as Associate Professor in Computer Engineering at Mukesh Patel School of Technology Management and Engineering, SVKM"s NMIMS (Deemed to be University), Vile Parle (W), Mumbai, INDIA. He is member of International Advisory Committee for many International Conferences, acting as reviewer for many referred international journals/transactions including IEEE and IET. His areas of interest are Image Processing and Biometric Identification. He has guided five M.Tech. projects and several B.tech projects. He has more than 130 papers in National/International Conferences/Journals to his credit with a Best Paper Award at International Conference SSPCCIN-2008, Second Best Paper Award at ThinkQuest-2009, Second Best Research Project Award at Manshodhan 2010, Best Paper Award for paper published in June 2011 issue of International Journal IJCSIS (USA), Editor's Choice Awards for papers published in International Journal IJCA (USA) in 2010 and 2011.

Sanchit Khandelwal is currently pursuing B.tech. (CE) from MPSTME, NMIMS University, Mumbai. His areas of interest are Image Processing and Computer Networks and security. He has 1 paper in an international journal to his credit.

Karan Dhamejani is currently pursuing B.tech. (CE) from MPSTME, NMIMS University, Mumbai. His areas of interest are Image Processing, Computer Networks and UNIX programming. He has 2 papers in an international journal to his credit.

Adnan Azmi is currently pursuing B.tech. (CE) from MPSTME, NMIMS University, Mumbai. His areas of interest are Image Processing and Computer Networks. He has 1 paper in an international journal to his credit.

Scenario-Based Software Reliability Testing Profile for Autonomous Control System

Jun Ai
School of Reliability and System Engineering
Beihang University
Beijing, China

Jingwei Shang
School of Reliability and System Engineering
Beihang University
Beijing, China

Peng Wang
School of Reliability and System Engineering
Beihang University
Beijing, China

***Abstract*—Operational profile is often used in software reliability testing, but it is limited to non-obvious-operation software such as Autonomous Control System. After analyzing the autonomous control system and scenario technology, a scenario-based profile constructing method for software reliability testing is presented. Two levels of scenario-based profile in the paper are introduced: system level and software level, and the scenario-based profile could be obtained through mapping them. With the method, the testing data for software reliability testing could be generated.**

Keywords- software reliability testing; scenario-based testing profile; autonomous control system.

I. INTRODUCTION

Software Reliability Testing(SRT) is used to ensure and verify software reliability requirement. The regular method is to perform a random statistical testing which is based on operational profile [1][2][3]. Operation profile delegates the probability distribution of the software usage, and the testing data could be generated by random sample with it[4][5][6]. But operational profile is limited to the Autonomous control system(ACS), in which the operation is not obvious.

Autonomous control system (ACS) usually has autonomy as it could run without people. The missile and unmanned plane are two typical examples [10]. It is a kind of non-obvious-operation system. The traditional operational profile could not be used in ACS directly. Consequently, it is necessary to develop a profile constructing method for ACS.

This paper describes a scenario-based software reliability testing profile (SRTP), which could solve the problem mentioned above about ACS. At first, the feature of ACS and the limitation of operational profile in SRT for ACS are analyzed. And then the conception and constructing method of scenario-based SRTP are presented combing with scenario technique. Finally, the feasibility would be validated by an instance.

II. SRTP REQUIREMENTS OF ACS

ACS does not have obvious operations. The running process of ACS is affected by system interactive, environment influencing and activity sequential. Therefore, there are some new technical requirements for the SRTP of ACS.

A. System-interactive:

System-interactive means the running of ACS software is affected by the cross-linking systems of ACS. At every moment, the stimulated inputs from the other cross-linking systems happen synchronously.

Different system interactions may cause different running modes and the states of the software are possibly determined by all inputs from the cross-linking systems. Consequently, all of the possible systems and their running processes should be presented in the SRTP.

B. Evironment-influencing:

Environment-influencing means ACS is sensitive to the change of environment. In other words, different strategies would be made along with the change by ACS. Autonomy control means online perceiving situation, making strategies and executing missions automatically according to definite assignments and principles [9]. Accordingly, to truly simulate software running processes, external environment in the life cycle of ACS should be described in testing profile.

C. Active-sequential:

Active-sequential means the system running periods could be apperceived by ACS in entire lifecycle, and all of the actives are sequentially. While working, real-time mission evaluation should be made, the mission in which period and the accomplishments should be determined by ACS. In addition, the next activity should be decided automatically too. Consequently, testing profile should be able to describe the mission periods and their time sequence.

According to these characteristics and SRT technical requirements of ACS, this paper proposes a testing profile constructing method based on scenario to support SRT for this kind of system.

III. SCENARIO-BASED SRTP

At first, the concept of usage scenario for software is made as follow:

usage scenario profile describes the possible time-series of software states and environments, which occur in the running process of software and system. For embedded software, the usage scenario profile should include both the active states of software and the environmental changes of embedded system.

Therefore, for the software of ACS, the usage scenario and its corresponding probability constitute its scenario-based SRTP. And the scenario-based SRTP can be express as:

$$\boldsymbol{SP}=\{<\boldsymbol{Sys}, \boldsymbol{So}>\} \quad (1)$$

$$\boldsymbol{Sys}=\{\text{Sc1, Sc2, Sc3，…，}\boldsymbol{Sci}\text{，…，Scn}\} \quad (2)$$

$$\boldsymbol{Sc}=\{<\boldsymbol{Pi}, \boldsymbol{Fj}>;\text{i=1,2,3,…,m; j=1,2,3,…,n }\} \quad (3)$$

$$\boldsymbol{P}=\{<\boldsymbol{Ti}, \boldsymbol{pi}>,\text{i=1,2,3,…,n}\} \quad (4)$$

$$\boldsymbol{So}=\{<\boldsymbol{Ai}, \boldsymbol{pi}, \boldsymbol{Lj},>\} \quad (5)$$

$$\boldsymbol{A}=\{<\boldsymbol{Ii}，\ \boldsymbol{Oi}, \boldsymbol{pi}>\} \quad (6)$$

In the formulas，***SP*** refers to the scenario-based SRTP; ***Sys*** refers to the system under test(SUT); ***So*** refers to the software under test; ***Sc*** refers to cross-linking system; ***P*** refers to parameter of the interactive system; ***F*** refers to function of the interactive system; ***T*** refers to equivalence class; ***A*** refers to activity; ***L*** refers to layer of the activity; ***I*** is a input of the activity; ***O*** refers to output of the activity; ***p*** refers to the corresponding probability.

As mentioned above, a complete usage scenario profile should include descriptions of both the system's and software's running process and their statistical distribution. The constitution of usage scenario profile is shown in Figure 1. :

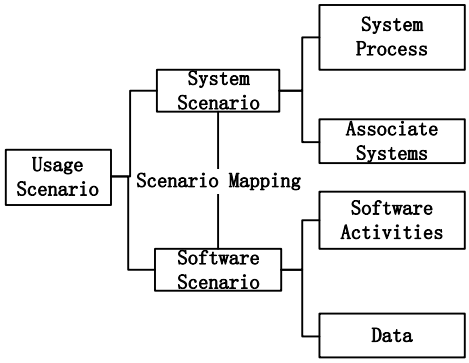


Figure 1. Constitution of usage scenario profile

In other words, the usage scenario profile consists of system scenario profile and software scenario profile. System scenario profile is used to describe the running process of ACS and its statistical distribution, including system running states, environments, and external driving sources. Software scenario profile is used to describe software running states and possible inputs and their distribution. System scenario profile is looked as the constraints of software scenario, which define the possible running condition of the software; while software scenario is the detail representation of system scenario. The combination of the two scenario profiles performs an expression of the software usage.

IV. SCENARIO-BASED SRTP CONSTRUCTION

According to the analysis of scenario-based SRTP, the main constructing process can be shown in Figure 2. .

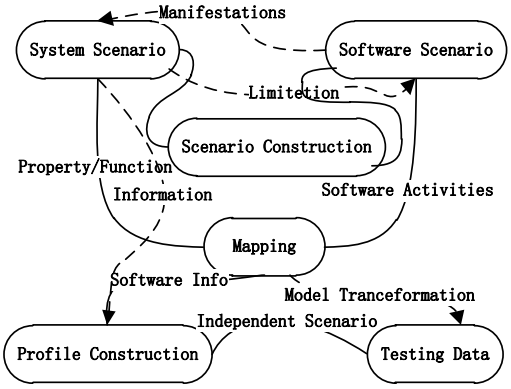


Figure 2. The constructing process of scenario-based SRTP

At first, the usage scenario should be analyzed and modeled in system level and software level. In this way, the system scenario profile and software scenario profile could be constructed. And then the mapping from system scenarios to software scenarios according to their properties would be made. In this step, independent system scenario should be extracted from the system scenario profile, which would be used to analyze software scenario profile. After the analysis, the input and output during the software process could be obtained, and the corresponding SRT data could be generated.

During the constructing process, some UML standard diagrams could be used to describe the usage scenario profile. The details are introduced as follows:

A. System Scenario profile Construction

Scenario-based SRTP includes description of the software under test and associate systems during the software process. As a top-to-bottom constructing method, analysis and construction of the associate systems should be done firstly. The main purpose of system scenario profile construction is describing system's states and environments during the system's running process and all the factors which could affect the running of system.

Commonly, there are four steps in the system scenario profile construction:

1) *Define associate systems:* including the SUT, associate system, user, and enviroment;
2) *Define external interactions of the systems:* including the interactions between SUT and other systems;
3) *Define the running input data:* including the input data lead to the change of the state of SUT;
4) *Define the factors and their distribution:* including all the factors which could affect the running of system and the occurrence probability of the factors.

The Class Diagram is used to describe system's states and environments. The variables in the class are the data which could be exchanged among the SUT and the state of system is decided by the variables, associate system, user and environment. The operations in the class could be used to describe the relationship among the variables, for example:

Data A is twice of Data B. An example of system and its interactions is expressed as Figure 3.

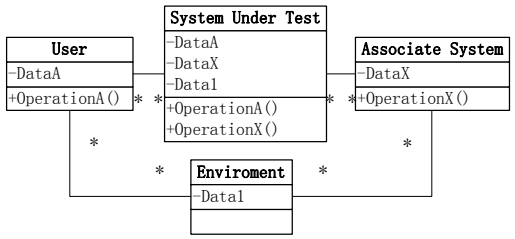


Figure 3. System and its interactions

In the system scenario, activities of users, interactions of cross-linking systems, effect of environment, and other influence from the SUT should be described. The factors which could affect the running of system should be listed and possible value range or value trend of the factors should be given. The factors and its value distribution could be expressed with a system scenario profile table, such as TABLEI.

TABLE I. EXAMPLE OF FACTOR TABLE

Factor	Value range	Probability	Source	Target	….
Altitude	1~100	0.3	Altitude meter	System	…
	100~1000	0.7	Altitude meter	System	…
Speed	<20km/h	0.4	Speed meter	System	
…	…	…	…	…	

B. Software Scenario Profile Construction

The main purpose of this period is to construct the scenario model of software running process and its input infomation.

1) Activity dividing: dividing task of the software according to the main purpose of each activity.

2) *Use case modeling:* constructing functional use case models for the software according to the mission and process.

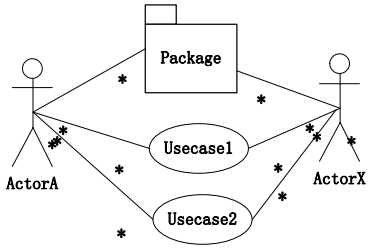


Figure 4. Use case model example of software function

As shown in Figure 4. , use case model (which is expressed by Use case Diagram) is used to analyze functional requirements of the software. However, the detailed description for each function is not necessary.

3) *Scenario flow analyzing:* software scenario can be determined by a path through use cases. A path starts from a use case and ends with a use case, traversaling through all of the basic-flow and alternative-flow.

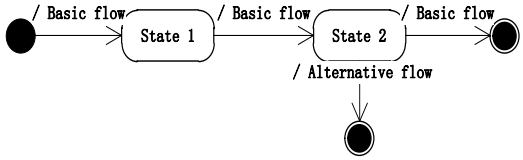


Figure 5. Software scenario flow example

As shown in Figure 5. , software scenario flow could be expressed by State chart Diagram. Each path describes a basic-flow or an alternative-flow with an arrow-line.

Basic-flow: an arrow-line which is the simplest path through the regular use cases; in this path the program should run from beginning to the end with no error.

Alternative-flow: an arrow-line which starts from a basic-flow or an alternative-flow in some specific conditions, and rejoin the basic-flow or ends with a use case.

Combining the basic-flow and alternative-flows, a scenario table can be obtained, such as TABLE II.

TABLE II. SOFTWARE SCENARIO FLOW

S 1	BF		
S 2	BF	AF 1	
…	…	…	…
S 8	BF	AF 3	AF 4

4) *Scenario flow Refine:* analyzing the basic-flow and alternative-flows according to the stimulation and sources, the equivalence class of inputs could be divided.

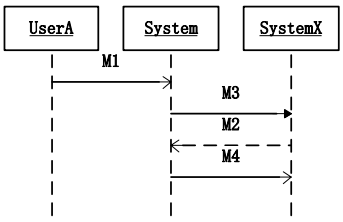


Figure 6. Software scenario flow refine example

It could be expressed by Sequence Diagram, as shown in Figure 6. The arrow-line in the diagram presents messages and data in the interactions.

5) *Scenario Splicing:* A software scenario with inputs and output could be obtained by splicing sequence diagrams in the "scenario flow refine" process according to the orders in the basic-flow or alternative-flow.

Generally, a software scenario is an instance of the use case. The software scenario flows through some use cases,

defines the relevant data and covers the use case flows through the path. Data associates with specific scenarios by means of input/output and some intermediate state. There would be multiple use cases in a complex scenario. The operations between use cases are described by execution orders, controlling conditions, parallels, or loops.

C. Scenario Mapping

System scenario and software scenario describe the running process of the software under test in macro and micro view. Using them without connections could not satisfy the testing requirements mentioned before.

Decision table could be used to map system scenarios with software scenarios. A decision table example is shown in TABLE III. , where each row represents software scenario and each column represent a system scenario. Each software scenario should contain the scenario ID, conditions (or states), all data and the expected results elements involved. The decision table should at least include necessary system scenario information in executing a software scenario.

TABLE III. SCENARIO MAPPING DECISION TABLE EXAMPLE

S ID	S	Data X1	Data Y1	Active A1	Result
S1.	BF	V	V	V	A1
S2.	BF-AF2	V	V	V	Active1-Active2
S3.	BF-AF	I	V	n/a	Active1-Active3-Active1

Here V means the basic-flow could be executed only when the condition is valid; the I means in the invalid condition the alternative-flow would be activated; and n/a means the condition is not applicable to this scenario.

By scenario mapping, the activities in software scenario could be Modular with details, and a scenario combing system scenario with software scenario could be obtained, which could be used to generate software reliability testing data.

D. Distributing probability

By distributing probabilities to each data, activity, and branches in the scenario diagrams and tables, a scenario profile with both scenarios and probabilities is produced.

Then we could obtain software reliability test cases by extracting a certain scenario and necessary data in describing scenario from the profile.

Therefore, scenario-based SRTP construction is an iterative process of refinement. Describe system and software behaviors in different points of view.

V. INSTANCE VALIDATION

An ACS software has been taken as an instance, which is a flight control system for a model airplane. In the issue, we simplified the detail data of the system. The scenario-based SRTP is constructed as follow:

A. System Scenario Profile Constructing

By analyzing the embedded ACS, the system scenario contains the system under test, seeker, fuse, environment and the carrier aircraft.

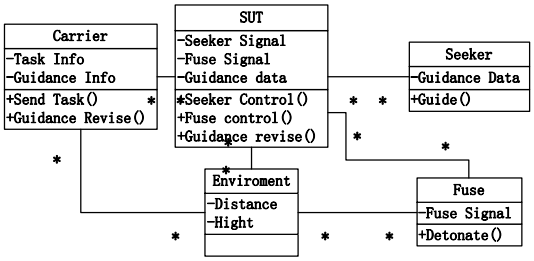


Figure 7. ACS system scenario construction diagram

We analyzing the input/output and activity of each object, and construct a system scenario as Figure 7, and the factor table is as TABLE I. ACS system scenario contains description of both objects and their factors.

B. Software Scenario Profile Construction

Construct the functional use case model as Figure 8. :

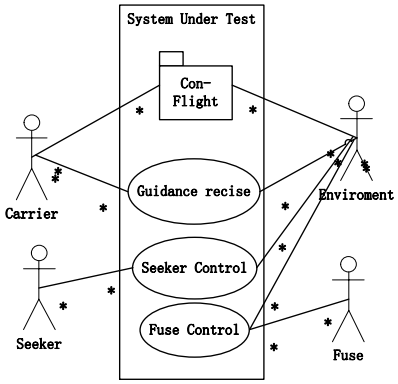


Figure 8. ACS functional use case diagram

A complex use case sequence contains guidance revise, seeker control, and fuse control is selected as an example for scenario flow analyzing, as in Figure 9. :

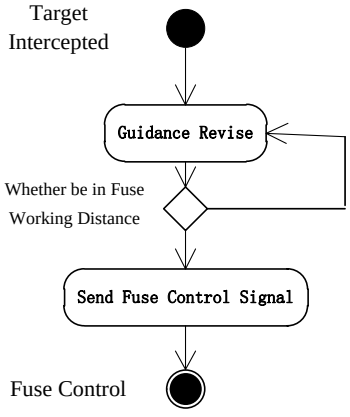


Figure 9. ACS scenario flow diagram

Form the system scenario flow table:

TABLE IV. SIMPLIFY ACS SCENARIO FLOW TABLE

S 1	Guidance Revise		
S 2	Guidance Revise	Seeker Control	Target Lost
…	…	…	…

The refined scenario flow sequence diagram will be express in the following process.

C. Scenario Mapping

Decision table based for scenario mapping can be demonstrated as TABLE V. :

TABLE V. SIMPLIFY ACS SCENARIO MAPPING DECISION TABLE

S ID	Guidance Revise Data	Seeker Working Distance	Fuse Working Distance	Fuse	Result
S1.	V	I	n/a	n/a	Guidance Flight
…	…	…	…	..	…
S4.	V	Seeker Control	Fuse Control	S 4	Guidance Revise

D. Scenario Profile and Test Data

By distributing probabilities to each data, activity, and branches in the scenario, the scenario profile with scenarios and probabilities is generated. Then a scenario such as figure 12 could be extracted.

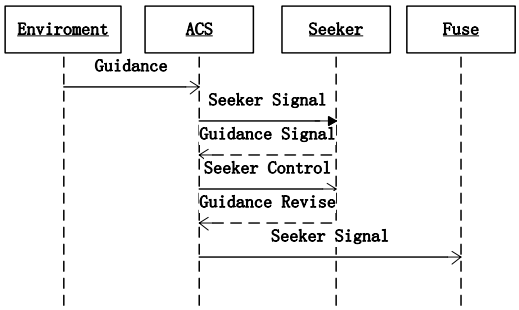


Figure 10. scenario sequence diagram

According to the system scenario factors in the scenario mapping table, the properties of system scenario are determined. By make a secant ling through the ACS scenario sequence diagram, data across the secant line is the test data in the very scenario.

A scenario-based SRTP constructing process for the ACS is demonstrated above. In particular, some data are simplified in the process.

VI. SUMMARY

A scenario-based SRTP constructing method is proposed in this paper. With system scenario profile and software scenario profile, the possible running state and usage of the SUT could be modeled.

TABLE VI. SIMPLIFY TEST CASE

Test case Id: T01-01
Input: 1) Input Guidance Data XXX，XXX……； 2) In Seeker Working Distance, Input Seeker Control Signal XX……； 3) Seeker Guidance Data XX……； 4) In Fuse Working Distance, Input Fuse Control Signal XX……； 5) Input Fuse Signal XX……

According to scenario profile, the software reliability test case could be generated. The problem that traditional operation profile could not be used in ACS has been resolved. With the instance of a flight control system for model airplane, the method in this paper is effective.

REFERENCES

[1] Lu Minyan, Chen Xuesong,"Software Reliability Testing and Practice".Testing and Control Technology, vol. 19, 2000,pp. 509-512.

[2] Musa J D."Operational profiles in software reliability engineering". IEEE Software , vol.10(2) ,1993, pp.14-32

[3] Ai, Jun ,Lu, Min-Yan, "The analysis and modeling for the input space of real-time embedded software", 8th International Conference on Reliability, Maintainability and Safety, 2009, pp. 774 - 777

[4] Elbaum, S.; Narla, S." A methodology for operational profile refinement", Reliability and Maintainability Symposium ,2001, pp. 142 - 149

[5] Saravana Kumar. K., Misra. R.B.and Goyal, N.K., "Development of Fuzzy Software Operational Profile",Secure System Integration and Reliability Improvement,Second International Conference on 2008, pp.195 – 196

[6] Gittens, M.,Lutfiyya, H.and Bauer, M."An extended operational profile model",15th International Symposium on Software Reliability Engineering, 2004, pp.314 - 325.

[7] Xiaofeng L,Chenggang B and Liang S,"Software Operational Profile Modeling and Reliability Prediction with an Open Environment ",10th International Conference on Quality Software (QSIC), 2010,pp.227 - 231

[8] Urem, Frane, Mikulic and Zelimir," Developing operational profile for ERP software module reliability prediction ",2010 Proceedings of the 33rd International Convention , pp. 409 - 413.

[9] Huang Yongfei, Peng Xinjie, "Missle Flight Control System Testing" , Missle and Guidance,Vol.29,2009, pp. 65-67.

[10] Ma Sasa,Chen Zili, Li Bing "UAV Flight Control Software Testing Technique ", Radio Engineering,Vol.34(10),2004. pp. 53-57

AUTHORS PROFILE

Jun Ai is the deputy director of Center of Software Dependability in School of Reliability and System Engineering, BeiHang University, China. He received a PHD in system engineering from BeiHang University. His research interests include software reliability and safety, software testing and software engineering.

Jingwei Shang and Peng Wang are graduate students in School of Reliability and System Engineering. Their research interests are about software reliability testing and evaluation.

Identification of Critical Node for the Efficient Performance in Manet

Shivashankar
Asst.Professor
Medical Electronics
Dr.Ambedkar Institute of Technology,
Bangalore 560 056, India

B.Sivakumar
Professor&HOD
Telecommunication Engg,
Dr.Ambedkar Institute of Technology

G.Varaprasad
Associate Professor
Computer Science Engg
B.M.S.College of Engineering
Bangalore 560 019, India.

