IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 6 Issue 8

August 2015

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

SAI

www.ijacsa.thesai.org

IJACSA

WHERE WISDOM SHARES

INTERNATIONAL JOURNAL OF
ADVANCED COMPUTER SCIENCE AND APPLICATIONS

SAI

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 6 Issue 8 August 2015
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

CONTENTS

Competitive Sparse Representation Classification for Face Recognition

Ying Liu
Chongqing Key Laboratory of Computational Intelligence
Chongqing University of Posts and Telecommunications
Chongqing, China

Cong Li
Chongqing Key Laboratory of Computational Intelligence
Chongqing University of Posts and Telecommunications
Chongqing, China

Jian-Xun Mi
Chongqing Key Laboratory of Computational Intelligence
Chongqing University of Posts and Telecommunications
Chongqing, China

Chao Li
Chongqing Key Laboratory of Computational Intelligence
Chongqing University of Posts and Telecommunications
Chongqing, China

***Abstract*—A method, named competitive sparse representation classification (CSRC), is proposed for face recognition in this paper. CSRC introduces a lowest competitive deletion mechanism which removes the lowest competitive sample based on the competitive ability of training samples for representing a probe in multiple rounds collaborative linear representation. In other words, in each round of competing, whether a training sample is retained or not in the next round depends on the ability of representing the input probe. Because of the number of training samples used for representing the probe decreases in CSRC, the coding vector is transformed into a low dimensional space comparing with the initial coding vector. Then the sparse representation makes CSRC discriminative for classifying the probe. In addition, due to the fast algorithm, the FR system has less computational cost. To verify the validity of CSRC, we conduct a series of experiments on AR, Extended YB, and ORL databases respectively.**

Keywords—face recognition; collaborative representation sparse representation; and competitive representation

I. Introduction

Face Recognition (FR) has become to a hot research area for its convenience in daily life. Recently, linear representation methods are very popular which represent the probe with training samples from gallery set. Collaborative representation (CR) method has achieved good performance for FR [1-4],in which a given testing image $\mathbf{y}$ can be represented by a training set **A** with a coding vector **x**, i.e. **y=Ax**. The training set **A** including all samples from all subjects is an over-complete dictionary. It is known that face images from a specific class lie in a linear subspace and a probe can be represented by images which have the same label as the probe. Comparing to the single-representation method, collaborative representation has more ability to compensate the pixels of probe. In order to make the coding vector more discriminative, the sparse constraint was introduced in regular term. Inspired by compressive sensing, the induced sparse constraint on coding vector uses l_0 -norm so that the representation problem is formulized as:

$$\min \| \mathbf{x} \|_0 \quad s.t. \quad \| \mathbf{y} - \mathbf{A}\mathbf{x} \|_2 \le \varepsilon \tag{1}$$

Where $\| . \|_0$ denotes the l_0 -norm, which counts the number of nonzero entries of the coding vector and ε is a small error tolerance.

The l_0 -norm is widely discussed and used in many researches. However, the problem of find the sparsest solution of an underdetermined system suffers the issue of NP-hard. Researchers put forward different solutions to l_0-norm [5-7]. Now, the sparse constraint methods in CR can be divided into two categories: first one uses l_1 -norm constraint instead of l_0-norm, and second one employs supervised sparse scheme. The first CR method uses l_1 -norm in FR is Sparse Representation Classification (SRC) [1]. In SRC, the sparse representation is solved using Lasso formulation in which the sparse degree of the coding vector can be adjusted by the norm constraint intensity. However, the l_1 -norm constraint could not bring good performance when the training samples have high correlation [8]. In addition, l_1 -norm based sparse representation problem is quite time consuming. A two-phase sparse representation (TPTSR) was proposed by Xu. et.al using supervised sparse constraint scheme [9]. The M nearest neighbors of a probe are selected based on the first phase of representation and then used as the new training set in the second phase of representation in TPTSR. However, the one-time deletion in TPTSR may lead to all samples from the classes as probe are removed when the probe is seriously distorted. In addition, TPTSR is sensitive to illumination because of the samples with negative coefficients are likely removed.

In this paper, we propose a supervised sparse constraint method named as competitive sparse representation classification (CSRC). Based on the competitive ability of each training sample for representing a probe, the proposed CSRC introduces a lowest competitive deletion mechanism which removes the lowest competitive samples based on the competitive ability of each training sample in collaborative linear representation. Only those samples with high competitive can be used in the next collaborative representation. Then the dimensionality of the coding vector in the next representation is bigger than the current one. The multi-phase deletion is more useful for classification than two-phases deletion [10]. According to that the probe is represented based on the multi-phase deletion and until the condition is satisfied. Meanwhile, the dimensionality of the final coding vector is much smaller than the first one's, i.e. the coding vector is sparse. In CSRC, the competitive ability of samples from correct class as probe is increasing as the lowest competitive ones are removed. In addition, the fast algorithm of CSRC enhances the efficiency of the FR system, because the algorithm avoids the procedure of finding inverse matrixes in each collaborative representation.

One advantage of CSRC is the multi-phases deletion lets the competitive ability of the correct class is strengthened gradually and avoids all samples form correct class of probe are removed in the one-time collaborative representation. The other one is that comparing with l_1 -norm sparse constraint, CSRC has lower computational complexity with the fast algorithm.

This paper is organized as followed: in section 2, three parts are described: the introduction about a basic general framework for classification using competitive sparse representation, description the optimization method, and the analysis of the computational cost of CSRC. The features of CSRC are described in section 3.We conduct a series of experiments to verify the good performance of our method in section 4 and the conclusions are demonstrated in section 5.

II. Competitive Sparse Representation Classification

Face images from a same class lie in a linear subspace, a query image can be represented by within-class samples [11, 12]. But face recognition is a lack of samples problem in general. When the number of training samples of each class is not big enough, it is hard to obtain a good representation of a query image by a small part of training samples that from a single-subject. That is to say, the representation has large distance with the query image. Thus, the recognition result is unstable. However, the query image can be represented faithfully by collaborative linear representation which each training samples are put in the dictionary (sometimes the dictionary is over-completed). Because more training samples participate in representing the query image in collaborative linear representation, the competitive of training samples that from the class as the query decreases. As a consequence, the query image is likely classified into the wrong class. However, in the collaborative representation, not every sample has high competitiveness (high coefficient value). So removing these less competitive images from the dictionary can increase the competitive of samples from correct class.

The lowest competitive deletion mechanism in CSRC, which delete the lowest competitive training samples from the dictionary in multi-phases. This mechanism can increase the competitive of the correct samples through removing the lowest competitive samples. In the meantime, as the samples are removed in CSRC, the dimensional of representation coefficients are smaller. Compared with the space of the initial representation coefficients, the final representation coefficients lie in a subspace of it. In other words, the representation coefficients are sparse. The sparse coding vector (representation coefficients) has more discriminate information.