***Abstract*— This paper considers a network where nodes are connected randomly and can fail at random times. The critical-node test detects nodes, whose failures are malicious behavior, disconnects or significantly degrades the performance of the network. The critical node is an element, position or control entity whose disruption, is immediately degrades the ability of a force to command, control or effectively conducts combat operations. If a node is critical node, then more attention must be paid to it to avoid its failure or removal of a network. So how to confirm critical nodes in the ad hoc network is the premise to predict the network partition. A critical node is the most important node within the entity of a network. This paper suggests methods that find the critical nodes of a network based on residual battery power, reliability, bandwidth, availability and service traffic type. The metrics for evaluation has been considered as packet delivery ratio, end-to-end delay and throughput.**

Keywords- Critical node; malicious; residual battery power; reliability; bandwidth; Mobile Ad hoc Network.

I. INTRODUCTION

Nodes in Mobile Ad hoc Networks (MANETs) are battery powered and hence have limited lifetime. Due to excessive utilization a node may die which can result in energy depletion problem and thereby affect overall network performance. Local rate of energy consumption based on application can be monitored. Thus early detection and avoidance of energy depletion problem is possible based on monitoring of power and remaining battery lifetime. As and when the rate of consumption based on the application returns to efficient levels, the mobile relay is released so that it can be made available for other critical nodes. This technique can be applied to any protocol used in MANETs such as DSDV, DSR, AODV, TORA etc.

In this work, we focus on maximizing avoidance of network partition, rather than after the network partition, then taking some remedial measures. We describe the critical nodes compensation approach to increase the network connectivity and improve packet delivery rate. First, we detect the critical nodes which may lead to the network partitioning. Secondly, propose the compensation algorithm to avoid the network partitioning. This approach described in this paper is built around the notion of a critical node in an ad hoc network. Our definition of a critical node is a node, whose failure or malicious behavior disconnects or significantly degrades the performance of the network. Once identified, a critical node can be monitored in terms of resource. If a node is not considered critical, this metric can be used to decide if the application or the risk environment warrant the expenditure of the additional resources required to monitor, diagnose, and alert other nodes about the problem. Determining the global network topology in a MANET gives the time delays of the diagnostic packets and the mobility of the nodes makes this task futile, but determining an approximation of this topology or subset of this topology, within a certain time frame may be useful.

An approximation of the network topology can still provide useful information about network the density, network mobility, critical paths, and critical nodes. Even with the uncertainty associated with correctly reconstructing the network topology for a given time period, this additional information can help to reduce the resources consumed to monitor all nodes in the absence of this information.

The node performing the test is referred to as the testing node and the node being tested is referred to as the node under test. Three steps are required to detect whether a testing node shares a critical link with its neighbor. First step is to temporarily modify the testing node's routing table to allow only one communication link to be operational at a time, while blocking communication through all others. The enabled communication link will be between the testing node and a node other than the node under test. Each communication link will be tested sequentially in this manner to determine if an alternative path to the link under test exists.

The responsibilities of a routing protocol include exchanging the route information, finding a feasible path to a destination based on criteria such as hop length, minimum power required and life time of the wireless link. Gathering information about the path breaks, mending the broken paths expending minimum processing power, bandwidth and utilizing the minimum bandwidth. Fig.1 shows the graphical view of identification of the critical node in the network.

If the critical node find in the network, then alternate path will be selected for the efficient network. The major challenges are:

A. Ability to Measure the Resource Availability

In order to handle the resources such as bandwidth efficiently and perform call admission control based on their availability, the MAC protocol should be able to provide an estimation of resource availability at every node.

B. Capability for Power Control

The transmission power control reduces the energy consumption at the nodes, causes a decreasing interference at neighboring nodes and increases frequency reuse.

C. Mobility

One of the most important properties of MANET is the mobility associated with the nodes. The mobility of the nodes results in frequent path breaks, packet, collisions, transient loops, stare routing information and difficulty in resource reservation. A good routing protocol should be able to efficiently solve all the above issues.

The rest of the paper is organized as follows. In section II, some previous work related to this paper and some applications are identified. In Section III, we present the different mechanisms with algorithms and full connectivity when the only randomness is due to random connections and there are no node failures. Section IV contains the simulation results, analyzing and comparison of different parameters and also it includes the discussion on the performance by the various protocols agents in our simulations. Finally, Section V concludes the paper.

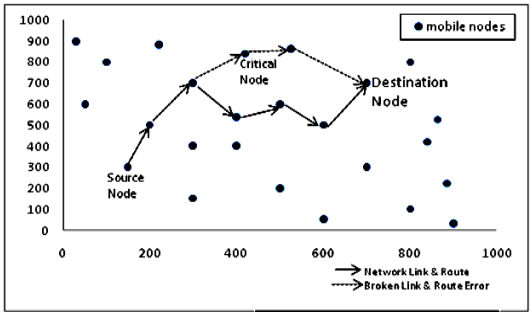


Figure 1. Critical Node representation in network.

II. Some of the Related Research Work

A. Critical Nodes Compensation Algorithm in Sparse Ad hoc Wireless network

Sparse ad hoc wireless networks [1] are an active application in recent years, though sparse environment often makes the network unconnected. However, most of the previous researches focus on connected networks where an end-to-end path exists between any two nodes in the network, therefore these technologies would not work well in sparse ad hoc networks, where restricted node communication radius can result in periods of intermittent connectivity. In this work, we develop a critical nodes compensation algorithm in order to prevent network from partitioning, thereby insure network connectivity and throughput.

But the drawback of previous studies is that they overlooked the importance of the service traffic flow [2] and non-articulation nodes. We propose a method that evaluates how important each node is based on service traffic flow and articulation node. Based on this evaluation, we can give priority to each node. In a complex network, it is useful for a network administrator to focus mainly on some high-ranked nodes, which are relatively important for security purposes.

B. Determining Node Priority Order Based on Traffic Volume

Assuming that there is a web server with a small amount of traffic volume and a backup server with huge amount of traffic volume, it is possible for the network administrator to value the web server because it needs to be alive all the time compared to backup server. In this case, the web server node is more important in spite of its low traffic volume [5].

C. Depth First Search Algorithm(DFS)

DFS was used to detect critical nodes in [3]. DFS is a centralized algorithm and can be also implemented in globalized distributed manner. The algorithms in [4] require that a node should be aware of global topology. In practice, this method is inefficient and involves a quadratic (in number of nodes) communication overhead. The node density is the average number of neighbor nodes of a node in the network. This value reflects the density of the network. The larger the value is, the denser the network is [6]. In this paper, we present a novel critical node detection algorithm for wireless ad hoc networks. It is an effective distributed localized algorithm, which greatly reduces communication overheads and the speed of detection. This is a distributed algorithm which adapts the dynamic topology adaptively, detects the critical node faster and more reliably, and decreases the detection overheads efficiently.

D. Bandwidth Balancing

The bandwidth balancing [7] solves the fairness problems suffered by long bandwidth in the networks. By constraining each node to take only a fraction of the available slots, bandwidth balancing can achieve a fair operation point when several nodes are performing large file transfers.

As routing protocols exchange routing data between nodes, as a result, they would maintain routing status in each node. Based on routing status, data packets are transmitted by mediated nodes along an established route to the destination [8]. M.K Rafsanjani, A Movaghar presents a scheme in which nodes do not need to exchange multiple messages to prove their identities [10]. However, most of the previous researches[11,9] focus on connected networks where an end to end path exists between any two nodes in the network, therefore these technologies would not work well in sparse ad hoc networks, where restricted node communication radius can result in periods of intermittent connectivity.

As routing protocols exchange routing data between nodes, as a result, they would maintain routing status in each node. Based on routing status, data packets are transmitted by mediated nodes along an established route to the destination [12]. In [13], the authors describe a distributed intrusion detection system for MANETs that consists of the local components data collection, detection and response and of the global components. Whereas their architecture is very promising and similar to the one we use in our paper, they

neglect the aspect how their local data collection should find out on incidents like dropped packets, concealed links, etc.

III. DESIGN AND IMPLEMENTATION

A. Bandwidth Constraint in MANET

Since the channel is shared by all nodes in the broadcast region (any reason in which all nodes can hear all other nodes), the bandwidth available per wireless link depends on the number of nodes and a traffic they handle. Thus only a fraction of the total bandwidth is available for every node. The control over head involved must be kept as minimal as possible.

Bandwidth efficiency can be defined as the ratio of the bandwidth used for actual data transmission to the total available bandwidth. The limited bandwidth availability also imposes a constraint on routing protocols in maintaining the topological information. Due to the critical nodes, maintaining consistence topological information at all the nodes involves more control over head which, in turn, results in more band width.

Available bandwidth calculation for the MANET:

Sufficient Bandwidth = $BW_{available} = \frac{Tidle}{T} \times W$ ------(1)

T=Sampling time window for calculating real time bandwidth.
T_{idle}= Time period when a mobile node in idle mode.
W= Maximum bandwidth for data transmission.

B. Location Dependent Contention

The load on the wireless channel varies with the number of critical nodes present in a given geographical region. This makes the contention for the channel high when the number of nodes increases. The high contention for the channel results in a high number of collisions and a subsequent wastage of bandwidth. A good routing protocol should have built in mechanism for distributing the network load uniformly across the network so that the information of regions where channel contention is high can be avoided.

C. Quick Route Reconfiguration

The unpredictable changes in the topology of the network require that the routing protocol be able to quickly perform route configuration in order to handle path break and subsequent packet losses because of critical nodes in the network.

D. Loop Free Routing

This is a fundamental requirement of any routing protocol to avoid unnecessary wastage of network bandwidth. In MANET, wireless network due to the random movement of critical nodes, transient loops may form in the route. A routing protocol should detect such critical nodes and take corrective actions.

E. Signal Strength Based Reliability

A node, N can measure signal strength of its active neighbors. The received signal strength is measured at physical layer and made available to the access of top layers.

$$R(\text{node}) = P(\text{new}) / [P(\text{old}) + P(\text{new})] \quad \ldots\ldots\ldots (2)$$

Where R(node) denotes node reliability and it will be a value between 0 and 1. Whatever new signal strength is more than before, R(node) approaches towards one and whatever new signal strength is less than before, R(node) approaches towards zero.

F. Reliability

The reliability of the real time data being transmitted can be enhanced by introducing a buffer of a fixed size calculated through monitoring the data transmitted by employing self-healing nodes.

G. Residual Battery Power

Power conservation in wireless ad hoc networks is a critical issue as energy resources are limited at the electronic devices used. Therefore to conserve battery energy of the nodes, there are various routing algorithms and schemes designed to select alternative routes. These algorithms and schemes are collectively known as 'power-aware routing protocols' and an example of a better choice of routes selected is one where packets get routed through paths that may be longer but that pass through nodes that have plenty of energy reserves. Some or all of the nodes in a MANET may rely on batteries or other exhaustible means for their energy. For these nodes, the most important system design criteria for optimization may be energy conservation. ..... (1)

H. Minimize Energy Consumed/Packet

This is one of the more obvious metrics. To conserve energy, we want to minimize the amount of energy consumed by all packets traversing from the source node to the destination node. That is, we want to know the total amount of energy the packets consumed when it travels from each and every node on the route to the next node. The energy consumed for one packet is thus given by the equation:

$$E = \sum_{i=1}^{k-1} T(n_i, n_{i+1}) \quad \text{---------- (3)}$$

In equation (1), n1 to nk are nodes in the route while T denotes the energy consumed in transmitting and receiving a packet over one hop. Then we find the minimum E for all packets. However, this metric has a drawback and i.e. nodes will tend to have widely differing energy consumption profiles resulting in early death for some nodes. Power aware routing schemes make routing decisions to optimize performance of power or energy related evaluation metric.

Excessively, conserving energy neglects power consumption at individual nodes, which speeds up network partition by draining batteries of the critical nodes in the network one by one. In effect, it shortens the network lifetime. On the other hand, overly conserving power expels energy consideration, which commits to paths with large number of hops and longer total distance. Consequently, the total energy dissipated is high and on average, the battery power decays faster. In effect, it also shortens the network lifetime. Energy efficiency is also an important design consideration due to the limited battery life of wireless node. Since, the network interface is a significant consumer of power, considerable research has been devoted to low power design of the entire

network protocol stack of wireless network in an effort to enhance energy efficiency.

All mobile should drain their power at equal rate as a minimal set of mobile exist such that their removal cause network to partition. Such node is called as a critical node. The route between these two partitions must go through one of these critical nodes. A routing procedure must divide the work among these nodes to maximize the life of the network. This problem is similar to load balancing problem. A packet to be routed through a path contains mobiles having grater amount of energy though it is not a shortest path. Delay is minimized as no congestion and nodes having less number of loads.

I. *Availability*

MANETs establishing trust relationships between the nodes in a decentralized fashion has been an important research issue for a long time. If the sender nodes accurately identify the legitimate nodes in the network, a robust routing can be provided while mitigating the effects of malicious nodes. Further, there is always a mutual interaction between a sender and its neighbor nodes during the communication. The scheme guarantees the availability of message as long as a legitimate path exists. Through simulations, we will show the efficiency of the scheme with respect to latency, availability and energy consumption in the presence of adversary.

J. *Residual Capacity*

Residual capacity at a node is the difference between the node's channel capacity and the sum of the bandwidth consumed by all contending flows of that node. It denotes the channel capacity that is not used and it constrains the rate that contending flows can acquire. To measure the residual capacity, each node in the network monitors the channel activity. The fraction of channel idle time during the past measuring period and the channel capacity, determine its residual capacity. The residual capacity at node k can be expressed as

$$R_k = \frac{T_{idle}}{T_P}(C_k) \qquad \text{----------(4)}$$

In equation(4), Ck is the channel capacity at node k; Tidle is the channel idle time during the last measuring period Tp. Larger Tp will give more accurate channel view but also longer response time for the source node to react to the change in the network. In this paper, Tp is set to 0.5 second.

IV. Simulation Setup and Result Discussion

Network Simulator (NS-2.33) has a very rich component library, which is compiled of two languages: C++ or Python is object oriented extension of TCL. First of all, we define the simulation in the 1,000 m×1,000 m region, random waypoint mobile model, the node number: 65; node original communication radius: 40m; according to the distance between compensation node and critical node, choose compensation transmission power; in order to maintain low-speed movement environment, node movement speed is not more than 2 m/s.

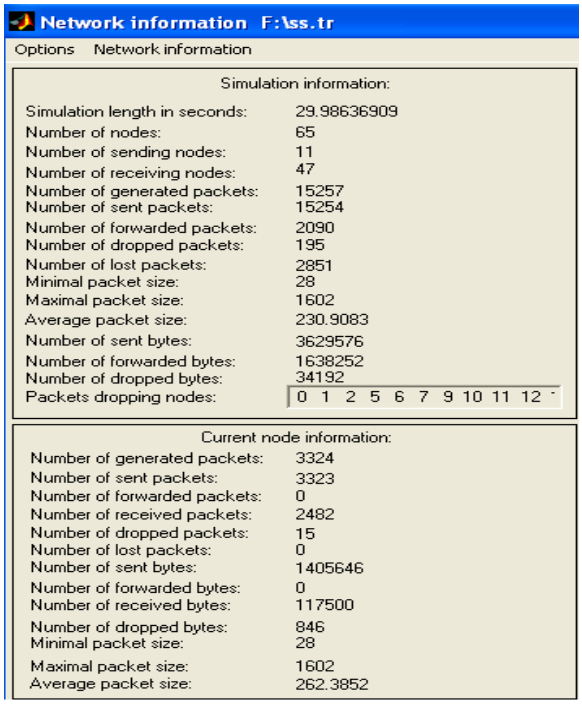
We choose two ad hoc network performance parameters: critical node number and network throughput. For our simulation, we used NS-2.33 with mobility extension. These extensions include the modeling of an IEEE 802.11/MAC.

Table 1 shows the simulation parameters used in the network setup for identifying the critical node and select the alternate path for maintaining the continuous efficient network connection in MANET.

TABLE 1. Simulation Parameters.

Simulation time	0-50 sec
Traffic type	TCP
Packet size	1460
Hello packet interval	2 sec
Node mobility	0 to 10 mts/sec
Frequency	1 Ghz
Channel capacity	2 M bps
Transmit power	2.0 Mw
Receiver power	2.0 Mw
Total number nodes	65
Communication system	MAC/IEEE 802.11G

TABLE 2. Overall network information with critical node and without critical node in MANET.



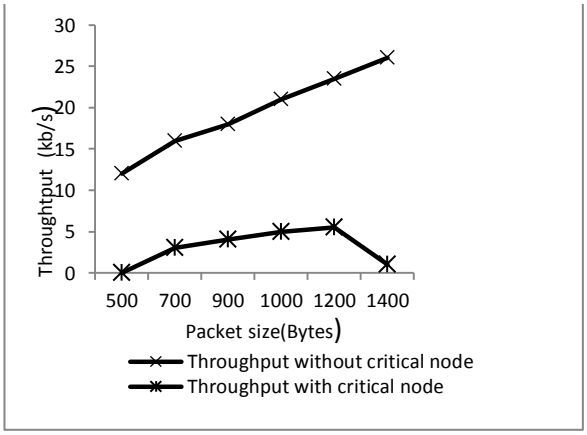


Figure 2. Throughput with and without critical node in MANET.

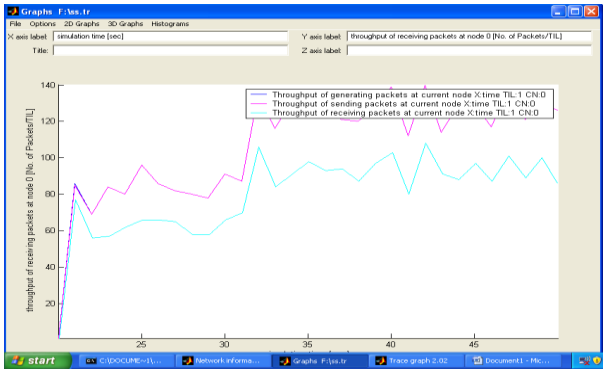


Figure 3. Throughput of receiving packets when there is critical node occurs during the transmission of data.

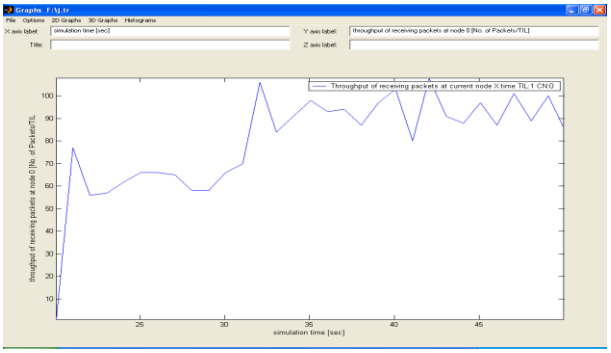


Figure 4. Throughput of receiving packets during the transmission of data.

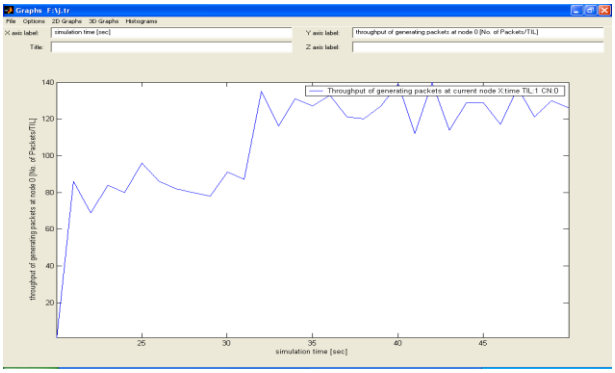


Figure 5. Throughput of generating packets at destination node.

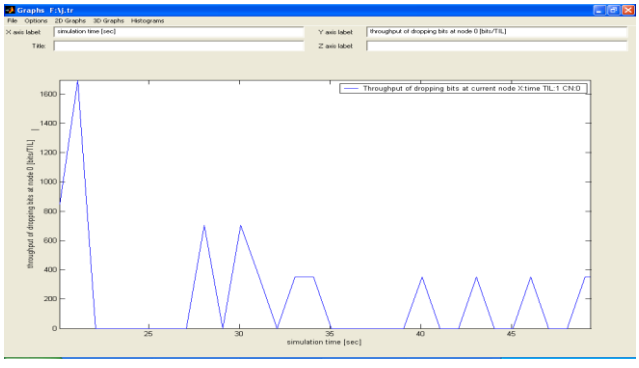


Figure 6. Throughput of dropping bits at destination node.

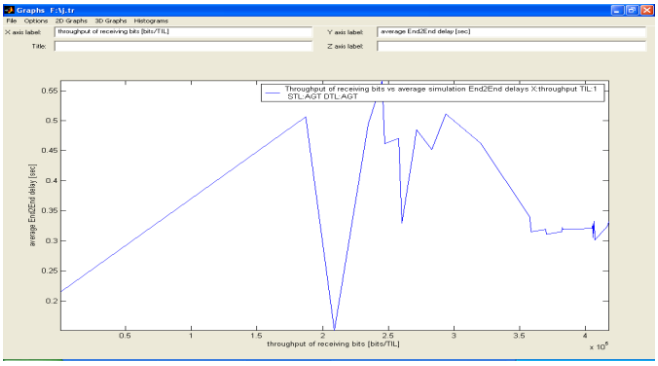


Figure 7. Average of end-to-end delay.

Figure-2 shows the throughput response of the MANET is 78% difference with and without critical nodes. In this paper, the performance analysis of MANET for different scenario as mention in Table-2 has been studied using trace graph. The first part of the table-2 shows more dropping packets, more number of lost packets in the presence of critical node as intermediate node. The second part of the table-2 shows that the improvements in MANET approximately 94% with more received packets, less dropped packets and less number of lost packets.

The figure-3 represents the throughput performance in MANET. The Y axis shows throughput of receiving packets at destination node in kilo bytes and the X axis shows simulation time in sec. The red line of the graph represents the best effort with 92% of the throughput. The green line of the graph shows the throughput of critical node presence and it is approximately 52% of the throughput. We start the best effort approximately started at 12 sec and going on until 50 sec. The presence of critical node in MANET, results in large delay. Since the delay has been increased, the throughput is gradually reduced. When the neighbor nodes are communicating with each other, the packets will move between source and destination and thus the throughput will be increased.

Figure-4 plots the receiving packets during the transmission time v/s simulation time after identified the critical node and alternate path has been selected. It shows the sum of numbers of all the intermediate nodes, receiving packets sent by the source node and number of received packets at the destination node. There is some packet losses (6% to 12%) for the rates of 78 to 90pkts/sec for both the schemes on some of the nodes. This may be due to insufficient bandwidth, less residual battery power and minimized energy consumed per packet. If these metrics are good, then throughput of the network will be increased.

Figure-5 shows the throughput of generating packets at any intermediate nodes v/s simulation time (sec).The graph reflects the simulation time for which an intermediate node in route generated packets. In other words, it depicts how long the route through the intermediate node was valid during the simulation after selected alternate path for the best efficiency.

The throughput for the proposed Networks as shown in figure-6, are calculated based on the distance. The throughput has dropped around 45.3%. The number of data packets dropped at any critical node present in the network. This is an important parameter because if the number of dropped packets increases, the throughput would decrease. Therefore the lower packets drop; lower would be delay in the network.

Figure-7 represents the average of end to end delay for both the schemes (with and without critical node). Though the nodes are at same distance from the route, they receive different bandwidth. This is the average delay of all the data packets. The delay different is 50% of the throughput of receiving bytes. It is calculated as the time taken between the generation of data packets and arrival of last bit of destination. There are possible delays caused by bandwidth, throughput, average number of node receiving delay between current and other node etc. This metric describes the packet delivery time. The lower end to end delay is the better application performance in MANET.

V. CONCLUSION

In this work, we develop a critical nodes compensation algorithm in order to prevent network from partitioning, thereby insure the network connectivity and throughput. Finally, we give simulation design and analysis of the critical nodes compensation algorithm using NS-2.33 model. Our simulation results show that the algorithm can effectively improve ad hoc networks performance. In addition, some time-critical applications may not be able to function properly in disconnected MANET as the end-to-end delay. However, there are lacks of full proof in theory and in practice. In these systems, nodes are subject to battery power, efficient bandwidth constraints and good throughput. The overall performance of the network is increased after detection of the critical node and selecting alternate path for the continuation of the same operation. By considering all the above metrics we come to a conclusion that while a network is set up, each node is leveled with a threshold value of energy. The node's energy is decreasing gradually as it is used in network connectivity. So the failure of critical node depends on less bandwidth, less residual battery power and poor throughput.

REFERENCES

[1] Shoudong Zou, Ioanis Nikolaidis and Janelle J. Harms "ENCAST: Energy–Critical Node Aware Spanning Tree for Sensor Networks", In Proceedings of the Communication Networks and Services Research Conference, pp.249-254, 2005.

[2] Michael Dion "Building in Reliability (BIR) with Critical Nodes", http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00493599

[3] Massimo Franceschetti and Ronald Meester, "Critical Node Lifetimes in Random Networks via the Chen–Stein Method", IEEE Transactions on Information Theory, vol.52, no.6, pp.2831-2837, June 2006.

[4] A. Karygiannis, E. Antonakakis, and A. Apostolopoulos, "Detecting Critical Nodes for MANET Intrusion Detection Systems", In Proceedings of IEEE Workshop on Security, Privacy and Trust in Pervasive and Ubiquitous Computing, pp.7-15, 2006.

[5] Vahid Tabatabaee and Leandros Tassiulas, "MNCM: A Critical Node Matching Approach to Scheduling for Input Buffered Switches with no Speedup", IEEE/ACM Transactions on Networking, Vol.17, no.1, pp.294-304, February 2009.

[6] Daisuke Kasamatsu, Norihiko Shinomiya and Tadashi Ohta, "A Broadcasting Method considering Battery Lifetime and Distance between Nodes in MANET", IEICE Transactions on Information and Systems, Vol. J91-B, No.4, pp.364-372, 2008

[7] Min Sheng, Jiandong Li and Yan Shi, "Critical Nodes Detection in Mobile Ad Hoc Network, ieeexplore.ieee.org/iel5/10777/33944/01620401.pdf.

[8] N.Komnios, D.Vergados and C. Douligeries. "Detecting unauthorized and compromised nodes in mobile adhoc network." Elseviewer Adhoc network,vol5,n0 3,pp.289-298,2007.

[9] Christian Bravo, Sonia A¨ıssa and Andr´e Girard, "Providing Quality of Service for Critical Nodes in Ad Hoc Networks", IEEE Vehicular Technology Conference, 2005.

[10] M.K Rafsanjani, A Movaghar, "Identifying monitoring nodes with selection of Authorized nodes in mobile Adhoc network",World Applied Sciences Journal,vol4,n03, pp.444-449,2008

[11] N.Komnios, D.Vergados and C. Douligeries, "Detecting Unauthorized and Compromised Nodes in Mobile Ad hoc Network," Adhoc Network(Elsevier), vol.5, no.3, pp.289-298, 2000.

[12] Yongguang Zhang, Wenke Lee, and Yi-An Huang, "Intrusion Detection Techniques for Mobile Wireless Networks", ACM Wireless Networks, vol.9, no.5, pp.545-556, September 2003.

Secret Key Agreement Over Multipath Channels Exploiting a Variable-Directional Antenna

Valery Korzhik, Viktor Yakovlev,Yuri Kovajkin
State University of Telecommunication
St. Petersburg, Russia

Guillermo Morales-Luna
Computer Science Department
CINVESTAV-IPN
Mexico City, Mexico

***Abstract*—We develop an approach of key distribution protocol(KDP) proposed recently by T.Aono et al., where the security of KDP is only partly estimated in terms of eavesdropper's key bit errors. Instead we calculate the Shannon's information leaking to a wire tapper and we also apply the privacy amplification procedure from the side of the legal users. A more general mathematical model based on the use of Variable-Directional Antenna (VDA) under the condition of multipath wave propagation is proposed. The new method can effectively be used even in noiseless interception channels providing thus a widened area with respect to practical applications. Statistical characteristics of the VDA are investigated by simulation, allowing to specify the model parameters. We prove that the proposed KDP provides both security and reliability of the shared keys even for very short distances between legal users and eavesdroppers. Antenna diversity is proposed as a mean to enhance the KDP security. In order to provide a better performance evaluation of the KDP, it is investigated the use of error correcting codes.**

Keywords-wireless communication; wave propagation; cryptography; key distribution.

I. INTRODUCTION

The problem of key distribution is still in focus of research activity especially for wireless LAN systems. This is due to the severe restriction of asymmetric (public key) cryptography WLAN implementation entailing a lower processing speed.

In order to solve this problem, quantum cryptography [1] which allows eavesdropping detection within the key sharing procedure seems useful. However, this approach does not reach a practical level due to many technical problems, such as the requirement of special quantum devices. There are well known key distribution protocols (KDP) based on the presence of noise in both legal and illegal channels [2], [3], [4]. But even though the eavesdropper's channel is less noisy than the legal ones and the eavesdroppers is passive, it is necessary to have the knowledge of the eavesdropper's noisy power in order to guarantee a fixed level of key security. Unfortunately this condition cannot be taken for granted because an eavesdropper may be able to get some advantage at the cost of better receiver sensitivity or a shorter distance of interception that it was considered by legal parties in the design of the secure KDP.

The most basic assumption on the executed KDP is that the legal and illegal users have different locations, and this fact has to be verified by physical means. (For that matter, an existing special zone surrounding each legal user shall be assumed where the presence of an eavesdropper is not allowed.)

This assumption is sufficient for secure key distribution if either the communication channel between the legal users and the eavesdropper have random parameters or one legal user generates some randomness, under the condition that this randomness is transmitted to other legal users over multipath channels and any eavesdropper is able to receive this information only on a multipath channel, but with some other parameters, due to different locations of the legal users and the eavesdropper.

The first approach is considered in [5], [6] for multipath channels with random parameters and in [7], [8] for ultra-wide band channels with random pulse responses. But the randomness exploiting of the fluctuation of channel parameters is very questionable because there may be such channel states in which a temporal variation of propagation characteristics is slow and small. In order to take for granted some given randomness level it would be better to create artificially this randomness by means of legal users.

Let us consider the following mathematical model of the channels between a source of randomness (the first legal user) and both the second legal user and the eavesdropper:

$\eta = \sum_{i=1}^{m} x_i \xi_i, \quad \zeta = \sum_{i=1}^{m} y_i \xi_i,$ where $\xi = (\xi_i)_{i=1}^{m}$ is the vector randomness, $x = (x_i)_{i=1}^{m}$ is the coefficient vector of the multipath propagation to the second legal user, and $y = (y_i)_{i=1}^{m}$ is the coefficient vector of multipath propagation to the eavesdropper. Let us assume for simplicity $E(\xi)=0$, then the following relation for the correlation coefficient between η and ζ results:

$$\rho(\eta,\zeta) = \frac{x^T R_\xi y}{\sqrt{(x^T R_\xi x)(y^T R_\xi y)}},$$

where R_ξ is correlation matrix of the random vector ξ. In a general case $\rho(\eta,\zeta) \le 1$. Moreover if x and y are orthogonal, (e.g.$\langle x, y\rangle=0$) and $R_\xi=\mathrm{Id}_m$ is the $(m\times m)$-identity matrix, then $\rho(\eta, \zeta)=0$.

The key bits of the second legal user can be generated after multiple repetition of the random independent vector (ξ) and the binary quantization of the random values (η). The first legal user can form key bits in a similar manner after a signal reception from the second user over the same multipath channel with the same randomness. If the variables η and ζ are Gaussian and non-correlated, then the shared key is provided secure due to the statistical independence of the variables. (A more general situation with non-zero correlation is considered in Section II.)

Common randomness can result from fluctuation of the cannel characteristics due to communicating object motion. Such approach has been proposed in [9], [10], [11]. But it still entails another problem: it is easy to break the secret key under an environment with small fluctuation of the channel characteristics or in the case when the communicating objects are stopped. In order to overcome these defects, a more sophisticated method, using smart antenna excited randomly by electronic means [12], has been proposed (although their results were obtained experimentally without an estimation of information-theoretic security of the shared keys).

This direction has been developed in many papers, [13], [14], [15], [16] are among the most important. In [14] some additional interference signals are simultaneously transmitted from the auxiliary antenna at legitimate access point. The authors of [13] change slightly the KDP based on Variable Directional Antenna (VDA) at the cost of a special selection of appropriate RSSI values in order to improve the key distribution security. In [15], an experimental scheme with the execution of dipole antennas was introduced. Such criterion of KDP security as Information Mutual Anti-tapping Condition (IMac) has been proposed in [16] but it was not proved that the IMac correctly estimates the security of KDP.

It is worth to note that in all above mentioned papers, there has not been considered the use of Privacy Amplification (PA) of the raw key bit strings and the application of the Privacy Amplification Theorem in order to correctly estimate the amount of Shannon's information leaking to an eavesdropper, although it is a very common technique used in the execution of different KDP [2], [3], [4], [17], [18], [19].

Our contribution consists first of all in an application of PA to VDA-based KDP that allows restricting reasonably the values of the required correlations between samples of legal users and eavesdroppers which are used for key bit generation. But in order to come closer to this main problem we have to solve a number of particular problems. The first attempt of such approach has been presented in [20].

In Section II, we describe the conditions of the physical channel and we introduce an exact mathematical model of the KDP. The results of the VDA simulation are presented in Section III. Section IV contains an optimization of the KDP in order to provide both reliability and security. Finally we conclude the main results and present some open problems in Section V.

II. KDP Based On Multipatch Wave Propagation And Randomly Excited VDA

The scheme of the communication system corresponding to the KDP is presented in Fig. 1.

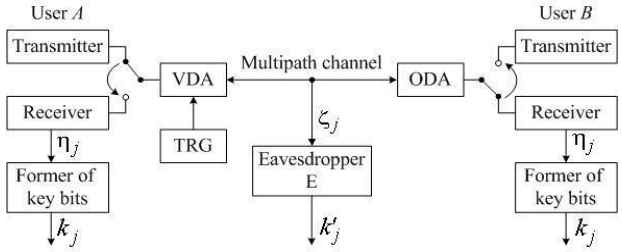


Figure 1. Scheme of communication system corresponding KDP

The KDP is described in the following steps:

1) The legal user A forms the random antenna diagram by exciting the VDA with output of *truly random generator* (TRG) and fix this diagram for some given time interval $[0,T_j]$ of the j-th key bit generation, $j=1,2,...,n$.

2) A transmits to B a harmonic signal $S_j(t) = \cos\omega_0 t$, $0 \le t \le T_j/2$, with the beam pattern obtained at step 1 over the multipath channel.

3) B receives a harmonic signal by an omni directional antenna (ODA) and forms the j-th key bit by comparing some functional η_j computed with the received signal on the time interval $[0, T_j/2]$ with a given threshold, forming the j-th key bit k_j.

4) The user B switches its ODA in a regime of radiation and transmits the same harmonic signal $S_j(t)=cos\omega_0 t$ within the time interval $T_j/2 \le t \le T_j$.

5) The user A switches its VDA to a receiver and processes the received signal in the same manner as B did, forming the j-th key bit k_j.

6) A and B repeat n times the steps 1-5 with new and independent outputs of TRG in order to create the desired number of key bits.

Thanks to the *Reciprocity Theorem* of radio wave propagation between uplink and downlink, the key sequences of A and B should be identical up to a random noise of receivers. Then the signal received by B at the time interval $T_j/2 \le t \le T_j$ can be expressed as:

$$y_j(t) = \sum_{i=1}^{m} \upsilon_{ij}\beta_{ij}\cos(\omega_0 t + \theta_{ij}), \qquad (1)$$

where with respect to the i-th ray at the j-th time interval, β_{ij} is the channel attenuation coefficient, υ_{ij} is the VDA amplitude gain, θ_{ij} is the phase shift, including both phases in antenna diagram and phase shift in i-th ray, and m is the number of paths (rays).

The signal received by E at time interval $T_j/2 \le t \le T_j$ is:

where the primed parameters have the same meaning as the

$$z_j(t) = \sum_{i=1}^{m} \upsilon'_{ij}\beta'_{ij}\cos(\omega_0 t + \theta'_{ij}), \qquad (2)$$

corresponding parameters in (1) but in possession of E. (We neglect initially the noise at the legal receivers, and we assume at this moment a noise absence at the eavesdropper E, in advantage with the illegal users.)

Later we will show that the probability distributions of the random values η_j and ζ_j, which are produced by executing some functionals from both $y_j(t)$ and $z_j(t)$ can have a good approximation by a zero mean Gaussian law.

It is easy to prove by a series of simple but tedious transforms that the probability of a bit disagreement between the j-th bit of the legal users and the eavesdropper key bits is:

$$p_e = 2\int_{-\infty}^{0} dy \int_{0}^{\infty} \frac{\exp\left(-\frac{x^2-2\rho xy+y^2}{2\sigma^2(1-\rho^2)}\right)}{2\pi\sigma^2\sqrt{1-\rho^2}} dx = \frac{1}{\pi}\arctan\left(\frac{\sqrt{1-\rho^2}}{\rho}\right) \qquad (3)$$

where ρ is the correlation coefficient between η_j and ζ_j, $\sigma^2=\mathrm{Var}(\eta_j)=\mathrm{Var}(\zeta_j)$.The dependence of p_e versus ρ is presented in Fig. 2. We can see that in contrast to our intuition, the probability $p_e = 0.1$ can be provided even when the correlation coefficient ρ has a significant value 0.95.

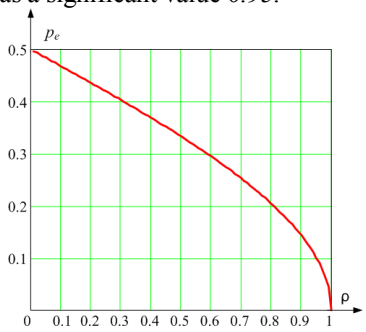


Figure 2. The probability of the key bit disagreement between legal and illegal users depending on the correlation coefficient ρ

In order to enhance the security of the legal user key string k shared after completion of the KDP it should be performed a privacy amplification [3], [17], [18], [19], or more specifically a mapping of the raw key string k to a shorter key string $\breve{k}$ of length l<n, using the so called hashing procedure $\breve{k}$ =h(k) taken from the universal class of hash functions [21]. Then the amount of Shannon's information leaking to E given her knowledge of the string k′ satisfies

$$I(\breve{\mathbf{k}};\mathbf{k}') \le \frac{l}{2^{n-l-t}\ln(2)} \qquad (4)$$

where $t = n+n\log_2(p^2_e+(1-p_e)^2)$ is the Renyi information under the assumption that the errors in the eavesdropper's key bits occur independently due to the independently generated VDA on each of the *j*-th time intervals. Hence in order to select the parameter *l* we should calculate the correlation coefficient ρ depending on the mutual location of the legal user and the eavesdropper, the properties of VDA and the characteristics of the multipath cannel. A solution for this problem will be presented in the next Section.

It is worth to remark that the quantized string k' has no redundancy and it is senseless to perform its soft decoding. As far as the use of a list decoding with the cipher text encrypted through the known key k, it looks as an completely intractable problem due to its large length (see Tables I and II at the end of Section IV below).

III. Correlation Between the Values η and ζ

Let us consider as VDA the so called ring antenna (RA) shown in Fig. 3 having N identical isotropic radiators excited by their random phases. Then the complex instant antenna diagram can be presented by the well-known formula [22]:

$$f(\varphi,\theta) = \sum_{i=1}^{m} \exp\left[ik_0 R\sin\left(\varphi - \frac{2\pi s}{N}\right) - i\psi_s\right] \qquad (5)$$

where ψ_s is a phase in the *s*-th radiator; $k_0=2\pi/\lambda$, λ is the length of the wave; *R* the radius of the RA; φ is the angle in the azimuthally plane; and θ is the angle in the vertical plane.

Both instant amplitude and the phase antenna diagrams can be obtained from (5) and they are random values providing random exciting to the RA. It would be possible to find different statistical characteristics of $f(\varphi,\theta)$ theoretically but it is rather more easy to solve the same problem by simulation. Since the current paper is limited in space, we present only the main conclusions based on the simulations for the case of independent and uniformly distributed phases ψ_s on $(0, 2\pi)$:

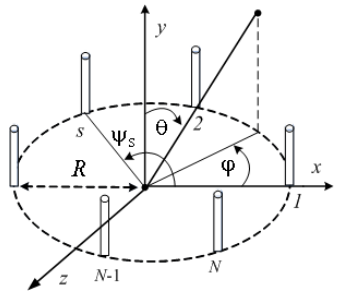


Figure 3. Ring antenna with N identical radiators

- the probability distribution of the amplitude antenna diagram has a good approximation through the Rice law which can be approximated in its turn by a Gaussian non-zero mean law;
- the probability distribution of the phase antenna diagram has a good approximation by an uniform law on the interval $(0,2\pi)$.

Next it is possible to compute theoretically the correlation coefficients between η_j and ζ_j for different functionals producing them and to find their probability distributions by simulation. However, it is necessary to specify the channel model and thereafter the functional description. To be more specific, let us consider a 3-ray channel model and a location of eavesdropper on the line connecting legal users (Fig. 4).

We select two functionals of $y_j(t)$ and $z_j(t)$ producing η_j and ζ_j respectively. Henceforth the functionals are compared with some thresholds in order to obtain the key bit k_j. The functionals are (see eq. (1)):

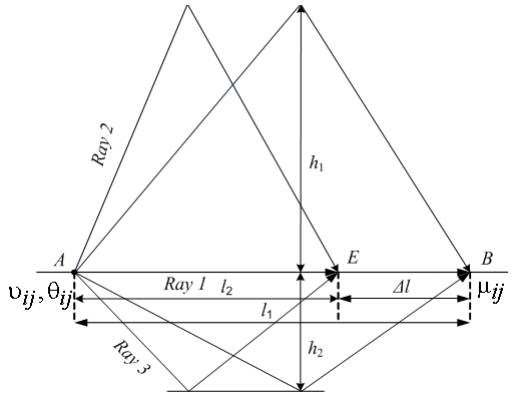


Figure 4. Channel model with 3-ray wave propagation

- envelope: $\mu_j=\sqrt{\mu_{cj}^2+\mu_{sj}^2}$,

where

$\mu_{cj}=\sum_{i=1}^{m} A_{ij}\cos\theta_{ij}$, $\mu_{sj}=\sum_{i=1}^{m} A_{ij}\sin\theta_{ij}$, $A_{ij}=\upsilon_{ij}\beta_{ij}$,

- phase difference

$$\Delta\psi_j = \Delta\psi_{j+1} - \Delta\psi_j = \arctan\frac{\mu_{s_{j+1}}}{\mu_{c_{j+1}}} - \arctan\frac{\mu_{s_j}}{\mu_{c_j}}.$$

In a similar manner, there can be presented the corresponding functionals for eavesdropper: μ'_j, μ'_{cj}, μ'_{sj}, $\Delta\psi'_j$.

We will be interested in a procedure to find the probability distributions of all functionals and correlations between similar functionals of any legal user *B* and the eavesdropper *E*. Because it is very hard to compute these values theoretically, we will find them by simulation for some given channel parameters.

Let us take distance between *AB* l_1=25m; distances to the first and to the second reflecting surfaces, respectively h_1=3m, h_2=3m, *N*=6, λ=12.5cm, *R*=λ/2 (see Fig's. 3 and 4). Assume that *E* is placed between legal users *A* and *B* within the interval Δl=3-22m. The dependences of the correlation coefficients $r_{\mu,\mu'}$ and $r_{\Delta\psi,\Delta\psi'}$ versus distance Δl between the eavesdropper *E* and the legal user *B* are shown in Fig. 5(a) and Fig. 5(b) respectively.

Similar dependences versus distance l_1 between legal users *A* and *B* where Δl=4m, h_1=h_2=3m are presented in Fig.6 (a) and 6(b). In Fig 7 (a) and 7(b) are shown the same dependences but versus distances to the first reflecting surface and for other parameters: l_1=25m, Δl=4m, h_2=3m.

As we can see from these figures, these dependences are looking very strange because if, for the thing, in some points on the line connecting *A* and *B* the correlation of amplitude is small enough, nevertheless a small shift of the eavesdropper location with respect to the locations of legal users results in strong correlation. Similar property holds also for the correlation of phase differences but the absolute values of this correlation are at most 0.8 for any conditions. Since the correlation between the values $\Delta\psi_j$ and $\Delta\psi'_j$ occurs less than the correlation between μ_j and μ'_j (see Fig. 5, 6, 7), it is reasonable to select the phase difference functional in order to form η_j and compare it with zero threshold for the k_j key- bit generation. (In order to coincide phases of support generators at users *A* and *B*, it is possible to transmit a special pilot signal and to tune phases of both users at the initial stage of KDP.)

In Fig. 8 there are presented empirical probability distributions for these functionals. It is evident that both cases can be approximated by appropriated Gaussian distributions (see solid curves). Therefore the relation (3) can be used to find the probability of disagreement between the key bits of the legal users and the eavesdropper. But before we address to eq. (4) in order to calculate security of KDP, it should be taken into account an opportunity for the presence of noise at the receivers of the legal users.