A. Competitive Sparse Representation Classification

Given sufficient training samples of the i th object class, $\mathbf{A}=\left\{\mathrm{a}_{i,n_i}\in R^m\right\}\in R^{m\times n}$ ($i=1,2,...,c$), where n denotes the number of training samples and each class contains n_i training samples. Meanwhile each training sample is an m -dimensionality feature vector. A probe image $\mathbf{y}\in R^m$ is represented by collaborative linear representation over the dictionary. Since CSRC is a multi-phase deletion method, the symbol $\mathbf{A}^t$ and $\mathbf{x}^t$ ($t=1,2,...$) denote the training samples and coding vector in t th collaborative representation respectively. The representation framework is written as following:

$$\min\|\mathbf{x}^1\|_2^2 \ \text{ s.t. } \ \|\mathbf{y}-\mathbf{A}^1\mathbf{x}^1\|_2^2\leq\varepsilon \qquad (2)$$

Where ε is the noisy term. The Eq. (2) can be written as ridge regression form:

$$\mathbf{x}^1=\arg\min_{\mathbf{x}}\left\{\|\mathbf{y}-\mathbf{A}^1\mathbf{x}^1\|_2^2+\lambda\|\mathbf{x}^1\|_2^2\right\} \qquad (3)$$

The coding vector can be computed as:

$$\mathbf{x}^1=(\mathbf{A}^{1^{\mathrm{T}}}\mathbf{A}^1+\lambda\mathbf{I}^1)^{-1}\mathbf{A}^{1^{\mathrm{T}}}\mathbf{y} \quad \mathbf{x}^1\in R^n \qquad (4)$$

Where the unit matrix $\mathbf{I}^1\in R^{n\times n}$ and λ is a Lagrangian coefficient. Since CSRC deletes only one training sample in each phase, CSRC removes the least competitive samples based on the corresponding entries of the coding vector, i.e., it finds the minimum absolute value $|x_{\mathrm{j}}|$ amongst the representation coefficients $|x_{\mathrm{j}}|=\min\{|x_1|,|x_2|...|x_n|\}$, and remove the corresponding training sample $\mathbf{a}_j$. So the dictionary is divided into two subsets: the first subset includes the deleted image, and the second subset includes the retained samples. Here the samples in the second part will be used as a dictionary in the second phase. Let $\mathbf{A}^{1r}$ and $\mathbf{A}^2$ denote the removed sample after the first representation and the training samples in 2th representation respectively. So the above two subsets can be described as $\mathbf{A}^{1r}=\left\{\mathbf{a}_j\,|\,\mathbf{a}_j\in R^m\right\}$ and $\mathbf{A}^2=\left\{\mathbf{A}^1\right\}-\left\{\mathbf{A}^{1r}\right\}$, $\mathbf{A}^2\in R^{m\times(\mathrm{n}-1)}$. Then the test $\mathbf{y}$ can be represented over the new dictionary $\mathbf{A}^2$. In the same way, repeatedly conduct the above operation in the next representation phase.

Assume the k th collaborative representation reaches the maximum number of the deletion phases. The final coding vector $\mathbf{x}^k$ can be represented as following:

$$\mathbf{x}^k=(\mathbf{A}^{k^{\mathrm{T}}}\mathbf{A}^k+\lambda\mathbf{I})^{-1}\mathbf{A}^{k^{\mathrm{T}}}\mathbf{y} \quad \mathbf{x}^k\in R^{n-k+1} \qquad (5)$$

Since many samples are removed from the dictionary $\mathbf{A}^1$

that is used in the first representation, it is very likely that $\mathbf{A}^k$ excludes all the samples of some classes. Therefore, the origin classification problem is weakened to a simpler problem which contains fewer classes. The coefficients of the deleted samples as zero and then select the coefficients associated with the i th class and mark it as $\boldsymbol{\delta}_i$, $i=1,2,...,c$. Then the Euclidean distance is used for measuring the distance between each class and the test image $\mathbf{y}$. The rule of classification is in favor of the class with minimum distance. The formula can be expressed as following:

$$\mathrm{ID(i)} = \arg\min_i d_i(\mathbf{y}) = \arg\min_i \|\mathbf{y}\text{ - }\mathbf{A}_i\boldsymbol{\delta}_i\|_2 \quad i=1,2,...,c \tag{6}$$

The detailed algorithm is given as following:

Algorithm: Competitive Sparse Representation Classification (CSRC)

```
1.  Input: an unidentified image $\mathbf{y} \in R^n$ , the initial dictionary $\mathbf{A}^1 \in R^{n\times m}$ .
2.  Initial value: $t=1$
3.  Repeat
4.  Compute the coding vector according to (3)
5.  Removing the training sample $\mathbf{a}_j$ from the dictionary $\mathbf{A}^t$
6.  Update the dictionary : $\mathbf{A}^{t+1}$, $t=t+1$
7.  Until satisfy termination conditions
8.  Identify : $\mathrm{ID}(i) = \arg_i \min(\|\mathbf{y}-\mathbf{A}_i\boldsymbol{\delta}_i\|_2)$
9.  Output: the identity of $\mathbf{y}$ as $i$ .
```

B. Optimization

As it known to all, the analytical solution of the above linear model is $\mathbf{x}=(\mathbf{A}^{\mathrm{T}}\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{y}$. Due to the deletion operation, the dictionary is updated in each phase. However, the new dictionary is the subset of the last dictionary. CSRC implements a fast algorithm to avoiding the repeated matrix inversion calculation.

Now let a new symbol $\tilde{\mathbf{A}}$ expresses the elementary transformation of $\mathbf{A}$, i.e. $\tilde{\mathbf{A}}=\mathbf{AE}$, where $\mathbf{E}$ is an elementary matrix. $\tilde{\mathbf{A}}$ can be treated that the matrix contains two matrices $\mathbf{A}^s$ and $\mathbf{A}^r$, i.e. $\tilde{\mathbf{A}}=\left\{\mathbf{A}^s,\ \mathbf{A}^r\right\}$. The two matrices $\mathbf{A}^r$ and $\mathbf{A}^s$ denote the deleted sample and the new training samples respectively. Since the matrix $\mathbf{A}$ is given, so the inverse matrix $(\mathbf{A}^{\mathrm{T}}\mathbf{A}+\lambda\mathbf{I})^{-1}$ is available. Then the matrix $(\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}+\lambda\mathbf{I})^{-1}$ can be derived from $(\mathbf{A}^{\mathrm{T}}\mathbf{A}+\lambda\mathbf{I})^{-1}$. The detailed derivation processes are written as follows:

$$\begin{aligned}(\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}+\lambda\mathbf{I})^{-1} &= ((\mathbf{AE})^{\mathrm{T}}(\mathbf{AE})+\lambda\mathbf{I})^{-1} = (\mathbf{E}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AE}+\lambda\mathbf{I})^{-1}\\ &= (\mathbf{E}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{AE}+\alpha\mathbf{E}^{\mathrm{T}}\mathbf{IE})^{-1} = (\mathbf{E}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{A}+\lambda\mathbf{I})\mathbf{E})^{-1}\\ &= (\mathbf{E}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{A}+\lambda\mathbf{I})\mathbf{E})^{-1} = \mathbf{E}^{-1}(\mathbf{A}^{\mathrm{T}}\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{E}\end{aligned} \tag{7}$$