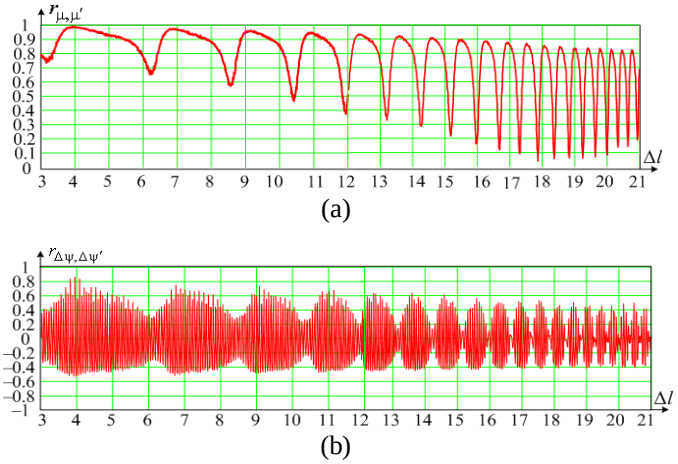


Figure 5. The dependence of correlation coefficients versus distances between legal use and eavesdropper.
a) for envelope, b) for phase difference

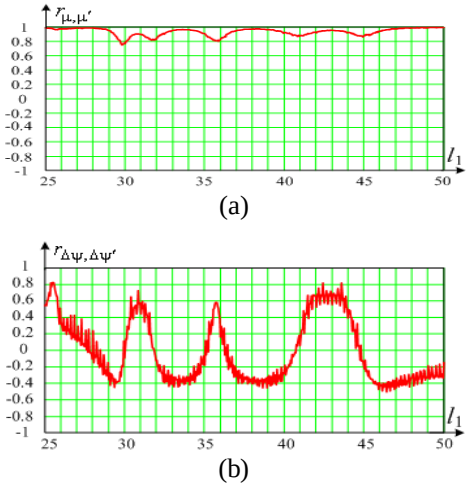
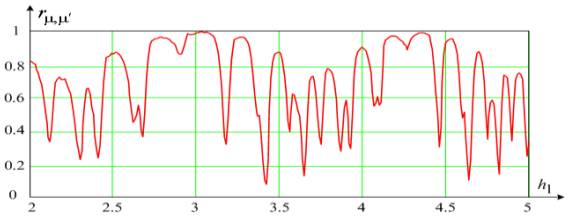
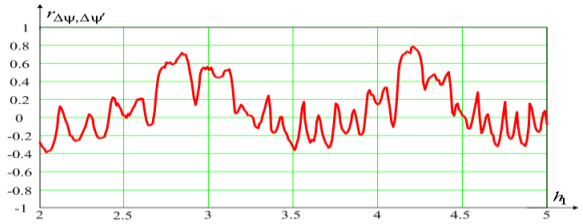


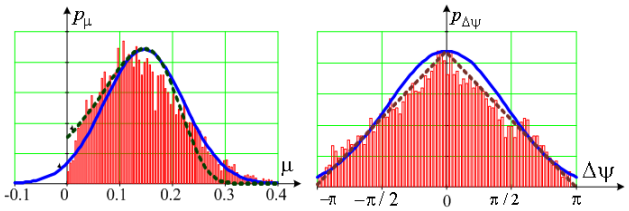
Figure 6. The dependence of correlation coefficients versus distances l1 between legal users for Δl=4m, h1=h2=3m.
a) for envelope, b) for phase difference





(b)

Figure 7. The dependence of correlation coefficients versus distances to the first reflecting surface and parameters: l1=25m, Δl=4m, h2=3m. a) for envelope, b) for phase difference



(a) envelope (b) phase difference

Figure 8. Empirical probability distribution for chosen functionals

IV. KDP Optimization Under Noisy Legal Channel

From now on we remove our previous assumption that the multipath channel among legal users *A* and *B* is noiseless but keep such condition for eavesdropper's channel. (Obviously, the last assumption cannot degrade the security of KDP.)

In this setting it is necessary to use some methods in order to correct disagreements in key bits of legal users. It is very reasonable to use firstly a selection of the most reliable key bits with a public discussion over a noiseless channel between legal users, and then to apply *forward error correction codes* (FEC) by sending of the check bits over the same but noiseless channel. (It is worth to note that a noiseless public channel among legal users can be arranged by the choice of special regime, namely large signal power or omni directional antenna of the user A that we were unable to use for the execution of KDP.)

The first method of the most reliable key bit selection is to take the decision according to the rule:

$$k_j = \begin{cases} 1, \text{ if } \eta_j \geq \alpha, \\ 0, \text{ if } \eta_j \leq -\alpha, \\ \text{erase otherwise,} \end{cases}$$

where η_j is the output of $\Delta\psi_j$, and α a threshold.

After a completion of the KDP including a production of the erased bits for both legal users it is necessary to mutually announce the numbers of these bits over public noiseless channels. In this case, the probability of a key bit disagreement between legal users and eavesdropper, given by (3), has to be corrected because an eavesdropper is able to intercept information about the numbers of accepted key bits over the public channel. We will take into account this fact later for the simulation procedure. The second method is to keep only the most reliable key bits, say *M*, and to remove the others. This means that the legal users form variation series of the values $|\eta_j|$ on a decreasing order and next to keep (after mutual public discussion) the first *M* members of this series to generate the key bits. Of course in this case the probability of key bit disagreement p_e is changed also against (3).

Let us denote by p1 and p2 the probability of legal key bit errors after the first and the second method, respectively. Next we use an error-correcting code (n_0+r, n_0) sending a sequence of *r* check symbols over public noiseless channel in order to correct eventually errors in the key sequence.

Then the probability of erroneous decoding P_{ed} by the modified Gallager's theorem is [19]: $P_{ed} \leq 2^{-n_0 E(R_C)}$, where

$$E(R_C) = \max_{\rho_0 \in (0,1)} \left[E_0(\rho_0) - \frac{\rho_0 (2R_C - 1)}{R_C} \right].$$

$$E_0(\rho_0) = \rho_0 - (1+\rho_0)\log_2 \left[p^{\frac{1}{1+\rho_0}} + (1-p)^{\frac{1}{1+\rho_0}} \right],$$

$R_C = n_0/(n_0+r)$, and no is the number of bits k_j which have been kept by legal users after erasing the unreliable bits following the first or the second procedures, and *p* is the error probability for the kept bits. In the case of check symbol sending, the Privacy Amplification Theorem against (4) becomes [19]:

$$I(\breve{\mathbf{k}};\mathbf{k}') \leq \frac{l}{2^{n_0 - l - t - r} \ln(2)}.$$

KDP optimization problem is to get the maximum key rate

$$R_C = \frac{l}{n_0 + n_{er}} = \frac{l}{n},$$

where n_{er} is the number of erased symbols after the use of the method 1 or 2 and given the values $I(\breve{\mathbf{k}};\mathbf{k}')$, P_{ed}, *l*, and different signal-to-noise ratio (*S*/*N*) at the receivers of the legal users. We solve this problem by simulation for the case of Gaussian noise at the legal receivers.

In Tables I and II there are presented the results of such optimization for typical conditions for the first and the second method of unreliable bits removal, respectively, where P_{er} is the probability of key bit erasing.

We can see from these tables that the second method is for large correlation a little bit better than the first one. However both methods provide sufficiently reliable and secure key sharing if eavesdropper is placed on 3-21m away from legal user *B* and phase difference is used as key generating method (see Fig. 5(b)).

A similar conclusion is drawn also for multipath channels with a greater number of rays and with other reasonable parameters and eavesdropper locations. In order to enhance the security of the KDP, antenna diversity can be used when *B* has m omni directional antennas and he selects randomly one of them at each time period T_j to receive and transmit signal. Then the relation finding the Renyi information used in (4) changes for:

$$t = n + \frac{n}{m}\log_2(\tilde{p}_e^2 + (1-\tilde{p}_e)^2). \qquad (5)$$

TABLE 1. KEY RATE MAXIMIZATION FOR THE FIRST METHOD GIVEN $I(\check{\mathbf{K}};\mathbf{K}')=10^{-9}$, $P_{ED}=10^{-5}$, S/N=10 AND DEFFERENT P

ρ	αopt	pe	Per	p1	l	n0	Rk
0.8	0.18	0.263	0.194	0.0087	128	539	0.243
			0.191	0.0083	256	940	0.272
			0.189	0.0082	512	1685	0.303
0.95	0.14	0.152	0.189	0.0083	128	1528	0.084
			0.188	0.0082	256	2484	0.103
			0.187	0.0082	512	4195	0.122
0.99	0.11	0.051	0.183	0.0078	128	7405	0.017
			0.181	0.0075	256	10977	0.023
			0.18	0.0075	512	15234	0.033

TABLE 2. KEY RATE MAXIMIZATION FOR THE SECOND METHOD GIVEN $I(\check{\mathbf{K}};\mathbf{K}')=10^{-9}$, $P_{ED}=10^{-5}$, S/N=10 AND DEFFERENT P

ρ	Mopt	pe	Per	p2	l	Rk
0.8	539	0.222	0.24	0.0075	128	0.245
	940		0.238		256	0.277
	1685		0.236		512	0.306
0.95	1528	0.115	0.236	0.0072	128	0.095
	2484		0.235		256	0.105
	4195		0.233		512	0.123
0.99	7405	0.049	0.23	0.0069	128	0.017
	10977		0.29		256	0.023
	15234		0.29		512	0.033

The relation (6) holds with the probability equal to the probability of the event in which with at least of one of antennas mutual location of the legal user and the eavesdropper is got such that $\rho \le \rho^*$, where ρ^* is found by (3) given β_0.

We considered so far a scenario when an eavesdropper uses the same omni directional antenna as the legal user *B*. But *E* can execute directional antenna to separate all rays and to process the best of them or even apply joint processing to all of them. We have performed a simulation of the case with single ray separation and it has been shown that the correlation coefficient even decreases in comparison with one presented before. The case of joint processing of separated rays is noteworthy. But we can remark that even under the very strong condition in which the eavesdropper knows exactly all channel parameters both for *E* and *B*, there is still uncertainty about VDA gains in the direction of *E* and *B*. Therefore, generally speaking, the correlation coefficient occurs even in this case with a value less than one.

V. CONCLUSION AND FUTURE WORK

We considered a method of key sharing based on the concept of a VDA under the condition of multipath channel and we showed that sufficient security and reliability of the shared keys can be provided even when the eavesdropper's channel is noiseless. (It is worth to remark that in order to get such result, the following two conditions are necessary: to create truly randomness with the help of a VDA and to have multipath channels.) The results of investigations show that the security of the KDP (in terms of Shannon's information leaking to eavesdropper) does not depend only on the distance between legal users and eavesdropper but also on the eavesdropper's location. This result somewhat contradicts to a very optimistic conclusion in [12].

We propose to use the difference-phase functional instead of either quadrature components or envelope in order to form key bits. This approach results in less mutual correlation between legal user and eavesdropper and simplifies a choice of threshold. The key sequence k is i.i.d if VDA is excited by independent random phases and threshold is chosen in an appropriate manner. (This fact has been confirmed by simulation using statistical tests.) Our contribution consists also in the proof of relation (3) which allows to connect the probability of disagreement between the key bits of legal users and eavesdropper with the correlation of corresponding values. Unfortunately, a limited space of the paper does not allow us to show all simulation results for different multipath channels and mutual location of legal users and eavesdroppers, which we have got at our disposition. It is pertinent to note that although some results were obtained by the use of computer simulation, it does not lead to a loss of generality because this is only an approach to reach the same goal by a simpler way. As far as the limitation due to the parameter selection (number of rays, position of eavesdropper, *S*/*N*, etc) it can be explained only by the space paper limitations. Indeed, following our theory, one can get the results corresponding to arbitrary parameters.

In the future we are going to investigate: *i*) the use of multitone signals in the KDP, ii) the optimal processing of the eavesdropper rays separation in order to provide the greatest correlation, iii) the use of real FEC and effective decoding algorithms with KDP (instead of extended Gallager's bounds);and, iv) the use of other types of VDA (like ESPAR or others).

REFERENCES

[1] C. H. Bennett and G. Brassard, "Quantum cryptography: Public key distribution and coin tossing," in Proceedings of International Conference on Computers, Systems and Signal Processing, December 1984.

[2] U. Maurer, "Protocols for secret key agreement by public discussion based on common information." in CRYPTO, ser. Lecture Notes in Computer Science, E. F. Brickell, Ed., vol. 740. Springer, 1992, pp.461-470.

[3] U. Maurer and S. Wolf, "Secret-key agreement over unauthenticated public channels II: the simulatability condition." IEEE Transactions onInformation Theory, vol. 49, no. 4, pp. 832-838, 2003.

[4] V. Yakovlev, V. Korzhik, and G. Morales-Luna, "Key distribution protocols based on noisy channels in presence of an active adversary: Conventional and new versions with parameter optimization," IEEE Transactions on Information Theory, vol. 54, no. 6, pp. 2535-2549, 2008.

[5] A. M. Sayeed and A. Perrig, "Secure wireless communications: Secret keys through multipath," in ICASSP. IEEE, 2008, pp. 3013-3016.

[6] Y. Liu, S. C. Draper, and A. M. Sayeed, "Secret key generation through ofdm multipath channel," in CISS. IEEE, 2011, pp. 1-6.

[7] M. G. Madiseh, M. L. McGuire, S. S. Neville, L. Cai, and M. Horie, "Secret key generation and agreement in uwbcommuniction channels," in GLOBECOM 2008.IEEE, 2008.

[8] M. G. Madiseh, S. W. Neville, and M. L. McGuire, "Time correlation analysis of secret key generation via uwb channels," in GLOBECOM. IEEE, 2010, pp. 1-6.

[9] A. A. Hassan, W. E. Stark, J. E. Hershey, and S. Chennakeshu, "Cryptographic key agreement for mobile radio," Digital Signal Processing, vol. 6, pp. 207-212, 1996.

[10] J. Hershey, A. Hassan, and R. Yarlagadda, "Unconventional cryptographic keying variable management," Communications, IEEE Transactions on, vol. 43, no. 1, pp. 3 -6, Jan. 1995.

[11] Ch.Ye, S. Mathur, AReznik, W. Trappe and N.B. Mandayam. "Information-Theoretically Secrete Key Generation for Fading Wireless Channels, IEEE Transactions on Information Forensics and Security , vol.5, No.2,June 2010,pp.240-254.

[12] T. Aono, K. Higuchi, T. Ohira, B. Komiyam, and H. Sasaoka, "Wireless secret key generation exploiting reactance-domain scalar response of multipath fading channels," IEEE Trans. Antennas &ropagation,vol. 53, pp. 3776-3784, Nov. 2005.

[13] T. Shimizu, H. Iwai, and H. Sasaoka, "Improvement of key agreement scheme using espar antenna," in Proc. 2008 Int. Symp. Antennas Propag.(ISAP2008). Taipei, Taiwan, 2008, pp. 1-4.

[14] M. Onishi, T. Kitano, H. Iwai, and H. Sasaoka, "Improvement of tolerance for eavesdropping in wireless key agreement scheme using espar antenna based on interference transmission," in The 2009 International Symposium on Antennas and Propagation (ISAP 2009).Bangkok, Thailand: ISAP, October 20-23 2009.

[15] T. Shimizu, N. Otani, T. Kitano, H. Iwai, and H. Sasaoka, "Experimental validation of wireless secret key agreement using array antennas," in XXX URSI General Assembly and Scientific Symposium. Istambul, Turkey: URSI, August 11-20 2011.

[16] T. Yoshida, T. Saito, K. Fujiki, K. Uematsu, and H. U. T. Ohira, "Impact of direct-path wave on imac in secret key agreement system using espar antennas," in XXX URSI General Assembly and Scientific Symposium. Istambul, Turkey: URSI, August 11-20 2011.

[17] C. H. Bennett, G. Brassard, C. Crepeau, and U. M. Maurer, "Generalized privacy amplification," IEEE Transactions on Information Theory, vol. 41, no. 6, pp. 1915-1923, 1995.

[18] U. Maurer and S. Wolf, "Privacy amplification secure against active adversaries," Lecture Notes in Computer Science, vol. 1294, pp. 307-321,1997.

[19] V. Korjik, G. Morales-Luna, and V. Balakirsky, "Privacy amplification theorem for noisy main channel," Lecture Notes in Computer Science, vol. 2200, pp. 18-26, 2001.

[20] V.Korzhik V. Yakovlev,,G. Morales-Luna, Yu. Kovajkin. "Wireless Secrete Key Sharing Based on on the Use of Variable-directional Antenna over Multipath Channels", in Proc.ELMAR' 2010 ,pp.277-280.

[21] L. Carter and M. N. Wegman, "Universal classes of hash functions," J. Comput. Syst. Sci., vol. 18, no. 2, pp. 143-154, 1979.

[22] R. E. Collin and F. J. Zucker, Antenna Theory - Part 1.McGraw-Hill,1969.

The Relationships of Soft Systems Methodology (SSM), Business Process Modeling and e-Government

Dana Indra Sensuse
Faculty of Computer Science
University of Indonesia
Depok, Indonesia

Arief Ramadhan
Faculty of Computer Science
University of Indonesia
Depok, Indonesia

***Abstract*—e-Government have emerged in several countries. Because of many aspects that must be considered, and because of there are exist some soft components in e-Government, then the Soft Systems Methodology (SSM) can be considered to use in e-Government systems development process. On the other hand, business process modeling is essential in many fields nowadays, as well as in e-Government. Some researchers have used SSM in e-Government. Several studies that relate the business processes modeling with e-Government have been conducted. This paper tries to reveal the relationship between SSM and business process modeling. Moreover, this paper also tries to explain how business process modeling is integrated within SSM, and further link that integration to the e-Government.**

Keywords- Soft Systems Methodology; Business Process Modeling; e-Government.

I. INTRODUCTION

e-Government is the use of Information Technology (IT) by public sector organizations [1]. The main orientation of e-Government is the accessibility of information by the public, rather than financial income, as implied in [1].

Because the target of e-Government is the public sector, then the e-Government systems are generally built based on the web technology. This technology is used because it has ability to reach people quickly and widely. Therefore, the e-Government can be associated as an attempt to put the public administration online. This means, e-Government is not simply replace all the equipment in the public sector, from a typewriter to a computer. e-Government is more than that. e-Government move the whole system along with existing business processes within the public sector to the online world.

Because of its relation with IT, then most people thought that e-Government is part of computer science. However, e-Government has become an emergent multidisciplinary field of research [2]. In addition to computer science, there are many other scientific fields in e-Government, for example public administration, management, politics, socio culture etc.

It is revealed in [2], that although theoretical ground is still under construction, e-Government certainly qualifies as a legitimate emerging scientific discipline. It also revealed in [2], that as technological innovations are continuously hitting the market, the frontiers of the e-government discipline are moving and its multidisciplinary nature confirmed [2].

Currently the development of e-Government systems have proliferated in several countries, both in developing countries and developed countries. It is stated in [3], that e-government is a useful tool for modernizing the state given that it enables government to offer higher-quality services to citizens and provide those services in a more efficient, effective, and transparent way.

Heeks in [1] says that e-Government is also an information system, but it is enriched with various aspects, such as the management aspects, political aspects, economical aspects and others. These aspects have to be considered by developers when developing an e-Government system.

Because of many aspects that must be considered, then the e-Government system development process can be very complex. These aspects cannot be observed separately, but should be observed as a whole, where there is interaction in it. Such characteristics can be solved using systems thinking.

e-Government is a socio-technical system that consists of soft components and hard components [1]. It could be argued that the soft component is the people who are involved in e-Government, whereas the hard component is the Information Technology (IT) that being used. The management approach of the soft component is likely inspired by social sciences, it tends to be subjective, qualitative, and further highlight by the aspects of humanism [1]. The management approach of the hard component is inspired by engineering science, it tends to be objective, quantitative and further highlight by the technical aspects [1].

It is implied in [1], that the most critical factor to determine whether an e-Government system development fail or not is the soft component. Because of the soft component is very dominating and is tend to be subjective, then we see that one type of systems thinking, that is soft systems thinking, can be used in e-Government system development.

Soft systems thinking do not assume that the world is systemic and well-ordered; on the contrary, it assumes social

reality to be "problematical", characterized by multiple angles of approaches and perspectives [4]. The understanding of reality is dependent upon the observer, his interpretations and what he chooses to focus on [4].

Some of the methodologies that can be used in soft systems thinking is a meta-synthesis approach used in [5], and Soft Systems Methodology (SSM). In this paper, we choose SSM and try to find the relationships between SSM and others.

II. SOFT SYSTEMS METHODOLOGY

Soft Systems Methodology (SSM) was proposed in 1981 by Peter Checkland [6]. As the name implies, SSM is based on soft systems thinking. The picture of SSM can be seen in Fig. 1. SSM consists of seven steps, i.e. (extracted from [6]):

1) The identification of a problem situation that demands attention

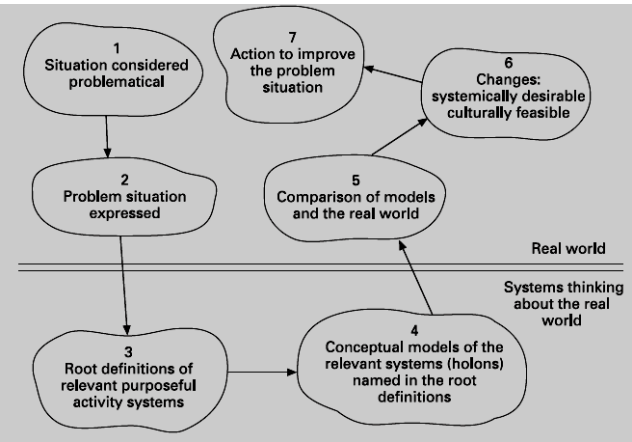


Figure 1. The seven-step of Soft System Methodology (SSM) [6].

2) Problem situation is expressed. The expression can be described using the Rich Picture Diagram. The examples of the Rich Picture Diagram can be seen in Fig. 2.

3) Some relevant human activity systems, potentially offering insight into the problem situation, are selected and from these 'root definitions' are built. In this step, CATWOE analysis is performed. CATWOE stands for Customers, Actors, Transformation process, World view, Owner, and Environmental constraints.

4) Construct conceptual models. This is the most important step in the SSM. Various modes of modeling techniques can be applied at this step.

5) Comparing the conceptual model with the real world. The aim is to provide material for debate about possible change among those interested in the problem situation. This step shows the social processes within the SSM.

6) Making changes to the model by accommodating the interests of several actors involved. Changes should be able to follow the desired model but still possible (feasible) historically, culturally and politically. Changes may include changes in attitudes, structures, or procedures.

7) Perform various activities to implement the model and fix the problem. In this step, the conclusions are drawn and long-term solution is formulated.