The key problem for solving the coding vector is inverse matrix. In order to have a convenience expression in (7), the equation can be represented by the four matrices ($\mathbf{O}$, $\mathbf{P}$, $\mathbf{C}$, and $\mathbf{V}$) as:

$$(\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}+\lambda\mathbf{I})^{-1} = \begin{bmatrix}\mathbf{A}^{s^T}\mathbf{A}^s+\lambda\mathbf{I}^{s\times s} & \mathbf{A}^{s^T}\mathbf{A}^r+\lambda\mathbf{I}^{s\times r}\\ \mathbf{A}^{r^T}\mathbf{A}^s+\lambda\mathbf{I}^{r\times s} & \mathbf{A}^{r^T}\mathbf{A}^r+\lambda\mathbf{I}^{r\times r}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{O} & \mathbf{P}\\ \mathbf{C} & \mathbf{V}\end{bmatrix}^{-1} \tag{8}$$

According the elementary transformation of matrix, the inverse matrix will be transformed as following:.

$$\begin{aligned}\left[\begin{array}{cc|cc}\mathbf{O} & \mathbf{P} & \mathbf{I}_s & \mathbf{0}\\ \mathbf{C} & \mathbf{V} & \mathbf{0} & \mathbf{I}_r\end{array}\right] &= \left[\begin{array}{cc|cc}\mathbf{O} & \mathbf{P} & \mathbf{I}_s & \mathbf{0}\\ \mathbf{0} & \mathbf{V}-\mathbf{CO}^{-1}\mathbf{P} & -\mathbf{CO}^{-1} & \mathbf{I}_r\end{array}\right]\\ &= \left[\begin{array}{cc|cc}\mathbf{A} & \mathbf{0} & \mathbf{I}_s+\mathbf{P}(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\mathbf{CO}^{-1} & \text{-}\mathbf{P}(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\\ \mathbf{0} & \mathbf{V}-\mathbf{CO}^{-1}\mathbf{P} & -\mathbf{CO}^{-1} & \mathbf{I}_r\end{array}\right]\\ &= \left[\begin{array}{cc|cc}\mathbf{O} & \mathbf{0} & \mathbf{I}_s+\mathbf{P}(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\mathbf{CO}^{-1} & \text{-}\mathbf{P}(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\\ \mathbf{0} & \mathbf{I}_r & -(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\mathbf{CO}^{-1} & (\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\end{array}\right]\\ &= \left[\begin{array}{cc|cc}\mathbf{I}_s & \mathbf{0} & \mathbf{O}^{-1}+\mathbf{O}^{-1}\mathbf{P}(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\mathbf{CO}^{-1} & -\mathbf{O}^{-1}\mathbf{P}(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\\ \mathbf{0} & \mathbf{I}_r & -(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\mathbf{CO}^{-1} & (\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\end{array}\right]\end{aligned} \tag{9}$$

Then

$$(\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}+\lambda\mathbf{I})^{-1} = \begin{bmatrix}\mathbf{O} & \mathbf{P}\\ \mathbf{C} & \mathbf{V}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{O}^{-1}+\mathbf{O}^{-1}\mathbf{P}(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\mathbf{CO}^{-1} & \text{-}\mathbf{O}^{-1}\mathbf{P}(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\\ -(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\mathbf{CO}^{-1} & (\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\end{bmatrix} \tag{10}$$

In fact, the ultimate goal is to obtain the solution inverse matrix about $\mathbf{A}^s$, which is used as the new dictionary in the next iteration, then the inverse matrix can be written as $(\mathbf{A}^{s^T}\mathbf{A}^s+\lambda\mathbf{I}^{s\times s})^{-1}$. Not hard to find that

$$(\mathbf{A}^{s^T}\mathbf{A}^s+\lambda\mathbf{I}^{s\times s})^{-1} = \mathbf{O}^{-1} = \mathbf{O}^{-1}+\mathbf{O}^{-1}\mathbf{P}(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\mathbf{CO}^{-1} \tag{11}$$

Where the delighted things are that $\mathbf{O}$, $\mathbf{P}$, $\mathbf{C}$ and $\mathbf{V}$ are already known in the matrix $\tilde{\mathbf{A}}$. A group of new symbols are introduced to express the four block of $(\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}+\lambda\mathbf{I})^{-1}$, i.e.,

$$\mathbf{Q}_{11} = \mathbf{O}^{-1}+\mathbf{O}^{-1}\mathbf{P}(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\mathbf{CO}^{-1} \tag{12}$$

$$\mathbf{Q}_{12} = \text{-}\mathbf{O}^{-1}\mathbf{P}(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1} \tag{13}$$

$$\mathbf{Q}_{21} = \text{-}(\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\mathbf{CO}^{-1} \tag{14}$$

$$\mathbf{Q}_{22} = (\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1} \tag{15}$$

Combining (12) and (13), (14), and (15) respectively, the four block matrices can be descripted as following

$$\begin{cases}\mathbf{Q}_{11} = \mathbf{O}^{-1}-\mathbf{O}^{-1}\mathbf{PQ}_{21}\\ \mathbf{Q}_{12} = \text{-}\mathbf{O}^{-1}\mathbf{PQ}_{22}\\ \mathbf{Q}_{21} = \text{-}\mathbf{Q}_{22}\mathbf{CO}^{-1}\\ \mathbf{Q}_{22} = (\mathbf{V}-\mathbf{CO}^{-1}\mathbf{P})^{-1}\end{cases} \tag{16}$$

After a few times iterate replacements, the $\mathbf{O}^{-1}$ can be expresses by other blocks, i.e.,

$$\mathbf{O}^{-1} = \mathbf{Q}_{11} - \mathbf{Q}_{12}\mathbf{Q}_{22}^{-1}\mathbf{Q}_{21} \tag{17}$$

Here $\mathbf{Q}_{22}$ is a invertible square matrix in general case. Since only one the lowest training sample is removed in each phase, so $\mathbf{Q}_{22}$ is a scalar. In other words, the analytical solution about the inverse matrix $(\mathbf{A}^{s^T}\mathbf{A}^s + \lambda\mathbf{I}^{s\times s})^{-1}$ is obtained, i.e.

$$(\mathbf{A}^{s^T}\mathbf{A}^s + \lambda\mathbf{I}^{s\times s})^{-1} = \mathbf{Q}_{11} - \mathbf{Q}_{12}\mathbf{Q}_{22}^{-1}\mathbf{Q}_{21} \tag{18}$$

C. Complexity analysis

We only analyze the time complexity of (5) in this section, since this process is the most time consuming in algorithm of CSRC. It is time consuming way to solve (5) directly in a certain collaborative representation and the time complexity is

$$O(n^2m + n^3)$$

However, the time complexity that CSRC obtains the coding vector is much less than (19). Since the matrix inversion in (5) is replaced by (18). Moreover, the matrix $\mathbf{Q}_{22}$ in (18) is only one element, it save more calculation. Therefore, the complexity of CSRC is denoted as $O(n^2m)$. In addition, it is much less than the complexity that SRC obtains the spares coding vector, i.e. $O(n^2m^{1.5})$.