SSM has been amended several times. The first change is made in 1990 in the form of "two-strands model" [6]. In this model, were added three types of inquiry, referred to as Analysis 1, 2 and 3 [6]. Analysis 1 considers the intervention itself and the roles of client, problem-solver and problem-owners. Analysis 2 is social system analysis. Analysis 3 examines the politics of the problem situation and how power is obtained and used [6].

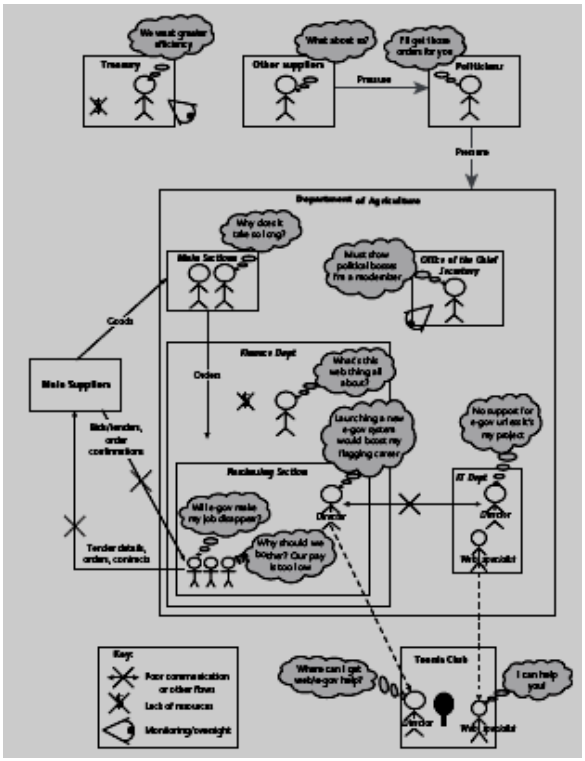


Figure 2. Rich Picture diagram for the system of procurement in the Department of Agriculture [1].

Subsequent change of SSM is made when the original seven-step is merged into just four steps. The new four-step is named as "learning cycle of SSM". Four new steps are [7]:

1) Finding out about a problem situation, including culturally/politically.

2) Formulating some relevant purposeful activity models

3) Debating the situation, using the models, seeking from that debate both :

a) changes which would improve the situation and are regarded as both desirable and (culturally) feasible

b) the accommodations between conflicting interests which will enable action-to-improve to be taken

4) Taking action in the situation to bring about improvement.

Although the SSM has been amended several times and although Checkland no longer favours it, the representation of SSM as a seven-step, which appeared in 1981, is still frequently used today [6]. Some researchers have used SSM in e-Government, for example can be seen in [24].

III. Business Process Modeling And E-Government

Business process is characterized by three key words, i.e. activities, linked, and objective. It can be seen from some of the definitions of business process. In [8], it is said that business process is a set of one or more linked procedures or activities which collectively realize a business objective or policy goal, normally within the context of an organizational structure defining functional roles and relationships. Other definition about business process is a set of coordinated tasks and activities, involving both human and system interactions, that will lead to accomplishing a set of specific organizational goals [9]. In [10], it is stated that a business process is a set of related activities or operations which, together, create value and assist organizations to achieve their strategic objectives.

A business process has a clear beginning and end, creating outputs by adding value to inputs [10]. It seems that business process is more likely a function. However, in [11], it was stated that a business process is not the same as a function [11]. It was said in [11], that the people and operations that include in a single business process may come from more than one traditional functional group.

The first stage of the analysis of a business process is concerned with constructing a model of the business process [11]. This "constructing" activity is commonly called business process modeling.

Business process modeling became popular in the context of enterprise reorganization and modernization in the early 1990's [12]. Business process modeling is the visual representation of business processes [13]. Visual representation is usually done in the form of pictures or notations with specific meanings.

Guizzardi et. al. in [14] stated that business process modeling is about the description of sequence of business activities carried out in organizations in order to make them explicit. It is implied in [15], that business process modeling is done for better understanding and analysis about the business process.

It is indicated in [16], that business process modeling can be used to communicate a wide variety of information to a wide variety of audiences. Some audiences who become the main focus are the stakeholders. The graphical nature of business process models can be used as a medium of communication between stakeholders (e.g., executives, developers, and employees) [15]. Business process modeling capabilities as a medium of communication is also revealed in [17].

Currently, there exists some software that can be used to manage the business processes, one of which is SAP [29]. Bider in [30] stated that there are four common views on the process development, i.e.: input/output flow, workflow, agent-related workflow, and state workflow.

Business process modeling is essential in many fields nowadays, and much research and many initiatives have been proposed in order to facilitate and improve its development [18]. One example of business process modeling activities regarding shipment process of a hardware retailer can be seen in Fig. 3.

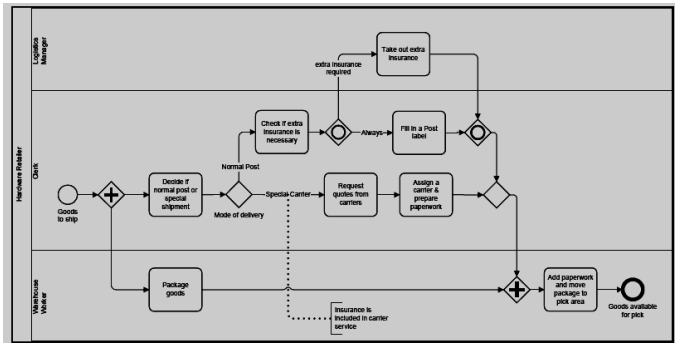


Figure 3. Shipment Process of a hardware retailer [16].

From its name, i.e. "Business Process", and from some of the above explanations, it is impressed that the "business process" is only needed and used within the business sector (private sector). However, there are some studies that link the business process modeling with e-Government (public sector). Several studies that relate the business processes modeling with e-Government can be seen in [25], [26] and [27].

IV. Soft Systems Methodology, Business Process Modeling And E-Government

In the section I and II it has been explained that there is a correlation between the SSM with e-Government, that is based on its characteristics, SSM can be used in the development of e-Government systems and that there has been many studies linking SSM with e-Government. Then, in section III it has been revealed that many studies has linked business process modeling with e-Government.

Within this section, we try to describe and propose how SSM can also be associated with business process modeling. In addition, at the same time, we also will reconcile these linkages with e-Government.

In [19], it is shown that the SSM can be associated with business process modeling. In [19], it is implied that the result of business process modeling is depicted using Unified Modeling Language (UML). However, in [19], the process of making business process model with the UML, is depicted as standing outside the SSM. In [19], that being compared with the rich picture diagram is still the conceptual model and not the business process model. It was further shown, that in [19], the process of making business process models with the UML, is conducted separately from the SSM.

In this paper, we propose that the process of making business process model can be fused within SSM. This integration process occurs by placing the process of making business process model into the step 4 of the SSM. So, the conceptual model that originally resulting from that step can be in the form of a business process model, which can be depicted using UML, Business Process Modeling Notation (BPMN) or

the other. Because of this integration, the result of step 1 (e.g. in the form of surveys results or interviews results), step 2 (e.g. in the form of rich picture diagram) and step 3 (e.g. CATWOE analysis) of the SSM, can be considered by business process modeler to create the business process model in step 4. Henceforth, in step 5, the business process model will be compared with the rich picture diagram of the real-world conditions. It can be seen that in our proposal the making of business process model is integrated within SSM, and not to put it separately. The reason for this integration proposal will be described in the explanation below.

In the conduct of business process modeling there are several obstacles that may be faced. In [28], it was said that defining a business process is a taxing, vexing, and iterative process. Business process modeling is often time-consuming and sometimes involves a certain amount of redundant work because of the similarity between modeling objects [20]. Business process modeling is a complicated process [21].

In addition to these constraints, the business process modeling is also ultimately depends on the modeler who is also a human being. As stated in [22] that the business process model is the result of mapping a business process, and that business process can be either a real-world business process as perceived by a modeler, or a business process conceptualized by a modeler. In this case it appears there are two key words which tend to be subjective and depends on the modeler's perspective, i.e. "to perceive" and "to conceptualize". Because of the subjective tendencies that are "soft" and the fact that business process modeling is very complicated, the SSM is really suitable to this case.

In [23], it is said that building a model of a real business process is a challenging task because:

- Business processes are not always clearly visible as they may go through the whole, often functionally structured organization.
- Written information about business processes is often non-existing or not reliable. The only practical way to obtain reliable information for creating a model of a real business process is by interviewing the people engaged in the process.

From the statement in [23] above, again it is shown that there is a very common thing carried out in SSM, i.e. the interview. In the SSM, the interview, usually carried out in the step 1. This again shows the relationship between SSM with business process modeling.

In addition, business process modeling involves an abstraction from the real-world business process [22]. It is highly compatible with SSM paradigm. Based on Fig. 1, it is shown that the SSM actually also tried to make abstraction of the real-world problems. SSM bring the real-world problems (upper half of the figure) into the conceptual model (bottom half of the figure). The fact that there are similarities between SSM paradigms with business process modeling further strengthen our proposal that the business process modeling can be integrated in the SSM.

Based on the above explanation and based on our explanations in section I, II and III, further, we can see that the SSM, business process modeling and e-Government can be mixed together. In other words, the making of business process model in the development of e-Government system can be done with the SSM.

V. CONCLUSIONS

This paper has explained how the business process modeling can be integrated into the SSM. Furthermore, this paper also has described the relationship between the SSM, business process modeling and e-Government. It can be concluded that the SSM can be used in the making of business process models in the development of e-Government system. Some further research can be conducted to enrich this paper, for example the research about how to do the real implementation of the integration of SSM and business process modeling in a country.

REFERENCES

[1] R. Heeks, Implementing and Managing eGovernment An International Text, London, England : SAGE Publications, 2006.

[2] S. Assar, I. Boughzala, and I. Boydens, "Back to Practice, a Decade of Research in E-Government", in Practical Studies in E-Government : Best Practices from Around the World, S. Assar, I. Boughzala, and I. Boydens, Eds. New York, USA: Springer, 2011.

[3] L. Herrera and J. R. Gil-Garcia, "Implementation of E-Government in Mexico: The Case of Infonavit", in Practical Studies in E-Government : Best Practices from Around the World, S. Assar, I. Boughzala, and I. Boydens, Eds. New York, USA: Springer, 2011.

[4] A. Mirijamdotter, "A Multi-Modal Systems Extension to Soft System Methodology," Ph. D. dissertation, Lulea Tekniska Universitet, Sweden, 1998.

[5] J. Gu and X. Tang, "Meta-Synthesis System Approach To Knowledge Science," International Journal of Information Technology & Decision Making, vol. 6, no. 3, pp. 559-572, 2007.

[6] M. C. Jackson, System Thinking Creative Holism for Managers, John Wiley & Sons Ltd, England, 2003.

[7] P. Checkland, "Soft Systems Methodology: A Thirty Year Retrospective," Systems Research and Behavioral Science, vol. 17, pp. S11-S58, 2000.

[8] [Workflow Management Coalition], "The Workflow Management Coalition Specification : Terminology & Glossary", Document Number WFMC-TC-1011, Document Status - Issue 3.0, 1999.

[9] [Oracle], "Oracle Application Integration Architecture: Business Process Modeling and Analysis", Oracle White Paper, April 2009.

[10] [Australian Government Information Management Office], "The Australian Government Business Process Interoperability Framework", July 2007.

[11] S. J. Childe, P. A. Smart, and A. M. Weaver, "The use of generic process models for process transformation," in Proceedings of the IFIP TC5 WG5.7 international workshop on Modelling techniques for business process re-engineering and benchmarking, pp. 51-60, 1997.

[12] R. von Ammon, C. Emmersberger, F. Springer, and C. Wolff, "Event-Driven Business Process Management and its Practical Application Taking the Example of DHL," in Proceedings of the 1st iCEP08 Workshop on Complex Event Processing for the Future Internet, 2008.

[13] M. Ramadan, H. G. Elmongui, and R. Hassan, "BPMN Formalisation using Coloured Petri Nets," in Proceedings of the 2nd GSTF Annual International Conference on Software Engineering & Applications (SEA 2011), 1997.

[14] R. S. S. Guizzardi, G. Guizzardi, J. P. A. Almeida, and E. Cardoso, "Ontological Foundations for Agent-Oriented Organizational Modeling," in Proceedings of the 3rd International i* Workshop - istar08, pp. 37 – 41, 2008.

[15] A. Lodhi, V. K¨oppen, and G. Saake, "Business Process Modeling: Active Research Areas and Challenges", Department of Technical and Business Information Systems, Faculty of Computer Science, Otto-von-Guericke University, Tech. Rep., Nr : FIN-001-2011, 2001.

[16] [Object Management Group], "Business Process Model and Notation (BPMN) Version 2.0", OMG Document Number : formal/2011-01-03, 2011.

[17] A. Popovic, M. I. Stemberger, and J. Jaklic, "Applicability of Process Maps for Simulation Modeling in Business Process Change Projects," Interdisciplinary Journal of Information, Knowledge, and Management, vol. 1, pp. 109-123, 2006.

[18] A. Koschmider, J. L. de la Vara, and J. Sanchez, "Measuring the Progress of Reference Model-Based Business Process Modeling," in Proceedings of 3rd International Conference on Business Process and Services Computing (BPSC 2010), pp. 218 – 229, 2010.

[19] M. Salahat and S. Wade, "Measuring the Progress of Reference Model-Based Business Process Modeling," in Proceedings of The 5th International Conference on Innovations in Information Technology, 2008.

[20] C. Ren, W. Wang, J. Dong, H. Ding, B. Shao, and Q. Wang, "Towards A Flexible Business Process Modeling And Simulation Environment," in Proceedings of the 2008 Winter Simulation Conference, IEEE Xplorer, pp. 1694-1701, 2008.

[21] R. Lu and S. Sadiq, "A Survey of Comparative Business Process Modeling Approaches," in Lecture Notes in Computer Science, Springer, Vol. 4439/2007, 82-94, 2007.

[22] J. Mendling, "Foundations of Business Process Modeling", in Handbook of Research on Modern Systems Analysis and Design Technologies and Applications, M. R. Syed and S. N. Syed, Eds. Hershey, USA: IGI Global, 2009.

[23] I. Bider, "State-Oriented Business Process Modeling: Principles, Theory and Practice," Ph. D. dissertation, Royal Institute of Technology and Stockholm University, Sweden, 2002.

[24] M. Alrazooqi and R. De Silva, "Mobile and Wireless Services and Technologies for M-Government Solution Proposal for Dubai Government," WSEAS Transactions on Information Science And Applications, Issue 8, Vol. 7, pp. 1037-1047, 2010.

[25] J. Becker, D. Pfeiffer, M. Rackers, and P. Fuchs, "Business Process Management in Public Administrations – The PICTRUE Approach," in Proceedings of 11th Pacific-Asia Conference on Information Systems, 2007.

[26] O. F. Aydınlı, S. Brinkkemper, and P. Ravesteyn, "Business Process Improvement in Organizational Design of E-Government Services", Department of Information and Computing Sciences Utrecht University, Tech. Rep. UU-CS-2007-041, 2007.

[27] P. Liegl et. al., "Modeling eGovernment processes with UMM," Informatica, Vol. 31, pp. 407–417, 2007.

[28] F. Nickols, "The Difficult Process of Identifying Processes : Why it isn't as easy as it sounds," Knowledge and Process Management, Vol. 5, No. 1, 1998.

[29] [www.winshuttle.com]. 2011. "How SAP Users hold the key to Business Process Improvement". http://www.winshuttle.com/White-Papers/Winshuttle-HowSAPUsersHoldtheKeytoBPI-whitepaper-EN.pdf

[30] I. Bider 2002. "Tutorial on: Business Process Modeling as a Method of Requirements Engineering". http://www.iceis.org/iceis2007/ Hall_Of_Fame/ ibider/ ibider_2002.pdf

AUTHORS PROFILE

Dana Indra Sensuse. B.Sc in Soil Science (Bogor Agricultural University, Indonesia, 1985), M.Sc in Library and Information Studies (Dalhousie University, Halifax, Canada, 1994), Ph.D in Information Studies (Toronto University, Canada, 2004), Lecturer at University of Indonesia, Head of e-Government Lab at University of Indonesia.

Arief Ramadhan. B.Sc in Computer Science (Bogor Agricultural University, Indonesia, 2005), M.Sc in Computer Science (Bogor Agricultural University, Indonesia, 2010), Ph.D Student in Computer Science (University of Indonesia), Research Assistant at University of Indonesia. Member of e-Government Lab at University of Indonesia.

Re-tooling Code Structure Based Analysis with Model-Driven Program Slicing for Software Maintenance

Oladipo Onaolapo Francisca (PhD)
Computer Science Department,
Nnamdi Azikiwe University
Awka, Nigeria

***Abstract*—Static code analysis is a methodology of detecting errors in program code based on the programmer's reviewing the code in areas within the program text where errors are likely to be found and since the process considers all syntactic program paths; there is the need for a model-based approach with slicing. This paper presented a model of high-level abstraction of code structure analysis for a large component based software system. The work leveraged on the most important advantage of static code structure analysis in re-tooling software maintenance for a developing economy. A program slicing technique was defined and deployed to partition the source text to manageable fragments to aid in the analysis and statecharts were deployed as visual formalism for viewing the dynamic slices. The resulting model was a high-tech static analysis process aimed at determining and confirming the expected behaviour of a software system using slices of the source text presented in the statecharts.**

Keywords- software maintenance; static analysis; syntactic program behavior; program slicing.

I. INTRODUCTION

There had been a growing use of static analysis both in commercial and academic areas in the verification of properties of software used and locating potentially vulnerable code in critical sensitive computer systems [1]. Static code analysis is the analysis of computer software that is performed without actually executing programs built from that software. This is in contrast to the analysis performed on executing programs which is known as dynamic analysis [2]. Empirical evidence from [3] and [4] showed that software maintenance consumed between 50% and 80% of the resources in the total software budget. In addition, according to [5] and [6]; an estimated 50% to 80% of the time and material involved in software development is devoted to maintenance of existing code, hence the main justification for this work was to leverage on the most important advantage of static code structure analysis in re-tooling software maintenance for a developing economy.

The most important advantage of static code analysis lied in the possibility of considerable cost saving by defects elimination in a program; this is expected to bring about some economic benefits to a developing country as technological innovations were known to do and they can be expected to save considerable costs in software maintenance. The earlier an error is determined; the lower the cost of its correction. According to [7], correction of an error at the testing stage is ten times more expensive than its correction at the construction (coding) stage.

This paper presented a model for code structure analysis for a large component based software system. A program slicing technique was deployed to partition the code to manageable fragments to aid in the analysis and statecharts were deployed as visual formalism to view the dynamism of the static slices. The model aimed at determining and confirming the expected behavior of a software system using the source text because research had shown that during maintenance, the most reliable and accurate description of the actual behavior of a software system is its source code [8]. The rest of this paper is organized as follows: A background to the concepts of static codes structure analysis and program slicing was presented in section II; section 3 described the materials and methods adopted in the research, the resultant models were discussed in section 4 and the section 5 concluded the paper.

II. RESEARCH BACKGROUND

Identifying errors in software during development is very important so that the end product can be error free and perform to its specification. Reference [9] believed that early identification of bugs in a developing program can be achieved through the concept of program slicing. Generally, program slice has a wider spectrum of applications that include debugging, testing, maintenance, code understanding, complexity measurement, security etc.

Static analysis is any form of analysis that does not require a system to be operated. The process complements dynamic analysis, where system operation is central. When applied to code, static analysis is typically referred to as white-box, glass-box, structural or implementation based techniques [10]. In most cases the analysis is performed on some version of the source code and in the other cases some form of the object code. The term is sometimes applied to the analysis performed by an automated tool, with human analysis being called program understanding, program comprehension or code

review. Static Analysis is performed without program execution and the process includes almost everything except conventional testing. This is because dynamic testing requires running code, some program properties such as race conditions are too hard to test for and though it may be impossible to program correctness, one can easily prove simple properties of simplified models. Static analysis can be applied earlier in development because some kinds of defects are hard to find by testing (e.g., timing-dependent errors) and because testing and analysis are complementary; each is best at finding different faults. The sophistication of the analysis performed by tools varies from those that only consider the behavior of individual statements and declarations, to those that include the complete source code of a program in their analysis. Uses of the information obtained from the analysis vary from highlighting possible coding errors to formal methods that mathematically prove properties about a given program [11].

One implementation technique of formal static analysis is model checking; it includes the consideration of systems that have finite state or may be reduced to finite state by abstraction. Data-flow analysis is a lattice-based static analysis technique for gathering information about the possible set of values and the abstract interpretation models the effect that every statement has on the state of an abstract machine. In this case, the model 'executes' the software based on the mathematical properties of each statement and declaration. This abstract machine over-approximates the behaviours of the system and the abstract system is thus made simpler to analyze, at the expense of incompleteness since not every property true of the original system is true of the abstract system. If properly done, though, abstract interpretation is sound as every property true of the abstract system can be mapped to a true property of the original system [12].

The sophistication of the analysis performed by the model varies from those that only consider the behavior of individual statements and declarations, to those that include the complete source code of a program in their analysis. Uses of the information obtained from the analysis vary from highlighting possible coding errors to formal methods that mathematically prove properties about a given program [13]. Static analysis can find weaknesses in the code at the exact location, it can be conducted by trained software assurance developers who fully understand the code and it allows a quicker turn around for fixes. In addition to all these, it permits weaknesses to be found earlier in the development life cycle, reducing the cost to fix.

Program slicing was initially proposed to guide programmers during program debugging, but had been found to be useful for the process of understanding programs. Dynamic slicing was used to identify those parts of the program that contributed to the computation of the selected function for a given program execution. This can be used to understand program execution by adopting a commonly used high level abstraction. Program slicing is the computation of the set of programs statements, the program slice that may affect the values at some point of interest, referred to as a slicing criterion. Program slicing can be used in debugging to locate source of errors more easily. Other applications of slicing include software maintenance, optimization, program analysis, and information flow control. [13]. Slicing techniques have been seeing a rapid development since the original definition by Mark Weiser. At first, slicing was only static, i.e., applied on the source code with no other information than the source code. Reference [14] introduced dynamic slicing, which worked on a specific execution of the program; for a given execution trace. Based on the original definition of [15], informally, a static program slice S consists of all statements in program P that may affect the value of variable v at some point p. The slice is defined for a slicing criterion C=(x,V), where x is a statement in program P and V is a subset of variables in P. A static slice includes all the statements that affect variable v for a set of all possible inputs at the point of interest (i.e., at the statement x). Static slices are computed by finding consecutive sets of indirectly relevant statements, according to data and control dependencies. Dynamic slicing techniques provided a means to prune unrelated computation, and it may help to narrow down this part of a program that contributed to the computation of a function of interest for a particular program input.