III. Analysis of CSRC

The method Collaborative representation based on Classification (CRC) uses l_2 -norm constraint coding vector in the collaborative representation [13]. Although the coding vector is not sparse, it fully embodies the competitive level of each training sample in CRC. However, CRC obtains the regression model uses only one collaborative representation. When the number of the training samples is small, it may lead to regression model over-fitted. The competitive representation, adopted in CSRC, removes the lowest competitive training sample from the training set in the current round and the rest training samples will be used in the nest round. After several rounds of competing, all samples of the subject which has a low correlation with the query may be removed. So CSRC reduces the scale of the FR problem. The Fig. 1, depicts the residuals between the probe and prediction of each class which are calculated by CRC and CSRC respectively, illustrates this phenomenon on ORL database. The upper one in the figure is obtained by CRC and the bottom one is obtained by CSRC. The distance between the probe and the predictions from over 20 classes is 1, which means the training samples from these classes are removed and the coefficients respect to them are zeros. In addition, in CSRC method, the lowest competitive deletion mechanism reduces residual the correct class of the probe and enlarges the residuals of the wrong classes. According to the figure it is easy to calculate that the ratio of two smallest residuals by CRC and CSRC are 1.579 and 1.717 respectively. The ratio of two smallest residuals is enlarged by CSRC, which means CSRC has better discriminative than CRC. It is easy to find that CSRC reduces the interference of the wrong classes.

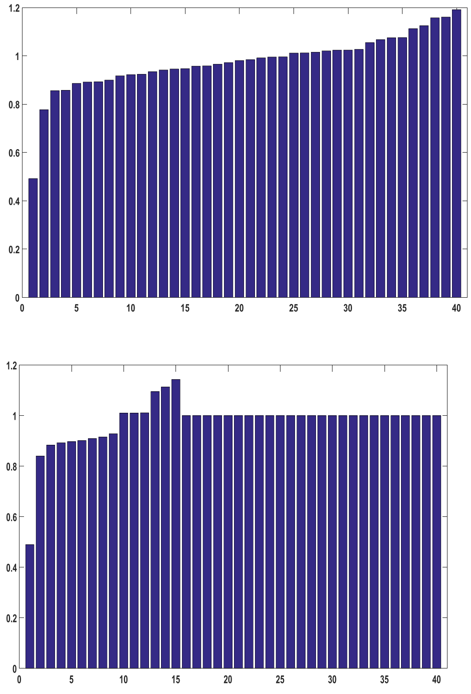


Fig. 1. The residual images by CRC and CSRC for a clear testing face on the ORL database. We select the fifth image of the first man as a probe and the first five images as training samples. In above two histograms, the horizontal axis denotes the number of the class and the vertical axis denotes the residuals between the probe and each class. The top one: the two smallest residual are 0.4922 and 0.7774 by CRC and the ratio of them is 1.579. The bottom one: two smallest residual are 0.4895 and 0.84.3 by CSRC and the ratio of them is 1.717

IV. Experimental Results

To evaluate the proposed CSRC algorithm, we conduct a serious of experiments on images from AR database, Extended YB database and ORL database respectively, as well as comparing with state-of-the-art methods including CRC, SRC (without extended matrix), and TPTSR (the candidate set is 10%). We also assess the recognition rate of CSRC(the candidate set is 10%) on the occluded testing faces. All experiments are performed in MATLAB on 2014b on desktop with 4GHz CPU and 8G RAM.

A. Face Recognition without Occlusion

AR database

More than 4000 color face images of 126 people (70 men and 56 women) consist in AR database [14]. Each people has 26 images include frontal views of face with different facial expression, illumination and occlusion. The pictures of each individual were taken in two sessions (separated by two weeks). Each section contains 13 color images and 120 individuals (65 men and 55 women) participated in both sessions. The images of these 100 individuals (50 women and 50 men) were selected and used in our experiment. Faces that are used to test these methods are gray and then normalized it to 50×40 pixels.

We select the first seven faces in session one as training samples and the first seven faces in the session two as the testing samples for each class and a specific class faces are shown in Fig. 2. The recognition rates of the four methods are shown in the Tab. 1. Since the testing samples and training samples are collected in different time, none of all methods have a 100% recognition rate. Since the training samples have high correlation, the sparse representation by SRC could not obtain a good performance. However, since CRC could not increase the competitive ability of the samples from the correct class as probe, so CRC has lower result than TPTSR and CSRC. Furthermore, TPTSR has lower recognition rate than CSRC, because of that TPTSR is likely to delete all images from the correct class in the first phase. From the experiment we can see that the lowest competitive deletion mechanism in CSRC makes the coding vector has more discriminant information indeed.

Extended Yale B database

The extended Yale B face database contains 38 persons under 64 illumination conditions [15,16]. A subset (contains 31 individuals) is used in this experiment. The 64 images of a person in a particular pose are acquired at camera frame rate of 30 frames/ second, so there is only small change in head pose and facial expression for those 64 images. Each image is resized to 50×40 pixels in our experiment. Several frontal faces of one person are shown in Fig. 3. As is known to all, illumination is another big challenge for face recognition. Faces were captured under carious laboratory-controlled lighting conditions. Samples in subset one (seven images per person) under nominal lighting condition was used as the gallery. Since the recognition rate for test subset 2and 3 (characterize slight-to-moderate luminance variations) are by all methods. Here we select faces in subset 4 are used for verify CSRC method. Due to the increasing illumination condition, the recognition results are not very high in subset 4. From the Tab. 2, CSRC is better than CRC for testing the illuminated images, which means the deletion mechanism makes the classification more discriminative. In addition, compared with TPTSR, CSRC has about 22% higher recognition result than it. The reason for which is that the training samples from the correct class as probe is easy removed in the two-phases deletion, as well as the samples with negative coefficient are not deleted. To the contrary CSRC reduces the risk that the correct training samples will be removed in one time through the multi-phase competitive deletion.

ORL database

ORL database, created by AT&T lab in Cambridge University, contains 400 face images of 40 subjects, i.e. each individual providing 10 face images, including expression variants, multiple directions of posture change within 20% of the scale of the change. Dimensionality of each face is reduced to 50×40. All face images are show in Fig. 4. For

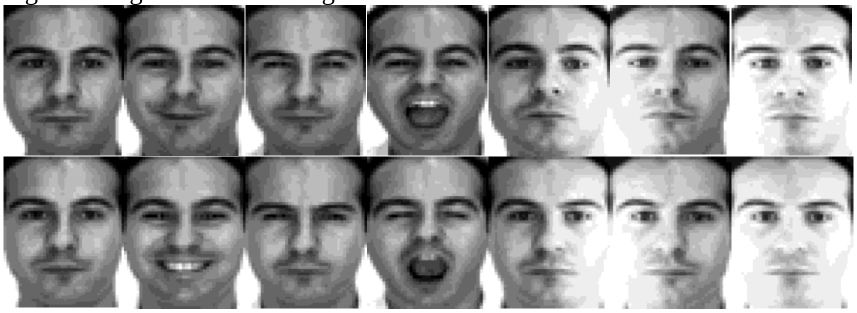


Fig. 2. Frontal faces with emotion and illumination changes on AR database. The top seven faces are from session one and the down seven samples are from session two

TABLE I. RECOGNITION RATES (%) ON AR DSTABASE

Methods	CRC	TPTSR	SRC	CSRC
Recognition rate	*91.571*	*91.857*	*82.714*	*92*

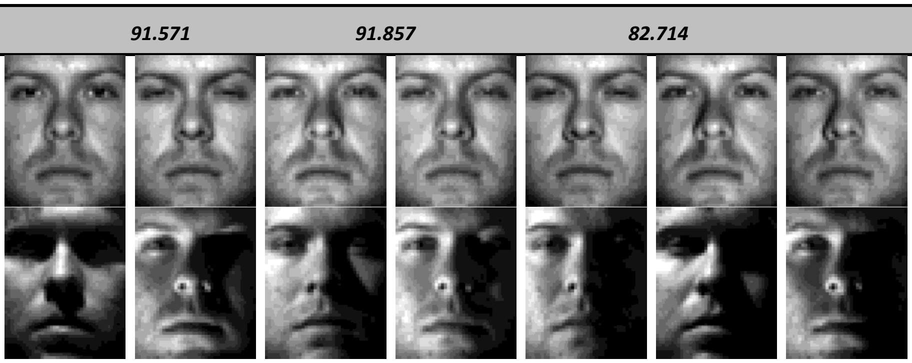


Fig. 3. Some sample faces of a subject from Extended Yale B database. The top row: seven images with moderate illuminance variations from subset 1. The down row: a part images with large illumination variations from subset 4

TABLE II. RECOGNITION RATES (%) ON EXTENDED YALE B DATABASE

methods	CRC	TPTSR	SRC	CSRC
Subset 4	***74.194***	***55.76***	***44.24***	***77.419***

each subject, we choose the first five images as the training images and the rest images are used for testing. From the Tab. 3, the recognition results of these four methods are close. Since the deletion operation in CSRC, the sparse coding vector has more discrimination than CRC, so CSRC has 1.5% higher recognition rate than CRC.

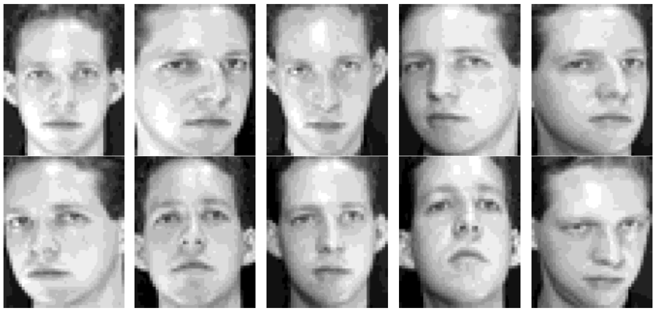


Fig. 4. Ten images of a specific class from the ORL database. The top row represents the training samples and the images in the bottom row are testing samples

TABLE III. RECOGNITION RATES (%) ON ORL DATABASE

methods	CRC	TPTSR	SRC	CSRC
Recognition rate	***87***	***91.5***	***88***	***88.5***

B. Recognition with sunglasses and scarf

In this section we test CSRC's ability to cope with real possibly malicious occlusions using a subset of AR database. The chosen subset consists of 1200 images of 100 subjects, 50 male and 50 female. For each subject, eight frontal faces (half face are from session one and another half are from session two) without occlusions are used as training samples. We select the testing face images with sunglasses (two samples for each subject and each sample with about 20 percent occlusion) and scarf (two samples for each subject and each sample with approximately 40 percent occlusion on the faces) respectively and the testing samples of a specific subject are show in Fig. 5. The recognition rates by TPTSR and CSRC are shown in Tab. 4. CSRC has little better than TPTSR for testing samples with sunglasses. In the scarf case, the recognition rate by CSRC is 26% higher than TPTSR's. Because that the proportion of the scarf almost reaches to 40%, it is likely that the images of correct class as probe are deleted in the first collaborative representation in TPTSR. On the contrary, CSRC makes sure the images of correct class of probe have high competitive.

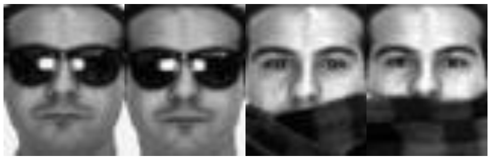


Fig. 5. Face images with sunglasses and scarf respectively on AR database

TABLE IV. RECOGNITION RATES (%) ON AR DATABASE FOR SUNGLASSES AND SCARF

Methods	TPTSR	CSRC
Sunglasses	***56***	***61***
Scarf	***52.5***	***78.5***

V. CONCLUSIONS

In this paper, a competitive representation framework is proposed to solve the sparse representation problem. The lowest competitive deletion mechanism ensures the competitive ability for representing a probe decrease and enhances the competitive ability of the correct class as the probe. What's more the fast algorithm makes the FR system more efficiency. According to the experiments, the multiple rounds of competitive representation has better performance in general than the two-phase deletion. In addition, SCRC adoptively reduces the over-fitting issue of the regression model. However, CSRC also has some disadvantages, such as has not enough robustness to deal with occlusions, disguises, and corruption. In the further, we will pay more attention on these disadvantages.

ACKNOWLEDGMENT

This work was supported partially by the National Nature Science Committee of China under Grant Nos. 61202276 and 61403053 and partially by Chongqing basic and frontier research project under Grant Nos. cstc2014jcyjA40018.

REFERENCES

[1] J. Wright et al., "Robust Face Recognition via Sparse Representation," Ieee Transactions on Pattern Analysis And Machine Intelligence, vol. 31, no. 2, pp. 210-227, Feb, 2009.

[2] M. Yang et al., "Regularized robust coding for face recognition," IEEE Trans Image Process, vol. 22, no. 5, pp. 1753-66, May, 2013.

[3] M. Yang et al., "Robust Sparse Coding for Face Recognition," 2011 Ieee Conference on Computer Vision And Pattern Recognition (Cvpr), pp. 625-632, 2011, 2011.

[4] L. Zhang et al., "Collaborative representation based classification for face recognition," arXiv preprint arXiv:1204.2358, 2012.

[5] A. Y. Yang et al., "Fast ℓ 1-minimization algorithms and an application in robust face recognition: A review." pp. 1849-1852.

[6] M. Yang, and L. Zhang, "Gabor feature based sparse representation for face recognition with gabor occlusion dictionary," Computer Vision–ECCV 2010, pp. 448-461: Springer, 2010.

[7] N. Kwak, "Principal component analysis based on l1-norm maximization," IEEE Trans Pattern Anal Mach Intell, vol. 30, no. 9, pp. 1672-80, Sep, 2008.

[8] J. Wang et al., "Robust face recognition via adaptive sparse representation," IEEE Trans Cybern, vol. 44, no. 12, pp. 2368-78, Dec, 2014.

[9] Y. Xu et al., "A Two-Phase Test Sample Sparse Representation Method for Use With Face Recognition," Ieee Transactions on Circuits And Systems for Video Technology, vol. 21, no. 9, pp. 1255-1262, Sep, 2011.

[10] J.-X. Mi, “Face image recognition via collaborative representation on selected training samples,” Optik - International Journal for Light and Electron Optics, vol. 124, no. 18, pp. 3310-3313, 9//, 2013.

[11] I. Naseem, R. Togneri, and M. Bennamoun, “Robust regression for face recognition,” Pattern Recognition, vol. 45, no. 1, pp. 104-118, Jan, 2012.

[12] I. Naseem, R. Togneri, and M. Bennamoun, “Linear Regression for Face Recognition,” Ieee Transactions on Pattern Analysis And Machine Intelligence, vol. 32, no. 11, pp. 2106-2112, Nov, 2010.