III. MATERIALS AND METHODS

The methodology adopted in the work was a modification of the Jakstab framework [16]. Jakstab is an Abstract interpretation-based, integrated disassembly and static analysis framework for designing analyses on executables and recovering reliable control flow graphs. In order to make the framework suitable for the research in this paper, the author added an extra layer of abstraction to the original framework to obtain a modified methodology suitable for the task at hand. While the starting point for the Jakstab framework is binary source, the approach in this paper performs the analysis on the program source code.

In addition to the modified Jakstab framework; the following materials were deployed in building the model-based high-tech source analysis system.

- A bottom-up dynamic slicing technique was defined in this work and deployed to obtain a hybrid-tech static analysis model (Fig. 1). A slice was constituted by an executable portion of the original program whose behavior is, under the same input, indistinguishable from that of the original program on a given variable ‘V’ at point ‘P’ in the program. Reference [13] had showed that bottom-up program slicing techniques could be successfully deployed to transform a large component-based program into a smaller one that contains only statements relevant to the computation of a given function.
- Statecharts notations used in [17] were deployed as a concise visual formalism that captured the dynamic behaviour of a system in representing program slices in this work. Illustrated below is a visualization of a login module Passwd.pas (Fig. 2).
- Goal models- a graph structure representing stakeholder goals and their inter-dependencies was deployed to decompose goals into sub-goals through AND/OR refinements (Fig. 3).

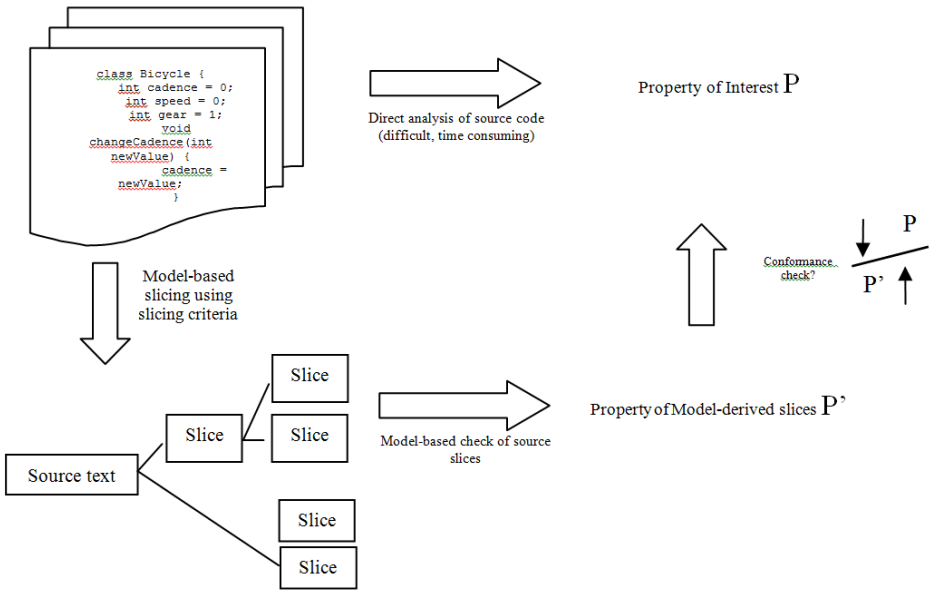


Figure 1. Source model-based slicing for static analysis

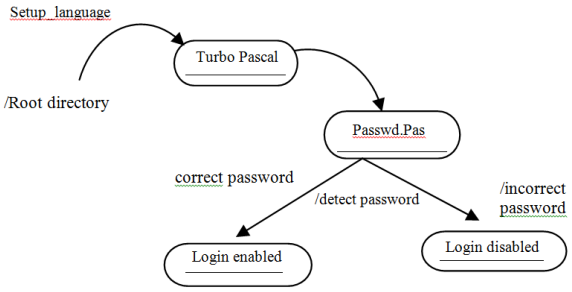


Figure 2. Sample statechart notation

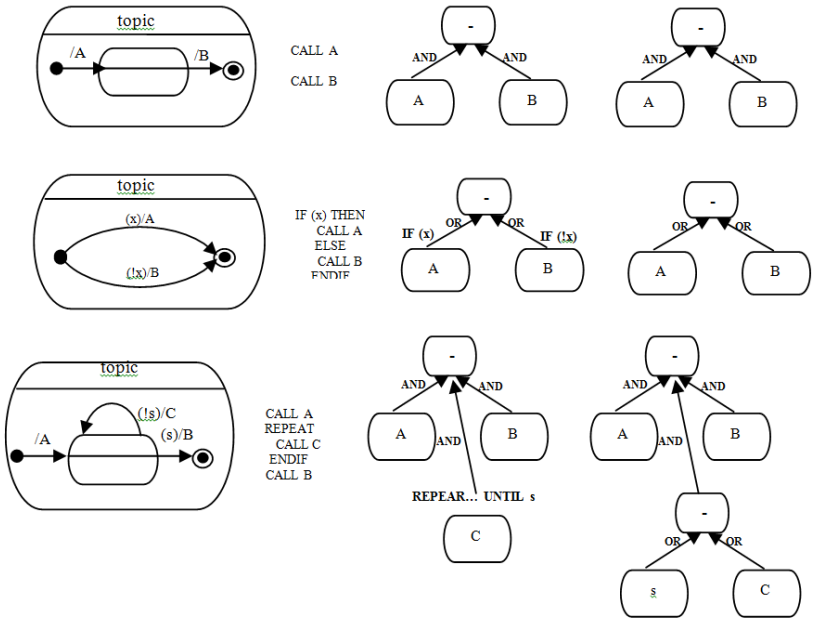


Figure 3. Patterns to extract goal models from abstract code [17]

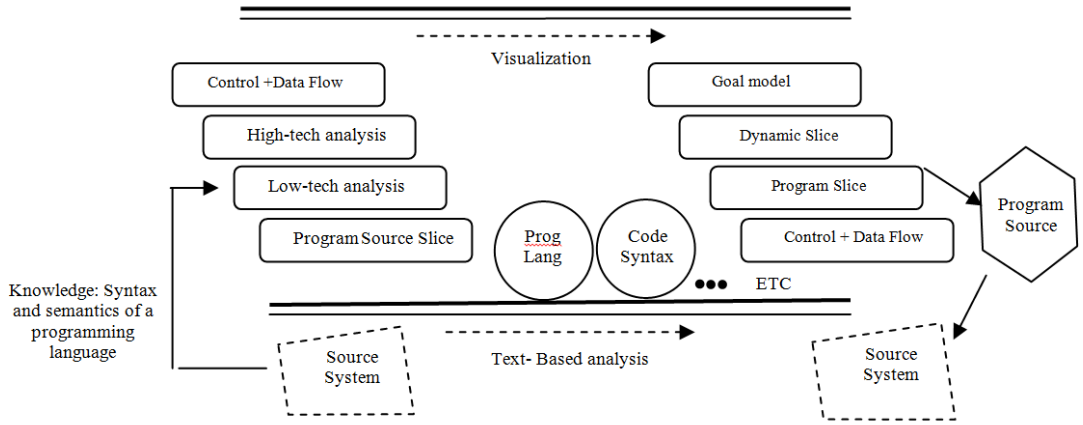


Figure 4. The model-based high-tech source analysis system

IV. Results and Discussions

This work described a model-based high-tech methodology for static source analysis which consisted of systematically using slices of code source models as primary source artifacts throughout the analysis process (Fig.1). This was because, though proving program correctness may be wrong at static analysis phase, it is possible to prove simple properties of simplified models. The process took as input, the program source text and generated slices based on slicing criterion.

The knowledge schema at this point comprised of the knowledge of the syntax and semantics of the programming language (Fig.4). Low-tech static analysis that involved simple software inspection and manual checking of simple syntactic standards were first carried out. This was followed by the high-tech static analysis that included enforced syntactic checks, check for conformance with respect to specifications, designs, and the code data flow analysis. The model was bottom-up and generated slices were analyzed, refined if necessary and evaluated. The sum of all analyzed parts (slices) gave the whole (entire source) at the end of the day which may eventually become a candidate for analysis at a later time. Generated slices were represented by statecharts and resolved into goal models.

The Static analysis model was applied to a real-life legacy application [13]. A procedural application developed for a commercial bank in Nigeria prior to the consolidation of the banking sector in the country was chosen due to its high availability and statecharts were deployed to show the abstract description of the behaviour of the source system. Fig. 5 showed the statechart that implemented the login page of the application, passwd.pas.

Visualizations like the one above (Fig. 5) were built for different slices of the application and the different visualizations were combined to obtain a high-level meta-model that contained the entire description of the original system (Fig. 6).

The top-level statechart above was converted to an annotated goal model using the conversion process described earlier (Fig. 3), all the transitions were converted into goals using the AND/OR decomposition rules. Some tasks in the goal model contributed to quality concerns modeled by the softgoals; for example, "correctPassword" contributed to the security concern while "ErrorMessage" contributed to the usability concern (Fig.7).

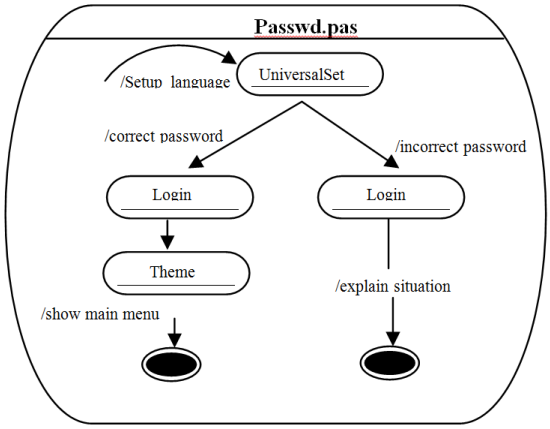


Figure 5. Sample visualization of the login module using Statechart

V. Conclusion

Previous research had showed that the maintenance of existing software source code consumed up to ¾ of the resources in the total software budget (time and material), and that the earlier an error is determined, the lower is the cost of its correction. In addition, [18] opined that code comprehension require 47% and 62% of the total time for enhancement and correction tasks, respectively. The main justification for this work therefore is to leverage on the most important advantage of static code structure analysis in re-tooling software maintenance for a developing economy- cost saving. In this work, a framework for code structure analysis for a large component based software system with program slicing was developed as a re-tooling technique for developing economies in order to enable an early detection of bugs during a software development process thereby saving significant costs in software maintenance.

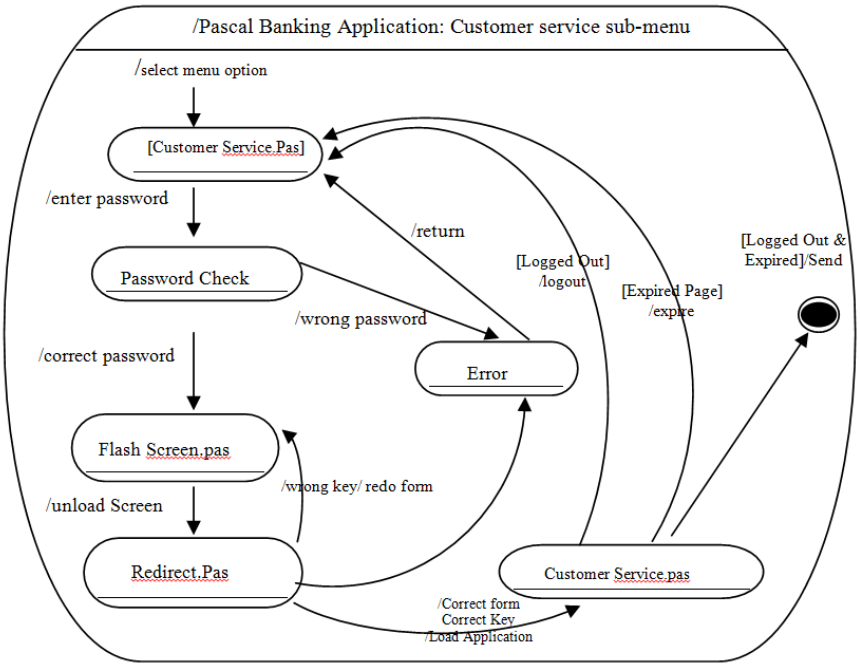


Figure 6. Sample top-level Statechart of the Procedural application

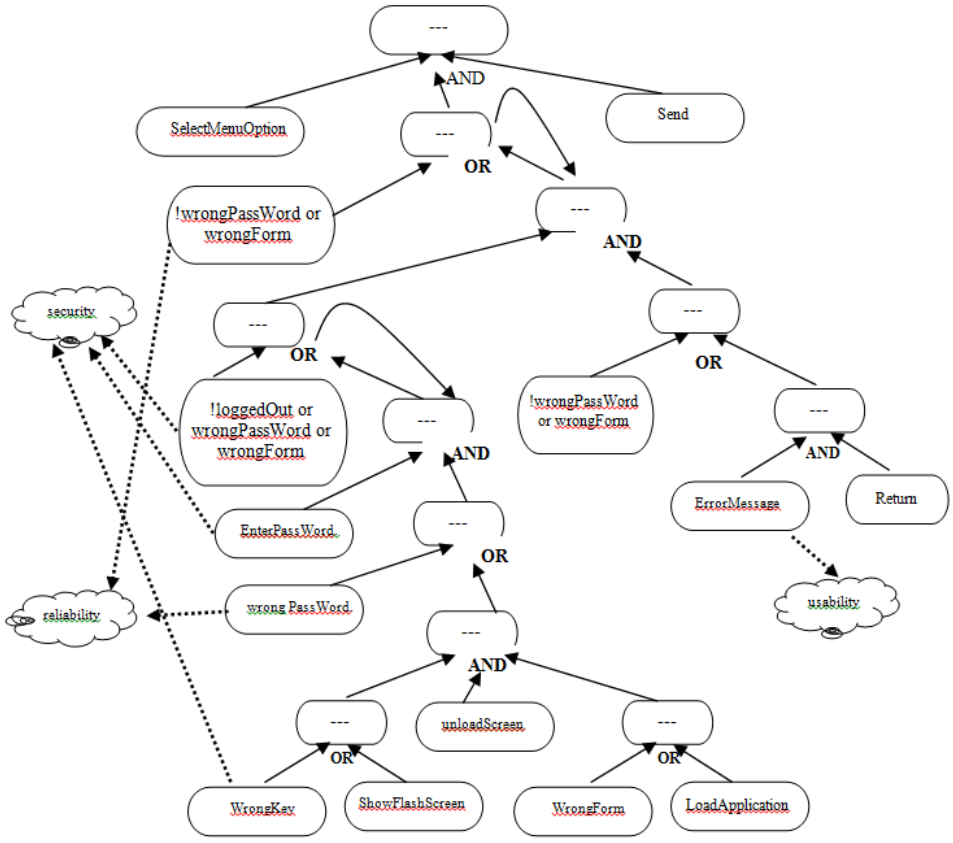


Figure 7. Sample goal model for the application [13]

REFERNCES

[1] B. Livshits, “Improving Software Security with Precise Static and Runtime Analysis” Section 7.3 “Static Techniques for Security,” Stanford doctoral thesis, 2006. http://research.microsoft.com/en-us/um/people/livshits/papers/pdf/thesis.pdf

[2] B. A.Wichmann, A. A. Canning, D. L. Clutterbuck, L. A. Winsbarrow, N. J. Ward, and D. W. R. Marsh, “Industrial Perspective on Static Analysis,” Software Engineering Journal vol 10, pp. 69-75, March, 1995.

[3] B. W. Boehm, “Software engineering economics,” Prentice-Hall, Englewood Cliffs, NH, 1981.

[4] C. McClure, “The three Rs of software automation,” Prentical Hall, Englewood Cliffs, NJ, 1992.

[5] A.F. Ackerman, L. S. Buchwald, and F. H. Lewski. “Software Inspections: An Effective Verification Process,” IEEE Software, Vol. 6, No. 3, May 1989, pp. 31-36.

[6] K. Erdos, & H. M. Sneed, "Partial Comprehension of Complex Programs enough to perform maintenance," Proceedings of the IEEE Sixth International Workshop on Program Comprehension, June 24 – 26, 1998.

[7] S. McConnell, Code Complete, 2nd ed., Microsoft Press: Paperback, 2004, 914 pages, ISBN: 0-7356-1967-0.

[8] R. Klosch, “Reverse Engineering: Why and How to Reverse Engineer Software”, Proceedings of the International Conference on Software Engineering, 2001, pp. 123-132.

[9] K. Thiagarajan, C.Saravanakumar, G. Poonkuzhali, Ponnammal Natarajan, and S.Jeyabharathi, “Static program slicing for composite data using FSM-Model,” World Academy of Science, Engineering and Technology Journal, Issue 32 Aug. 2009, pp. 820-824. Downloaded December 2011 from http://www.waset.org/journals/waset/v32.php

[10] A. Ireland, “Static Analysis Techniques,” Lecture notes on F28SD2: “Software Design”, School of Mathematical and Computer Science, Heriot-Watt University, Edinburgh

[11] Wikipedia the free encyclopedia, “Static Program Analysis,” Downloaded November 2011 from http://en.wikipedia.org/wiki/Static_program_analysis

[12] P. Jones, "A formal methods-based verification approach to medical device software analysis". Journal of Embedded Systems Design, 2010.

[13] O.F. Oladipo, “Software reverse engineering of legacy applications,” Ph.D. Dissertation Computer Science Department, Nnamdi Azikiwe University, Awka Nigeria, March 2010, unpublished

[14] B. Korel and J. Laski. “Dynamic program slicing,” Information Processing. Letters, vol. 29, no 3, pp.155-163, Oct. 1988.

[15] M. Weiser. "Program slicing". IEEE Transactions on Software Engineering, vol. 10, Issue 4, pages 352–357, IEEE Computer Society Press, July 1984.

[16] Jakstab Framework homepage http://www.jakstab.org/

[17] Y. Yu, Y. Wang, J. Mylopoulos, S. Liaskos, A. Lapouchnian, and J. Cesar Sampaio do Prado Leite, “Reverse Engineering Goal Models from Legacy Code,” In: 13th IEEE International Conference on Requirements Engineering (RE'05), 29 Aug-2 Sept 2005, Paris, France, Downloaded October 2009 from www.cs.toronto.edu/~alexei/pub/re05re.pdf

[18] M. L. Nelson, “A Survey of Reverse Engineering and Program Comprehension,” Lecture notes on CS 551: “Software Engineering Survey,” Old Dominion University, April 19, 1996, Downloaded November 2011 from arxiv.org/pdf/cs/0503068

AUTHOR’S PROFILE

Oladipo, Onaolapo Francisca holds a Ph.D in Computer Science from Nnamdi Azikiwe University, Awka, Nigeria; where she is currently a faculty member. Her research interests spanned various areas of Computer Science and Applied Computing. She has published numerous papers detailing her research experiences in both local and international journals and presented research papers in a number of international conferences. She is also a reviewer for many international journals and conferences. She is a member of several professional and scientific associations both within Nigeria and beyond; they include the British Computer Society, Nigerian Computer Society, Computer Professionals (Regulatory Council) of Nigeria, the Global Internet Governance Academic Network (GigaNet), International Association of Computer Science and Information Technology (IACSIT), the Internet Society (ISOC), Diplo Internet Governance Community and the Africa ICT Network.

Transform Domain Fingerprint Identification Based on DTCWT

Jossy P. George
Department of Computer Science
Christ University
Bangalore, Karnataka, India

Abhilash S. K., Raja K. B.
Dept. of Electronics and Communication Engineering
University Visvesvaraya College of Engineering
Bangalore, Karnataka, India

***Abstract*— The physiological biometric characteristics are better compared to behavioral biometric identification of human beings to identify a person. In this paper, we propose Transform Domain Fingerprint Identification Based on DTCWT. The original Fingerprint is cropped and resized to suitable dimension to apply DTCWT. The DTCWT is applied on Fingerprint to generate coefficient which form features. The performance analysis is discussed with different levels of DTCWT and also with different sizes of Fingerprint database. It is observed that the recognition rate is better in the case of level 7 compared to other levels of DTCWT.**

Keywords-Fingerprint; DTCWT; Euclidean Distance; Preprocessing

I. INTRODUCTION

The term biometric is derived from the Greek word bio (life) and metrics (to measure). The biometrics identifies the person, based on feature vectors derived from physiological or behavioral characteristics. The biometric traits must satisfy universality, uniqueness, permanence, accessibility, collectability. The physiological biometrics are Fingerprint, Hand Scan, Iris Scan, Facial Scan and Retina Scan etc. and behavioral bio-metric are Voice, Keystroke, Gait, Signature etc. The physiological biometrics traits are almost constant throughout the life span of a person [1], even for identical twins. The behavior biometric trait varies with mood and environment.

The bio-metric authentication is better than traditional methods. The existing traditional methods to authenticate are Passwords, Personnel Identification Numbers (PINs), Tokens and Smart Cards. The disadvantages of traditional methods are (i) Passwords and Pins are difficult to remember. (ii) More chances of losing tokens and smart cards. (iii) The misuse of traditional methods of authentication by miscreants is very high, especially in the case of money transaction through ATMs, access to the unauthorized places, etc.

Biometric traits are identified in different domains, such as spatial, transformation and hybrid domain. In spatial domain biometric traits are considered as image itself, wherein features of image are obtained by computing the area, height, width, pixel density, mean, variance and standard deviation etc. In transform domain, the spatial domain image is converted into other domain using Fast Fourier Transform (FFT), The Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), Dual-Tree Complex Wavelet Transform (DTCWT), Empirical Mode Decomposition (EMD), Singular Value Decomposition (SVD), Principal Components Analysis (PCA), Independent Component Analysis (ICA), etc. These coefficients of transform domain form feature vectors. In hybrid domain, the combination of spatial and transformation domain features are obtained. The features of spatial, transformation and hybrid domains are compared using distance vector formulae such as Equivalent Distance (ED), Hamming Distance (HD), Chi-square test, or classifiers such as Neural Networks (NN), Support Vector Machine (SVM) and Random Forest (RF).

Fingerprint is a physiological trait which is constant throughout the life span of a person. The fingerprint is unique even for the twins. It has been used as the identification technique for over a century. The fingerprint authentication officially established as a means of identifying people around 1900s. This authentication got the popularity because; the devices for accessing the fingerprints are small and inexpensive. When a biometric verification is to occur, a scan of the biometric of a person is made and which is to be compared with the stored data of the same person.