[13] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?." pp. 471-478.

[14] A. M. Martinez, “The AR face database,” CVC Technical Report, vol. 24, 1998.

[15] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, “From few to many: Illumination cone models for face recognition under variable lighting and pose,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 23, no. 6, pp. 643-660, 2001.

[16] K. C. Lee, J. Ho, and D. J. Kriegman, “Acquiring linear subspaces for face recognition under variable lighting,” IEEE Trans Pattern Anal Mach Intell, vol. 27, no. 5, pp. 684-98, May, 2005

Finite Element Analysis based Optimization of Magnetic Adhesion Module for Concrete Wall Climbing Robot

MD Omar faruq Howlader
Robotics and NDT Research Center
London South Bank University
London, UK

Traiq Pervez Sattar
Robotics and NDT Research Center
London South Bank University
London, UK

***Abstract*—Wall climbing robot can provide easier accessibility to tall structures for Non Destructive Testing (NDT) and improve working environments of human operators. However, existing adhesion mechanism for climbing robots such as vortex, electromagnet etc. are still at development stage and offer no feasible adhesion mechanism. As a result, few practical products have been developed for reinforced concrete surfaces, though wall-climbing robots have been researched for many years. This paper proposes a novel magnetic adhesion mechanism for wall-climbing robot for reinforced concrete surface. Mechanical design parameters such as distance between magnets, the yoke thickness, and magnet arrangements have been investigated by Finite Element Analysis (FEA). The adhesion module can be attached under the chassis of a prototype robot. The magnetic flux can penetrate maximum concrete cover of 30 mm and attain adhesion force of 121.26 N. The prototype provides high Force-to-Weight ratio compared to other reported permanent magnet based robotic systems. Both experiment and simulation results prove that the magnetic adhesion mechanism can generate efficient adhesion force for the climbing robot to operate on vertical reinforced concrete structures.**

Keywords—Finite Element Analysis (FEA); Magnetic Adhesion System; Non Destructive Testing (NDT); Wall Climbing Robot

I. Introduction

This research aims to develop a novel adhesion mechanism for concrete wall climbing robot. In this paper, the design parameters and development of a prototype robot are reported. Many of the structures such as high-rise buildings, bridges, dams, complex nuclear power plants etc. are constructed using concrete because of its high strength and durability. Over the lifetime, natural phenomenon like moisture and chlorides present in the atmosphere cause the concrete structures to lose their strength and that induce structural faults like reinforcement bar corrosion, cracking, and delamination of the concrete surface. Currently manual Non Destructive Testing (NDT) using handheld devices are used to detect faults in concrete surface and overall structural integrity. Figure 1 shows the scenario of an operator carrying out manual NDT of a 92 m tall concrete chimney using handheld ground penetrating radar.

Likewise, current method of carrying out NDT process is manual and less efficient as access to test sites on these structures is obtained by constructing scaffolding or abseiling down from the top. Operators deal with hazardous environments such as nuclear industrial environments, working at high altitude, and limited manoeuvrability [2]. In terms of safety, cost and efficiency, climbing robots can offer reliable performance and access to large structures in hazardous environments that may not be accessible to human inspectors.

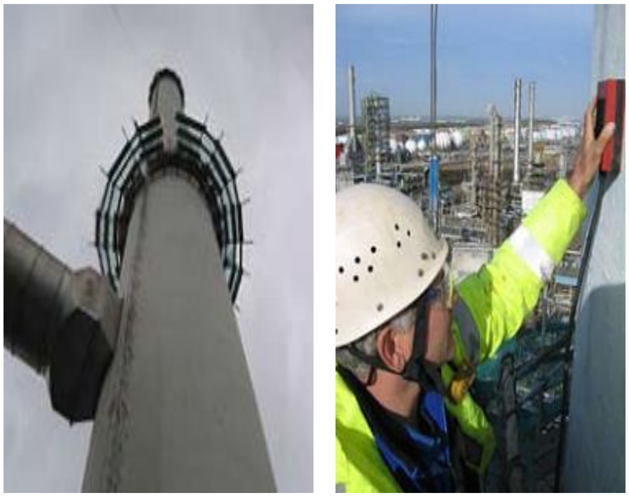


Fig. 1. Manual NDT of tall concrete structure by erecting scaffolding [1]

The proposed prototype robot discussed in this paper is designed to provide robotic platform to attach variety of NDT equipment for concrete wall inspection. The general requirements of the robot mainly include:

a) The robot should have climbing capability of vertical steel surfaces of up to 100 m high and curved surfaces where the minimum diameter is down to 10 m which is common for concrete based nuclear power plants.

b) Keep the robot weight to a minimum as possible. Should the robot fail for any reason, the impact of a heavy robot would cause damage to plant equipment and surroundings.

c) Retain a minimum payload capacity of 2-5 kg for mounting NDT equipment.

d) It should be able to make floor to wall transition and vice versa with a maximum angle of 90^0.

This paper proposes a permanent magnet based adhesion system for concrete surface climbing robot that is distinctive compared to other systems found in the literature. The permanent magnets are arranged in a special way that the magnetic flux is concentrated and magnified to couple with reinforcement bars (rebars) buried under the concrete. The paper consists of the following sections: section II gives an overview of the existing climbing robots. Section III describes the construction method of concrete structures. Dynamic force analysis to establish climbing force requirement and optimization of magnetic adhesion module using Finite Element Analysis (FEA) are given in section IV and V respectively. Section VI presents the experiment setup for adhesion performance verification with comparison of this proposed system with other available systems in section VII and finally, the paper concludes with a brief summary and recommendation in section VIII.

II. Current State of Climbing Robot

Most important considerations in developing wall-climbing robots are its mechanical design, locomotion and adhesion mechanism. Climbing robots need to have the same locomotion mechanism as mobile robots however, the adhesion mechanism significantly affects the system reliability, payload ability, and power consumption of robots. That makes it more challenging to develop a climbing robot than a mobile robot.

To overcome the limiting factors of manual NDT process, application of special climbing robots has been seen in fields such as steel oil tank inspection, weld line inspection, steel pipe structure inspection [3]. There are number of adhesion techniques for climbing robot such as vacuum suction cups, magnets, negative pressure vortex that could be found in the literature. Permanent magnets have been used as the primary adhesion method in designing ferrous wall-climbing robots. Magnetic wheel-type [4], track-type mechanisms [5], legged robots with magnetic feet [6] are used for robot inspecting complex shaped structures. The benefit of wheel-type locomotion is that it can move flexibly with a small contact surface between the wheels and the wall. Therefore, the robot's energy use ratio is low. The track-type mechanism has a larger contact area and can generate higher attraction force however, it is hard to change directions. A non-contact adhesion mechanism has been proposed in [7]. It mounts arrays of magnets under the chassis of the robot. There is a gap between the magnet and the wall surface. However, the effect of multiple magnet layer and the distance between the magnets on total adhesion force are not investigated there. Reference [8] also proposes permanent magnet based system. It investigates the effect of air gap between multiple magnets and the steel surface, the effect of distance between the magnets, effect of magnet dimension on overall adhesion force have not been investigated though. Suction cups and pneumatic adhesion are other popular methods of adhesion. Robots using more traditional suction mechanism are the ROBICEN [9], NINJA-II [10], ROBIN [11]. The Clarifying Climber III [12] is a robot that works by creating negative pressure called vortex between the surface and the robot body. Another robotic system developed for the purpose of concrete NDT is Tile-Wall robot as shown in fig. 2. This includes a mobile module which carries the sensors and NDT devices, a ground platform and a roof platform which work together to form a vertical conveyor belt system to slide the robot up and down [13].