A fingerprint usually appears as a series of dark lines that represents the high, peaking portion of the friction ridge skin, while the valleys between these ridges appears as white space and are the low, shallow portion of the friction ridge skin, minutiae, or the location and direction of the ridge endings and bifurcations along a ridge path. The upper most point on the inner most ridge of the fingerprint image is known as core. The fingerprints are mainly classified as Arch, Tented Arch, Whorl, Left Loop, and Right Loop. There are separating point between pattern area and non-pattern area in many of the fingerprint images. These points are generally known as delta.

A. Contribution: In this paper the Transform Domain Fingerprint Identification Based on DTCWT at different levels is introduced. DTCWT generates complex coefficients by using a dual tree of wavelet filters to obtain two parts of the images, i.e., real and imaginary part. DTCWT is applied on Fingerprint to generate different levels to obtain Fingerprint features.

B. Organization: The Introduction is given in section I, the existing research papers are discussed in section II, the proposed model is explained in section III, the algorithm is described in section IV, the performance analysis is discussed in section V and finally, conclusion is given in section VI.

II. LITERATURE SURVEY

The Fingerprint for individual identification and verification with existing different technique and applications are described in this section.

S. Vasuki et. al., [2] proposed a model for segmentation of a color textured regions of a given images. This is obtained by the segmentation and by applying DTCWT. This model works in two levels where in first level, after applying DTCWT, the image is divided into 16 sub images and from where the maximum energy sub image is selected as an optimum feature space. In the second level, K-means spatial refinement algorithm is applied. The main advantage of this model is the accuracy. Zhao Song and Liu Yuanpeng [3] gave a novel image diagnosing scheme by applying 2-D DTCWT to the second bandlet transform. This is obtained based on the shift – invariance and better directionality of the DTCWT. The bandlet reconstruction recovers the transitions and directional textures. This improves significantly image diagnosing results. Chen Feng and Yu Song – nian [4] proposed a model which retrieves the multi scale image by using a new class of image features as the image descriptors from DTCWT. From this work, we can conclude that the performance of DTCWT is better in the experiment on the stander COREL image database due to the rotation invariance, translation invariance, robust to noise and getting the key point according to people's cognitive habits. Sathesh and Samuel Manoharan [5] discussed on advantages and disadvantages of DWT. Also they give the methods to overcome the limitations of DWT and the theoretical analysis of complex wavelet transform and its verification using the simulated images. V. Lulian and B. Monica [6] proposed a method to design and optimize separately two channels perfect reconstruction filter banks. This method ensures the good quality for the two levels. This method is more useful where the only the few levels of decomposition is required.

Shahid and Gupta S [7] proposed a novel method to fuse an image using the DTCWT. This is achieved through the formation of a fused pyramid using DTCWT coefficients from the decomposed pyramids of a source image. This method gives better qualitative and quantitative results than the DWT methods. Yun and Cho [8] proposed an adaptive preprocessing method, which extracts five features from the fingerprint images, analyzes image quality with clustering method, and enhances the images according to their characteristics. The preprocessing is performed after distinguishing the fingerprint image quality according to its characteristics. Brankica M. Popovi´c and L. Maskovic [9] used multi scale directional information obtained from orientation field image to filter the spurious minutiae. The feature extraction in pattern recognition system is to extract information from the input data and depends greatly on the quality of the images. Multi scale directional information estimated based on orientation field estimation. F. A. Afar et. al., [10] presented the minutiae based Automatic Fingerprint Identification Systems. The technique is based on the extraction of minutiae from the thinned, binarized and segmented version of a fingerprint image. The system uses fingerprint classification for indexing during fingerprint matching. G. Jagadeeswar Reddy et. al., [11] presented fingerprint diagnosing using both wavelet and Curvelet Transforms. The search-rearrangement method performs better than minutiae based matching for fingerprint binary constraint graph matching since it does not require implicit alignment of two fingerprint images.

K. Zebbiche and F. Khelifi [12] presented biometric images as one Region of Interest (ROI) that has the data processed by most biometric based system. The scheme consists of embedding the watermark into ROI in fingerprint images. Discrete Wavelet Transform and Discrete Fourier Transform are used. Bhupesh Gour et. al., [13] introduced midpoint ridge contour representation in order to extract the minutiae from fingerprint images. Color coding scheme is used to scan each ridge only once. Seung Hoon chae and Jong Ku Kim [14] proposed Fingerprint Verification in which both minutiae and ridge information are used to reduce the errors due to incomplete alignment or distortion. This works gives more importance to the areas where we get errors in processing the algorithms. Aparecido Nilcau Marana and A. K. Jain [15] proposed Ridge Based Fingerprint matching using the Hough transform. The major straight lines that match the fingerprint ridges are used to estimate rotation and translation parameters. This method gives better and accurate results. Anil K Jain et al., [16] described the use of logistic regression method to integrate multiple fingerprint matching algorithms. The integration of Hough transform based matching, string distance based matching and 2D dynamic programming based matching using the logistic regression has minimized the False Rejection Rate for a specified level of False Acceptance Ratio.

Fanglin Chen and Jie Zhou [17] proposed an algorithm for reconstructing fingerprint orientation field from saved minutiae and are used in the matching stage to compare with the minutiae from the query fingerprint. The orientation fields computed from the saved minutiae is a global feature and the saved minutia is the local feature, are used to get more information. Chunxian Ren and Yilong Yin [18] used the hybrid algorithm based on linear classifier to segregate foreground and background blocks. The pixel wise classifier uses three pixel features such as Coherence, mean and variance. Hartwig Front haler and Klaus Kollreider [19] used a multi grid representation of a discrete differential scale space enhancement strategy of fingerprint recognition system. The fingerprint image is decomposed using Laplacian Pyramid as relevant information is concentrated within a few frequency bands. The Fausian Directional Filtering is used to enhance ridge valley pattern of fingerprint using 1-D filtering on higher pyramid level. The linear symmetric features are used to extract the local ridge –valley orientation. D. R. Shashi Kumar et. al., [20] proposed fingerprint verification based on fusion of minutiae and ridges using strength factors. In this model, to extract minutiae and ridges, block filter and Hough Transform are used. This proposed algorithm gives better results than the many other existing ones. This is achieved by fusing the minutiae and the ridge parameters using strength factors.

III. MODEL

In this section, definitions of Performance Analysis and proposed model are discussed.

A. Definitons

a) *False Accept Rate (FAR):* It is the probability that system incorrectly matches with images stored with input image database. The FAR can be calculated using Equation 1.

$$FAR = \frac{\text{No. of persons accepted from out of database}}{\text{Total no. of persons in database.}} \quad (1)$$

b) *False Rejection Rate (FRR):* It is the probability that system fails to recognize the correct pattern to match with the database images. It is the ratio of number of correct persons rejected in the database to the total number of persons in database and can be calculated using Equation 2.

$$FRR = \frac{\text{No. of correct persons rejected}}{\text{Total no. of persons in database.}} \quad (2)$$

c) Equal Error Rate (EER): It is the value where both the reject and accept rates are equal.

d) *True Success Rate (TSR):* It is the ratio of total number of persons correctly matched in the database to the total number of persons in the database and is given by Equation 3.

$$TSR = \frac{\text{No. of persons correctly Matched in the database}}{\text{Total no. of persons in database.}} \quad (3)$$

B. Proposed Model

The proposed model of Fingerprint Identification using DTCWT is given in the Fig. 1.

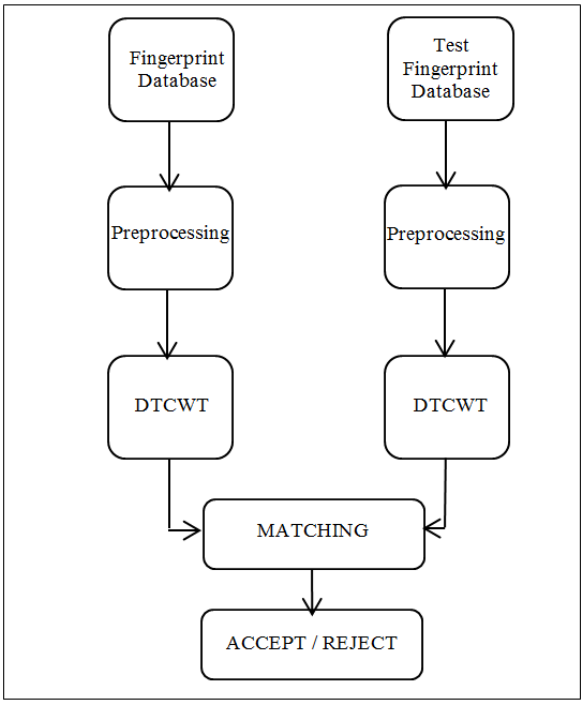


Figure 1. Proposed model

a) *Fingerprint Database:* Database collection is one of the important works in testing the biometric system. The best and advisable way of collecting the database is with a different sensor. There are different databases made available for the researchers to study on the different biometrics systems. The Fingerprint database available are the first, second and third International Competition on Fingerprint verification such as FVC 2000, FVC 2002 and FVC 2004 [21] respectively.

Four distinct databases for FVC 2004 provided by the organizers constitute the benchmark: DB1, DB2, DB3 and DB4. Each database is 110 fingers wide and 8 samples per finger in depth i.e., it consists of 880 fingerprint images. Each database is partitioned into disjoint subsets A and B.

The subsets DB1-A, DB2-A, DB3-A and DB4-A, which contain the first 100 fingers (800 images) of DB1, DB2, DB3 and DB4, respectively, is used for the algorithm performance evaluation. The subsets DB1-B, DB2-B, DB3-B and DB4-B, containing the last 10 fingers (80 images) of DB1, DB2, DB3 and DB4 respectively made available to the participants to allow parameter tuning before executable(s) submission.

The Fig. 2 shows a sample image from each database. The DB3_A database is considered to test our algorithm due to its high resolution and size compatibility. A sample of finger print with eight impressions of DB3 is given in Fig. 3.

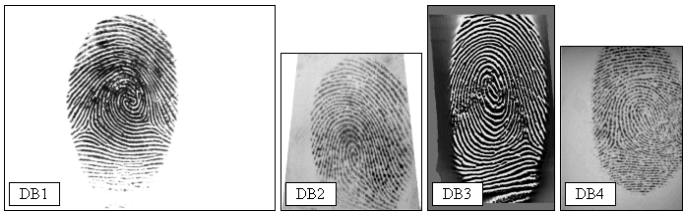


Figure 2. One fingerprint image from each database

Figure 3. Fingerprint sample of DB3_A

- Source Database: The first seven Fingerprint images of each person from DB3 _A database of FVC 2004 are stored.
- Test Template: The eighth Fingerprint of each person from DB3 _A database of FVC 2004 are used in the test template and is compared with source database to compute FRR and TSR.
- Mismatch template database: The DB3_B of FVC 2004 database of 10 fingers are stored in Mismatch template database and compared with source database to compute FAR.

b) Pre-processing: The original Fingerprint image is of size 480 X 300. An observing the DB3_A of FVC 2004, we crop the input image to the size 401 X 201 in order to remove the unwanted portion in the image. And then the cropped image is resized into 512 X 256 for the DTCWT requirement.

c) DTCWT: Dual Tree Complex Wavelet Transform is a recent enhancement technique to the Discrete Wavelet Transform with some additional properties and changes. It is a effective method for implementing an analytical wavelet transform, introduced by Kings bury in 1998 [22], [23], [24]. DTCWT gives the complex transform of a signal using two separate DWT decompositions ie., tree 'a' and tree 'b'. DTCWT produces complex coefficients by using a dual tree of wavelet filters and gives real and imaginary parts which are shown in Fig 4.

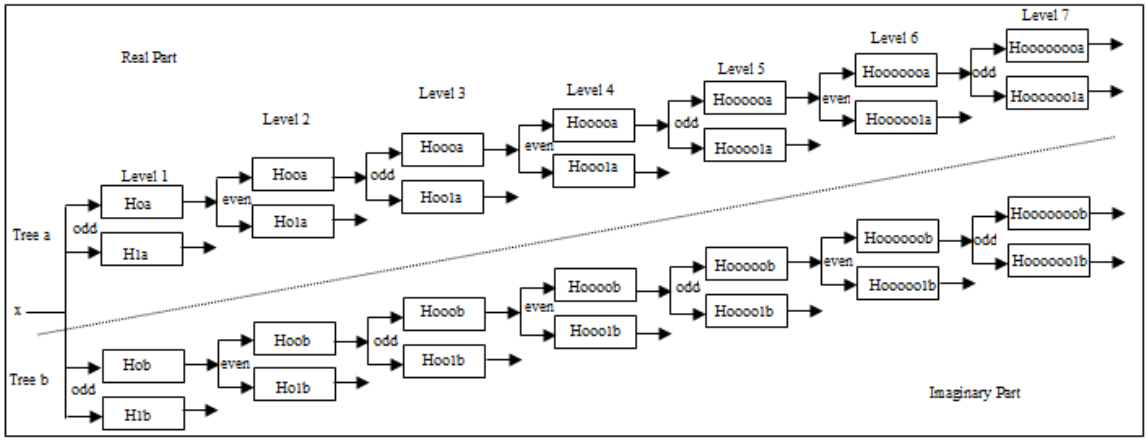


Figure 4. Real and imaginary parts of the complex coefficients

The DTCWT has following properties:

- Approximate shift invariance;
- Good directional selectivity in 2-dimensions (2-D) with Gabor-like filters also true for higher dimensionality: m-D);
- Perfect reconstruction (PR) using short linear-phase filters;
- Limited redundancy: independent of the number of scales: 2:1 for 1-D (2m :1 for m-D);
- Efficient order-N computation - only.
- DTCWT differentiates positive and negative frequencies and generates six sub bands oriented in ±15°, ±45°, and ±75 °. The different levels of DTCWT such as levels 5, 6, and 7 are applied on pre-processed Fingerprint Image.

d) Matching: The Euclidian Distance (ED) is used to verify the test image with the database images using the Equation 4.

$$d1(p,q) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(p_i - q_i)^2} \qquad (4)$$

Where,

M = the dimension of feature vector.
Pi = is the database feature vector.
qi = is the test feature vector.

IV. ALGORITHM

A. Problem Definition

The physiological trait Fingerprint is used to identify a person using the features obtained by the coefficients of DTCWT. The proposed algorithm for the performance analysis of the fingerprint identification for different levels of DTCWT is given in the Table1.

The objectives are;

- Fingerprint verification to authenticate a person using DTCWT
- To achieve high TSR
- To have FRR and FAR very low

While applying the DTCWT in different levels, the number of features and dimensions are reduces. The Fingerprint images for level-1, level-2, level-3, level-4 and level-5, level-6, level-7 are shown in Fig. 5.

TABLE 1. PROPOSED ALGORITHM.

1) FVC 2004, DB3_A database is considered.
2) Pre-processing is done by cropping the input fingerprint image to 401 X 201.
3) Cropped image is resized to 512 X 256 for DTCWT requirement.
4) DTCWT is applied on Fingerprint with levels 5, 6 and 7.
5) Magnitude and phase resulted from DTCWT are concatenated and considered as features.
6) Source database is created with the features obtained by step 5.
7) For the test fingerprint DTCWT is applied and features obtained using step 5.
8) Test Fingerprint is compared with the database fingerprint using ED to verify a person

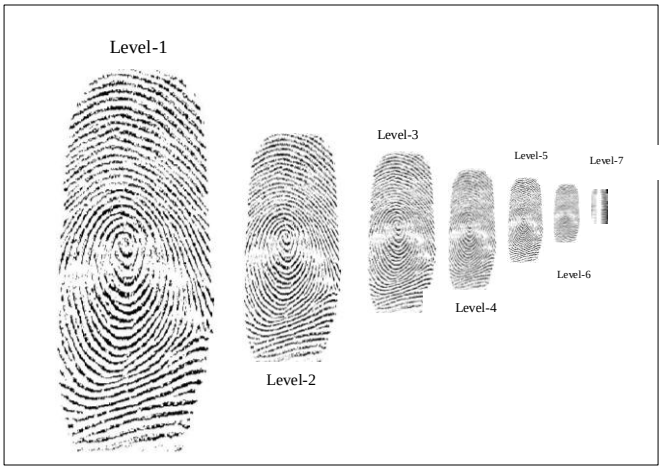


Figure 5. DTCWT images at different levels.

V. Performance Analysis

For the performance analysis, DB3_A of FVC 2004 Fingerprint database is considered. The number of persons inside the database (PIDB) to compute FRR and TSR are varied from 30 to 90 and the number of persons outside the database (PODB) are 10 to compute FAR is given in Table 2.

It is observed from the table 2 that the values of EER and TSR depend on the quality of Fingerprint image than the number of images in PIDB and PODB. The values of EER and TSR are better in the case of PIDB: PODB of 40:10. The performance of recognition rate is better in DTCWT level 7 compared to other lower levels of DTCWT. The TSR and EER is 85% and 0.15 respectively for DTCWT level 7 with PIDB: PODB of 40:10.

The graph for FRR, FAR and TSR is given in Fig. 6 and the variations of FRR and TSR with threshold for POIB: PODB of 40:10 is tabulated in Table 3 and It is noticed that as threshold increases, the value of FRR decreases, whereas the values of FAR and TSR increases. The highest success rate of recognition of 85% is achieved for the threshold value of 2.4.

VI. Conclusion

The Fingerprint biometric is used to authenticate a person. In this paper, Transform Domain Fingerprint Identification Based on DTCWT is proposed. The Fingerprint is preprocessed to a suitable size that suit DTCWT. The Fingerprint features are obtained by applying DTCWT with different levels. The test image features are compared with Database images using Euclidean Distance. It is observed that the recognition rate is better in the case of DTCWT level 7 compared to lower levels with PIDB: PODB of 40:10. In future, the algorithm may be combined with spatial domain features such as global and local features to enhance recognition rate.

TABLE 2. EER AND TSR FOR DIFFERENT LEVELS OF DTCWT

levels		PIDB:PODB					
		30:10	40:10	60:10	70:10	80:10	90:10
5	EER	0.5	0.2	0.573	0.34	0.36	0.33
	TSR	50%	80%	42.7%	66%	64%	67%
6	EER	0.45	0.2	0.59	0.3	0.282	0.3
	TSR	55%	80%	41%	70%	71.8%	70%
7	EER	0.36	0.15	0.228	0.21	0.197	0.197
	TSR	64%	85%	77.2%	79%	80.3%	82.1%

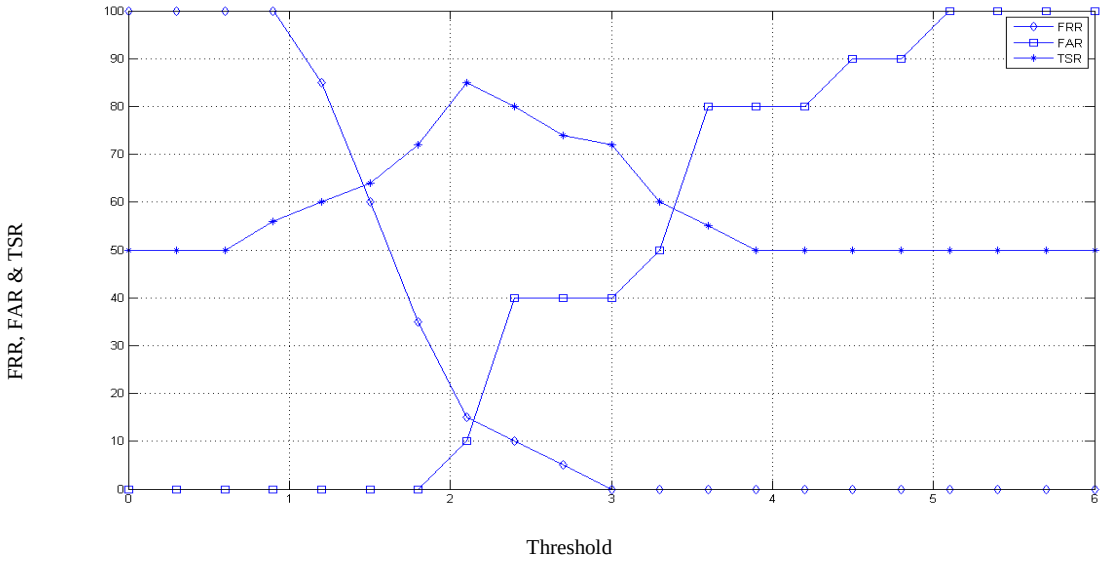


Figure 6. Variations of FRR, FAR and TSR with respect to different thresholds

TABLE 3. VALUES OF FRR, FAR AND TSR FOR DIFFERENT THRESHOLDS

Threshold	FRR	FAR	TSR %
0	1	0	50
0.3	1	0	50
0.6	1	0	50
0.9	1	0	50
1.2	0.85	0	50
1.5	0.6	0	60
1.8	0.35	0	64
2.1	0.15	0.1	72
2.4	0.1	0.4	85
2.7	0.05	0.4	80
3	0	0.4	74
3.3	0	0.5	72
3.6	0	0.8	60
3.9	0	0.8	55
4.2	0	0.8	50
4.3	0	0.9	50
4.8	0	0.9	50
5.1	0	1	50
5.4	0	1	50
5.7	0	1	50
6	0	1	50

REFRENCES

[1] B. N. Lavanya, K. B. Raja and K. R. Venugopal, "Fingerprint verification based on gabor filter enhancement," International Journal of Computer Science and Information Security,vol. 6, no. 2, pp. 138-144, 2009.

[2] S. Vasuki, L. Ganesan and R. Florin Raja Singh, "DT- WT based segmentation algorithm for color images," RTCSP Conference Proceedings, Coimbatore, pp. 207-212, 2009.

[3] Zhao Song and Liu Yuanpeng, "A novel image denosing scheme via combining dual tree complex wavelet transform and bandelets," IEEE International Symposium on Intellignet Information Technology Application, pp. 509-512, 2009.

[4] Chen Feng and Yu Song-nian, "Content – based image retrieval by DTCWT feature," IEEE International Conference on Computer Research and Development, vol. 4, pp. 283–286, 2011.

[5] Sathesh and Samuel Manoharan, "A dual tree complex wavelet transform construction and its application to image denosing," International Journal of Image Processing, Vol. 3, pp. 293-300, 2010.

[6] Lulian Voicu and Monica Borda, "New method of filters design for dual tree complex wavelet transform," IEEE International Symposium on Signals, Circuits and Systems, vol. 2, pp. 437-440, 2005.

[7] Mohd Shahid and Sumna Gupta, "Novel masks for multimodality image fusion using dtcwt," IEEE International Conference, TENCON, pp. 1-6, 2005.

[8] E. K. Yun and S. B. Cho, "Adaptive fingerprint image enhancement with fingerprint image quality analysis," International conference of Image and Vision Computing, pp. 101–110, 2006.