Even though such system solves the difficulty of climbing vertically but it still needs the conveyor system to be installed on the roof and ground. This feature makes this system ineffective in case of very tall structures.

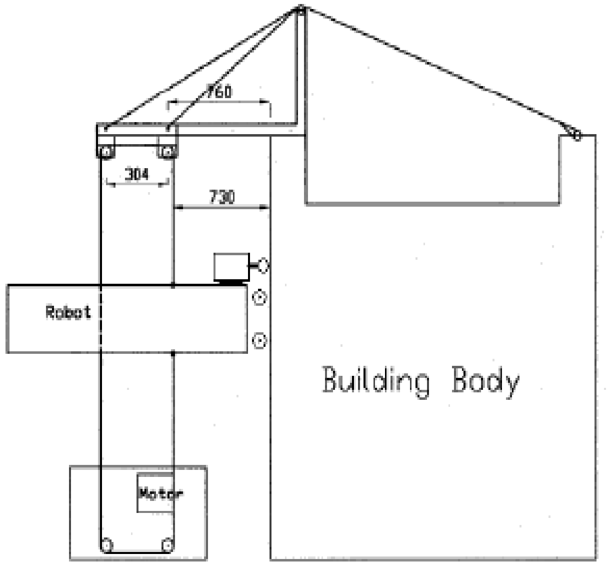


Fig. 2. Block diagram of Tile-Wall robot system [13]

Developing climbing robot for reinforced concrete surface presents a new aspect in adhesion technique as adhesion principles of vortex and electro adhesion are still in development stage and not completely understood as well as suction cups, which operation is limited to smooth surfaces. Moreover, tracked vehicles having two degrees of freedom are less maneuverable. As a result, few practical products have been developed for reinforced concrete surfaces, though wall-climbing robots have been researched for many years.

III. Construction Method of Concrete Structures

Permanent magnets can establish magnetic coupling with ferromagnetic materials present within the magnetic field. All safety critical concrete structures are reinforced with steel rebars. Steel is highly ferromagnetic material therefore, means to establish magnetic adhesion between rebars buried under concrete and magnets could provide a simple, low-energy adhesion mechanism. However, the main challenge when adhering to such structures is to shape the magnetic flux to flow into the concrete as deeply as possible. Hence, it is important to investigate the construction methods of safety critical concrete structures.

Concrete is a brittle material. It has strong compression force but cannot withstand strong tensile stress [14]. As a consequent, cracking on a non-reinforced concrete plane could easily occur if excessive force if applied by external loading as shown in fig. 3. Therefore, raw concrete is reinforced with strong materials such as steel to counter the tensile forces resulting from induced loading. Dense meshing of steel rebars on the vertical wall of a nuclear power plant is shown in fig. 4.

However, a correct rebar positioning is critical. For instance, if rebars are located near the bottom surface of a concrete slab and if the surface is subjected to excessive force applied by the load then this will lead to appearance of cracks on the bottom plate as in fig. 5 and must follow a standard to ensure overall structural strength.

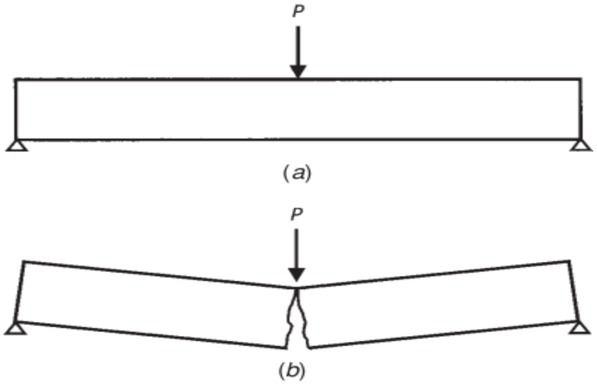


Fig. 3. Cracks on a concrete plane (a) Bending Force < Tensile, (b) Bending Force > Tensile [14]

Fig. 4. Proposed workspace for the concrete wall-climbing robot

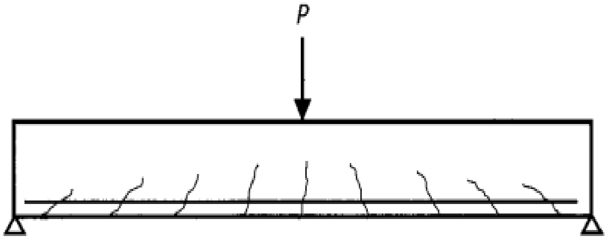


Fig. 5. Appearance of cracks due to inaccurate rebar positioning [14]

The standard applicable in UK that determines the minimum concrete cover requirement is BS 850 and Eurocode 2 [15].

In Eurocode 2, different structures are categorized in different classes such as buildings and power plants are XC3 and XC4 class structures as they are exposed to cyclic wet, dry and moderate humid conditions whereas bridge columns that are submerged in water are XC1 class. The nominal concrete cover is determined based on the type of environmental exposure, concrete quality, and intended working life of the structure. Therefore, the following measurement should apply for nominal concrete cover where an allowance of 10 mm in design deviation is introduced.

$$C_{nominal} = C_{minimum} + 10 \text{ mm} \qquad (1)$$

The nominal cover requirement for atmospheric exposure of 50 years working life extracted from BS 850 is given in table I.

TABLE I. NOMINAL CONCRETE COVER FOR DIFFERENT STRUCTURES BASED ON BS 850

Structures	**Minimum range, mm (Cement quality dependent)**	**Nominal range, mm (Cement quality dependent)**
Buildings (Vertical elements exposed to rain and snow)	25-35	30-45
Bridge Columns	25-35	30-45
Dam	25-50	35-60
Nuclear Power Plant	25-35	30-45

The rebars diameter can vary from 8-55mm. The primary requirement of the adhesion module is to penetrate magnetic flux to a maximum of 60 mm depth to couple with the rebars and generate sufficient adhesion force to ascend on the vertical wall carrying the payload.

IV. DYNAMIC FORCE ANALYSIS

The ideal operation of a climbing robot involves vertical climbing, transitioning and climbing angled slopes. A force analysis could provide insights into stability of the climbing robot and determine minimum adhesion force requirement to avoid sliding and roll-over failure.