[9] M. P. Brankica and L. Maskovic, "Fingerprint minutiae filtering based on multiscale directional information," FACTA Universitatis-Series: Electronics and Energetics, vol. 20, pp.233-244, August 2007.

[10] F. A. Afsar, M. Arif and M. Hussain, "Fingerprint identification and verification system using minutiae matching," National Conference on Emerging Technologies, pp.141-146, 2004.

[11] G. Jagadeeswar Reddy, T. Jaya Chandra Prasad and M. N. Giri Prasad, "Fingerprint image denoising using curvelet transform," Proceeding of Asian Research Publishing Network Journal of Engineering and Applied Sciences, vol 3, no 3, pp. 31-35, June 2008.

[12] K. Zebbiche and F. Khelifi "Region-based watermarking of biometrics images: Case study in fingerprint images," Proceedings of International Journal of Digital Multimedia Broadcasting, pp. 1-13, 2008.

[13] Bhupesh Gour, T. K. Bandopadhyaya and Sudhir Sharma, "Fingerprint feature extraction using midpoint ridge contour method and neural network," Proceedings of International Journal of Computer Science and Network Security, vol.8, no.7, pp. 99-103, July 2008.

[14] Seung-Hoon Chae and Jong Ku Kim "Ridge-based fingerprint verification for enhanced security," Digest of Technical Papers International Conference on Consumer Electronics, pp 1-2, 2009.

[15] A. N. Marana, and A. K. Jain, "Ridge-based fingerprint matching using hough transform," IEEE Proceedings of the Brazilliab Symposium on Computer Graphica and Image Processing, pp. 112-119, October 2005.

[16] A. K. Jain, S. Prabhakar, and A. Chen, "Combining multiple matchers for a high security fingerprint verification system," Pattern Recognition Letters, Elsevier Science Direct, vol.20, pp 1371- 1379, 1999.

[17] Fanglin Chen and Jie Zhou, "Reconstructing orientation field from fingerprint minutiae to improve minutiae-matching accuracy," IEEE Transactions on image processing, vol. 18, no 4, pp 1665-1670, 2009.

[18] Chunxiao Ren and Yilong Yin, "A linear hybrid classifier for fingerprint segmentation," Fourth International Conference on Neural Computation, pp 33-37, 2008.

[19] Hartwig Fronthaler and Klaus Kollreider, "Local features for enhancement and minutiae extraction in fingerprints," IEEE Transactions on Image Processing, vol. 17, no 3, pp 354-363. 2008.

[20] D. R. Shashi Kumar, R. K. Chhotaray, K.B. Raja and Sabyasachi Pattanaik, "Fingerprint verification based on fusion of minutiae and ridges using strength factors," International Journal of Computer Applications, vol. 4, no.1, 2010.

[21] D. Maio, D. Maltoni, R. Cappelli, J.L. Wayman and A. K. Jain, "FVC2004: Third Fingerprint Verification Competition," Lecture Notes in computer science, pp.1-5, 2004.

[22] N.G. Kingsbury, "The dual-tree complex wavelet transform: A new technique for shift invariance and directional filters," Proceeding of 8th IEEE DSP Workshop, Utah, 1998.

[23] N.G. Kingsbury, "Image processing with complex wavelets," Philos. Trans. R. Soc. London A, Math. Phys. Sci., vol. 357, no. 1760, pp. 2543–2560, 1999.

[24] N.G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," Appl. Comput. Harmon. Anal., vol. 10, pp. 234–253, 2001.

AUTHORS PROFILE

Jossy P. George currently serves as Assistant Professor of the Department of Computer Science at Christ University, Bangalore. He received the B. Sc. In Computer Science, Bachelor of Philosophy (B. Ph), Bachelor of Theology (B. Th) and Master of Computer Application (MCA). He has done his FDPM from IIM, Ahmedabad. He is pursuing his Ph.D. in Computer Science of Christ University under the guidance of Dr. K. B. Raja, Assistant Professor, Visvesvaraya College of Engineering. His research activities focus on algorithms for improved accuracy in fingerprint and iris biometrics.

Abhilash S. K. obtained his BE in Electronics and Communication Engineering from DR Ambedkar Institute of Technology, Bangalore and is pursing his Postgraduate (ME) in Visvesvaraya college of Engineering, Bangalore. His research interests include Image processing and Biometrics.

Dr. K B Raja is an Assistant Professor, Department of Electronics and Communication Engineering, University Visvesvaraya college of Engineering, Bangalore University, Bangalore. He obtained his BE and ME in Electronics and Communication Engineering from University Visvesvaraya College of Engineering, Bangalore. He was awarded Ph.D. in Computer Science and Engineering from Bangalore University. He has 85 research publications in refereed International Journals and Conference Proceedings. His research interests include Image Processing, Biometrics, VLSI Signal Processing, computer networks.

Effective Security Architecture for Virtualized Data Center Networks

[1]Udeze Chidiebele. C, [3] Okafor Kennedy .C
[1,3] R & D Department, Electronics Development Institute (FMST-NASENI), Awka, Nigeria.

[2]Prof. H. C Inyiama,[4]Dr C. C. Okezie,
[2,4] Electronics and Computer Engineering Department, Nnamdi Azikiwe University, Awka, Nigeria

***Abstract*—This work presents a candidate scheme for effective security policy that defines the requirements that will facilitate protection of network resources from internal and external security threats. Also, it ensures data privacy and integrity in a virtualized data center network (VDCN). An integration of Open Flow Software Defined Networking (OFSDN) with VLAN Virtual Server Security (VVSS) architecture is presented to address distinct security issues in virtualized data centers. The OFSDN with VVSS is proposed to create a more secured protection and maintain compliance integrity of servers and applications in the DCN. This proposal though still on the prototype phase, calls for community driven responses.**

Keywords- Infrastructure; Virtualization; VDCN; OFSDN; VVSS; VLAN; Virtual Server.

I. INTRODUCTION

Recently, data center networks (DCNs) have attracted a lot of interest in the enterprise networking industry. DCNs are used to provide data storage and files transfer where end stations are interconnected as clusters and bladed systems [1]. A data center represents the heart of any organization's network [2]. Companies rely on the data stored in the data center to interact with its employees and customers.

The proliferation of the Web-based technologies makes the data center more vulnerable to security attacks. Any security attack on the data center can destroy the whole organization's network and data [2]. Besides throughput and low latency required in DCNs, the security considerations of enterprise data centers is also very critical. Several researches were dedicated to the security issues and the design constraints of large scale data centers from different points of view [2]. The authors in [2], [3], [4], [5], [6] discussed on the data center security problems, technologies, security strategies such as consolidation, relocation, migration, expansion and review of asset management policies. The authors of [4] carried out an overview of the communication network design problems that arise with large numbers of nodes, links and switch costs. Some layered security models for addressing complex security issues are discussed in [5] and [6]. With fast changing technologies and service demands in DCNs, the need for an effective open platform secure model becomes very imperative.

In this paper with detailed study on the security proposals existing in literature, and having considered all the requirements of network security management for a virtualized data center model, we propose an effective secured model: Open Flow Software Defined Networking (OFSDN) with VLAN Virtual Server Security (VVSS). The design is based on layered security architecture for virtual servers and open flow switch architecture. Operational mechanism is presented in section V with other details. By allowing the MAC controllers in the virtual open flow switch in our DCN to house the flow tables for each virtual port, this work creates lines of defense against any security threat. Unicast, broadcast and multicast traffic are characterized and monitored by the modeled switch architecture which serves as an aggregation link buffer.

The paper is organized as follows. In Section II, we discussed virtualization in data center network, data center security problems as presented in [2]. In section III, the proposed security model (OFSDN) is shown with the Virtual server security system. Section IV gives the experimental setup for VLAN open flow switch. The paper ends with conclusions and future directions

II. VIRTUALIZATION IN DATA CENTER NETWORKS

Server virtualization has become popular in data centers since it provides an easy mechanism to cleanly partition physical resources, allowing multiple applications to run in isolation on a single server [7]. Virtualization helps with server consolidation and provides flexible resource management mechanisms [7] in DCNs particularly. We quickly add that Virtualization is not a new technology, but it has regained popularity in recent years because of the promise of improved resource utilization through server consolidation. According to [8], a Data Center is the consolidation point for provisioning multiple services that drive an Enterprise business. In [2], the authors enlist the data center hardware and software components. The hardware components are: firewalls, Intrusion Detection Systems, contents switches, access switches and core switches. The software components are: IPSec and VPN, antivirus software, network management systems and access control server. However, for effective security implementation in a virtualized DCN, this work goes further to propose a more secured data center design that is programmable, secured with strong isolation, and flexible using the OFSDN approach in our context.

III. DATA CENTER SECURITY PROBLEMS

Data center networks usually have its security threats. The work carried out in [8], [9]and[10] discussed some of these problems, viz: Unauthorized Access, MAC Flooding, ARP Spoofing, IP Spoofing, Denial of Service (DOS), Viruses, Worms, Trojans, and internal Security threats. However, sampled solutions to these problems were given in [2]. We still argue that these solutions do not completely eradicate security vulnerabilities in contemporary data center networks.

For a virtualized data center domain, a restructured architecture which will address the possible lapses in addition to the outlined remedies in [2], will serve in securing today's enterprise networks.

IV. DATA CENTER SECURITY TECHNOLOGIES

Information stored at the data center must be protected from any security threat that may destroy or modify it in any unwanted way [2]. These security threats can originate from hackers outside or from inside the data center network. Different solutions to the security threats can be used together to achieve the highest possible data protection. Some of these technologies are:

- Firewalls.
- Network intrusion detection and prevention systems.
- Virtual Local Area Networks (VLAN).
- Virtual Private Network (VPN) and IPSec.

Leveraging on these four technologies, our contribution is shown in the Open Flow Software Defined Network model in Fig. 2. OFSDN is a layer 2 protocol in the virtual Software Defined Network (SDN) switch that allows for policy control via its open flow visor (virtualization layer). This model creates multiple layers of security for the virtualized DCN controlling unicast, broadcast and multicast traffics. Section IV and V discussed in details the security models for highly scalable and secure virtualized DCN.

V. VLAN VIRTUAL SERVER SECURITY SYSTEM

VLAN Virtual Server Security (VVSS) system proposed in this work for the server VM provides multi-layered workgroup segmentation while utilizing the underlying hardware technology to protect the virtual data center. The VVSS solution is a generic purpose-built framework proposed for large scale enterprises. The virtual environment at the core of the infrastructure is the Vm server running on ESX platform with its VMware.

Fig. 1 shows the VVSS model while Fig. 5 and Fig. 6 show the packet tracer simulation. Again, in our architecture shown in Fig. 2, MAC controllers were assigned to all the network entities to house their flow tables. For active participation in the network, the open flow visor must uniquely identify and authenticate the client node else, the terminal is dropped for access.

As shown in Fig. 1, VLAN virtual security model was modeled to be deployed on a virtualized server for various applications (Vm1…Vm5). The kernel utilizes the hypervisor API to inspect and control the virtual switch network and VM behavior. Virtual Security Service (VSS) utilizes a subnetted IP mapping, which is provided as VMsafe for various user groups. For demonstration in this work, each VM server on virtualized server is managed and configured through packet tracer environment.

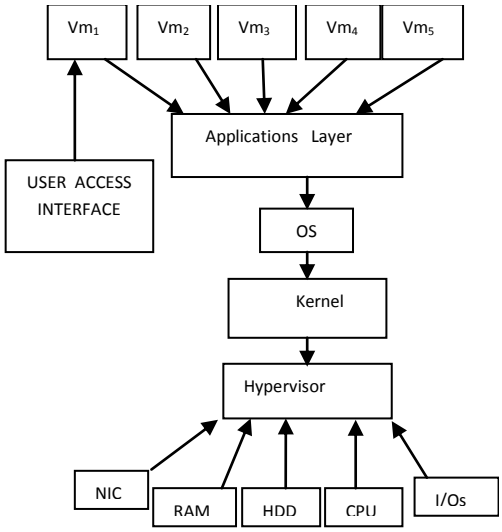


Figure 1. VLAN Virtual Security System Model.

A VLAN backbone which hosts the Vm server is the central manager for the applications. VVSS has the following functions:

- Inter-VM migration of Applications and services for compatibility issues
- Virtual Machine generator and monitor
- Network Access Control (NAC)
- Discovery and Broadcast Isolation
- License and Update Management (LUM)

VI. OPEN FLOW SOFTWARE DEFINED NETWORK MODEL FOR DCN

The Open flow software defined networking switch in figure 2 is a speed redundant device with isolated MAC controllers housing the flow tables shown in Fig. 4. An open flow protocol (OFP) which can be enabled in the switch carries out control policy (CP), reaction execution (RE) and history tracking (HT). Once OFP is enabled on the switch, any device interfaced with the switch is actively monitored as a software robot, thereby securing the overall network against any form of threat. This is proposed for virtualized data center in context. The key security metric is the MAC ID of the interfacing devices.

The security policy of the flow table in Fig. 4 controls activities that is handled by conventional VLAN and Access control list (ACL) such as traffic denial or flow allowance, routing, broadcast isolation flow, flow detection and suppression in the OFSDN switch. All servers, etc shown in Fig. 2 are mapped in the MAC controllers. Fig. 3 shows the open virtual isolation in the OFSDN switch. This model offers a highly secured security layer to existing security approaches in literature.

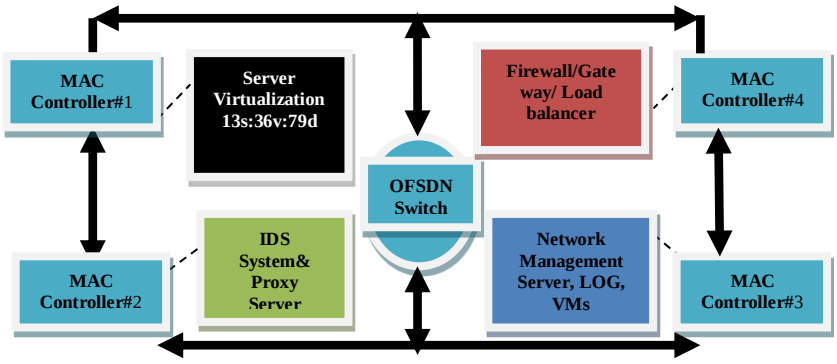


Figure 2. OFSDN Security Model for DCN

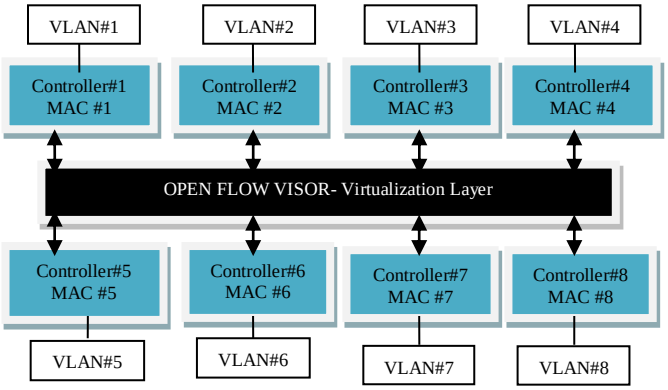


Figure 3. A Virtualized Open Flow Switch

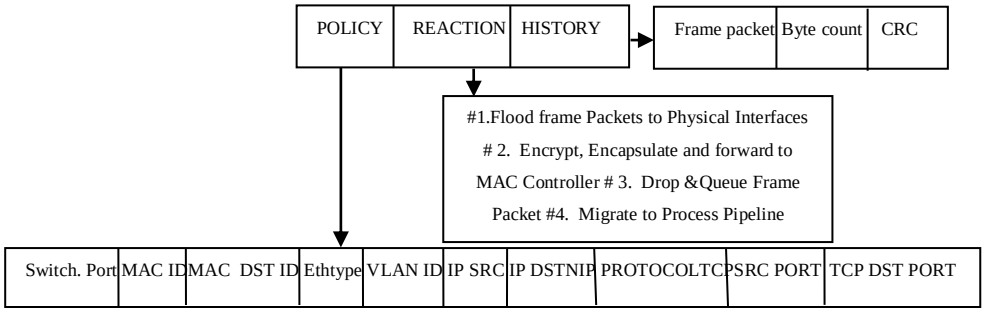


Figure 4. Flow table Ingress in Open Flow Switch Model

VII. EXPERIMENTAL SETUP

The first phase experiment involves virtualizing the server test bed consisting of one standard HP machine with a dual-core Intel Xeon processor connected to a rack-mounted disk enclosure with a Small Computer Scale Interface (SCSI) backplane running on ESX linux sever.

For the purposes of trace security, six VLANs were created for the server and simulated with packet trace tool. In the server, a Seagate model 15,000RPM disks: of size 1TB was considered with a RAM of 6GB. The server was connected via a switched (OFSDN) 1Gbps Ethernet link.

This work provides three fundamental security services:

- Data confidentiality: protecting against unauthorized access to data being transmitted.
- Data integrity: protecting against alteration or future replay of traffic.
- Source authentication: network addresses are authenticated as part of the protocol.

We deployed Classless Inter-domain Routing (CIDR) approach to generate usable IP for the VM server and users on the network. For valid IP range for 200 users with 128Vm servers, we used a class valid host range: 192.168.10.1 to 192.168.10.199 with a subnet mask of 255.255.255.0. For effective security and broadcast isolation, virtual IP mapping on the Vm server enables the hosts, guests and clients to communicate with each other. Fig. 5 and Fig. 6 show the packet flow in the packet tracer integrated development environment (IDE).

TABLE 1: DATA CENTER VM SERVERS (13 SERVERS, 36 VOLUMES, 79 DISKS)

VmServers	**VLAN**	**Volumes**	**IP Mapping**
$UserV_M$	10	3	192.168.10.2
$ProjectV_M$	10	3	192.168.10.3
$PrtrV_M$	20	4	192.168.10.4
$HrdmV_M$	20	5	192.168.10.24
RDVm	20	1	192.168.10.20
$PrxyV_M$	30	2	192.168.10.22
$ScrV_M$	30	3	192.168.10.50
WebVm	40	2	192.168.10.24
MdSVm	40	4	192.168.10.23
ERPVm	40	2	192.168.10.68
NACVm	50	4	192.168.10.70
$E\text{-}ComV_M$	30	2	192.168.10.58
IntrantVm	60	1	192.168.10.78

TABLE 2: AVERAGE UTILIZATION RATES.

Resource	Utilization
CPU	6%
MEMORY	40%
NETWORK I/O	<5%
DISK I/O	<5%

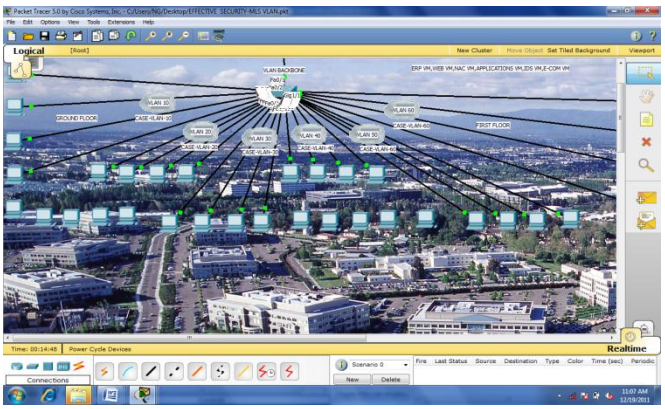


Figure 5. DCN VLAN workgroup Model with ESx server

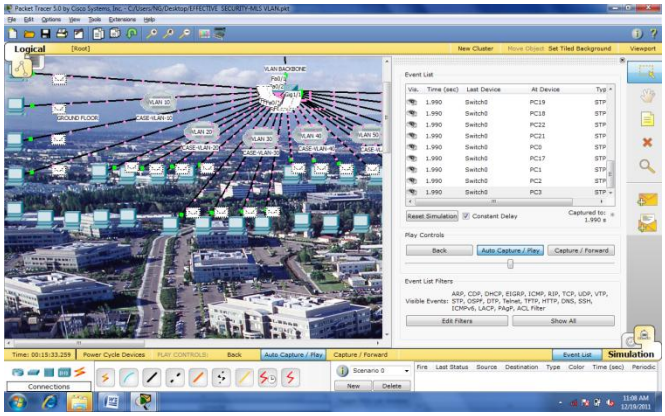


Figure 6. Simulated Packet Flow for secured DCN

VIII. CONCLUSION

The effective security architecture discussed in this paper is conceived to achieve the best possible solution for virtualized data center networks. Owing to advancements in virtualization technology, the security methodologies for traditional data centers which includes: firewalls, intrusion detection system/intrusion protection system, virtual local area network (VLAN) and virtual private network (VPN) cannot effectively handle security implications of a virtualized data center networks. This work presents an effective open flow software defined network switch with VVSS model and with emphasis on VLAN virtualization on ESX server to ensure total security of the critical data in the virtualized data center network.

The analytical model and validation of the proposed models in Fig. 2 and 3 will be clearly shown in the future work; however this work seeks to use the presented approaches to enhance the security design of a virtualized data center network.

REFERENCES

[1] Jinjing. jiang and R.Jain, " Analysis of backward congestion notification (BCN) for ethernet in datacenter applications. IEEE communications Society INFOCOM 2007 proceedings.

[2] Jalal Frihati, Florica Moldoveanu, Alin Moldoveanu , General guidelines for the security of a large scale data centre design, U.P.B. Sci. Bull., Series C, Vol. 71, Issue 3, 2009.

[3] Data centre services, URL, http://www.sun.com/service/storage/datacenterdatasheet.pdf

[4] Practical Large-Scale Network Design With Variable Costs for Links and Switches, URL:http://whitepapers.silicon.com/0,39024759,60304468p,00.htm

[5] Mitchell Ashley "LAYERED NETWORK SECURITY 2006: A best-practices approach",URL:http://www.stillsecure.com/docs/StillSecure_LayeredSecurity.pdf.

[6] Juniper networks layered security solution, URL:http://cn.juniper.net/solutions/literature /white _papers/2005.pdf

[7] Timothy Wood, "Improving data center resource Management, deployment, and availability with virtualization", PHD thesis June,2009,(Unpublished).

[8] http://www.cisco.com/application/pdf/en/us/guest/netsol/ns107/c649/ccmigration_09186a008073377d.pdf

[9] Data center: infrastructure architecture SRND,URL:http://www.cisco.com/application/pdf/en/us/guest/netsol/ns304/c649/cdccont_0900aecd800e4d2e.pdf

[10] Data Center: Securing Server Farms , URL:www.cisco.com/application/pdf/en/us/guest/netsol/ns304/c649/ccmigration_09186a008014edf3.pdf

[11] Data center security topologies: www.cisco.com/application/pdf/en/ us/guest/netsol/ns376/c649/cdccont_0900aecd800ebd1d.pdf