A. Sliding avoidance

To understand the forces acting on a robot, consider the forces acting on a robot resting on an inclined plane as shown in fig. 6. If robot weight is = W, distance of front and back wheel center = L, height of the center of gravity = d, acceleration of the robot = a, the coefficient of surface friction = μ, slope of the inclined surface = θ.

$$\sum F_x = W sin\ \Theta - \ \mu N$$

$$N = \frac{W sin\ \Theta}{\mu}$$

$$W cos\ \Theta + \ F_a = \frac{W sin\ \Theta}{\mu}$$

Therefore, for the robot to avoid slipping, required adhesion force should be

$$F_a > \frac{W sin\ \Theta}{\mu} - \ W cos\ \Theta \qquad (2)$$

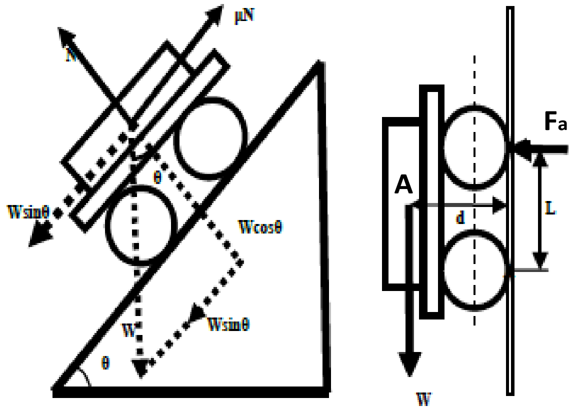


Fig. 6. Free body diagram of robot moving on an angled surface

If the robot is operating on a completely vertical surface then $\Theta = 90^0$ and equation stands as

$$F_a > \frac{W}{\mu} \tag{3}$$

B. Capsizing avoidance

The torque induced by the front set of wheels should be greater than that induced by the robot's weight. So, by taking moment about point A from fig. 6,

$$\sum M = W \times d + F_a \times L = 0$$

$$F_a = -\frac{W \times d}{L}$$

Here negative sign denotes the force direction. To avoid capsizing, the adhesion force should be,

$$F_a > \frac{W \times d}{L} \tag{4}$$

In order to avoid slipping, the center of gravity of the robot should be as close to ground as possible and the distance between the wheels should be large enough. Moreover, increasing the coefficient of surface friction and reducing the robot weight will ensure capsizing avoidance. In addition, the calculations suggest that as long as the robot can avoid slipping, it can avoid capsizing as the force required to avoid capsizing is significantly lower than slipping.

V. Design Optimization of Amgnetic Adhesion Module

The main challenge when adhering to concrete structure using permanent magnet is to shape the magnetic flux to flow into the concrete as deeply as possible to couple with rebars. FEA could be used to investigate the magnetic flux propagation behavior and also measure the resultant adhesion force.

Mathematical model of FEA based magneto-static simulation

Magneto-static models are better understood by Maxwell's equations. If magnetic field intensity is *H*, magnetic induction is *B* then according to Maxwell equation the relations between them in a static magnetic field are [16]:

$$\nabla \times H = J \tag{5}$$

$$\nabla . B = 0 \tag{6}$$

$$B = \mu H \tag{7}$$

$$J = \sigma E \tag{8}$$

Here, *J* is current density, μ is dielectric permeability, *E* is electric field intensity, and σ the conductivity of the medium. The magnetic permeability of ferrous materials such as steel and iron is non-linear and in reverse relation with the magnetic field intensity, *H*. The magnetic field generated the proposed adhesion unit is regarded as static and the electromagnetic field generates a uniform magnetic field effect in this situation. The field vectors and source vectors in Eq. (2) – Eq. (8) are all space co-ordinate functions and do not change with respect with time.

The double rotation equation of the equivalent vector magnetic potential function is:

$$\nabla^2 \times A = \mu J \tag{9}$$

Where, *A* is a vector

In a three-dimensional co-ordinate system, Eq. (6) can be translated into

$$\nabla^2 \times A_x = \frac{\partial^2 A_x}{\partial x^2} + \frac{\partial^2 A_x}{\partial y^2} + \frac{\partial^2 A_x}{\partial z^2} = -\mu J_x \tag{10}$$

$$\nabla^2 \times A_y = \frac{\partial^2 A_y}{\partial x^2} + \frac{\partial^2 A_y}{\partial y^2} + \frac{\partial^2 A_y}{\partial z^2} = -\mu J_y \tag{11}$$

$$\nabla^2 \times A_z = \frac{\partial^2 A_z}{\partial x^2} + \frac{\partial^2 A_z}{\partial y^2} + \frac{\partial^2 A_z}{\partial z^2} = -\mu J_z \tag{12}$$

The FEA divides the solution area into small areas by using tetrahedral meshing method and calculates the resultant force as an integral of Maxwell's surface stress tensor equation as below [17],

$$n_1 T_2 = -\frac{1}{2}(H.B)n_1 + (n_1 . H)B^T \tag{13}$$

Where, n_1 is the boundary normal pointing out from the rebars and T_2 is the stress tensor of air.

A. Optimization of magnetic adhesion system

To understand the design criteria of an adhesion module for reinforced concrete structures, 3D model simulations are carried out using industry standard software Comsol Multiphysics. Here parameters such as distance between multiple magnets, implementation of flux concentrator, variable concrete cover etc. are investigated to determine their effect on resultant adhesion force, F_a. The main properties of the materials used in simulations are listed in table II.

TABLE II. Main Properties of Ndfeb N42

Properties	Value
Magnetic induction intensity B_r (T)	1.31
Coercive force H_{cb} (KA/m)	915
Intrinsic coercive force H_{ci} (KA/m)	955
Magnetic energy product HB (KJ/m^3)	318
Relative permeability (μ_r)	1.068 - 1.113
Relative permeability of yoke (μ_r)	5000
Relative permeability of steel concrete surface (μr)	1
Relative permeability of steel rebar (μr)	1500

The primary focus here is to achieve maximum adhesion force to support the payload by using minimum number of magnets. Therefore, keeping the robot's net weight, W to a minimum will help to increase the performance of the adhesion module. So, the value of magnetic force ratio, η directs the performance of adhesion unit where,

$$\eta = \frac{F_a}{W} \qquad (14)$$

1) Effects of distance between multiple magnets

If two magnets are placed close to each other, then the distance between them affects the distribution of magnetic flux and thus the total adhesion force. To evaluate this scenario, two neodymium grade N42 magnets are attached to a grey cast iron (97% pure) plate called "yoke" with reversed polarity. Dimensions used are magnet length, M_l=50 mm, width, M_w=50 mm, thickness, M_t=12 mm, rebar diameter, R_d=12 mm, yoke thickness, Y_t=10 mm; concrete cover, C_c=30 mm. The distance between the two magnets are varied from 10 mm to 150 mm and the magnetic flux density norms are observed from simulations. The model setup is given in fig. 7.

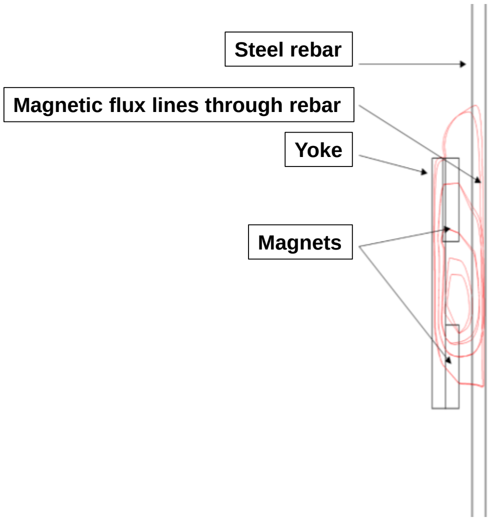
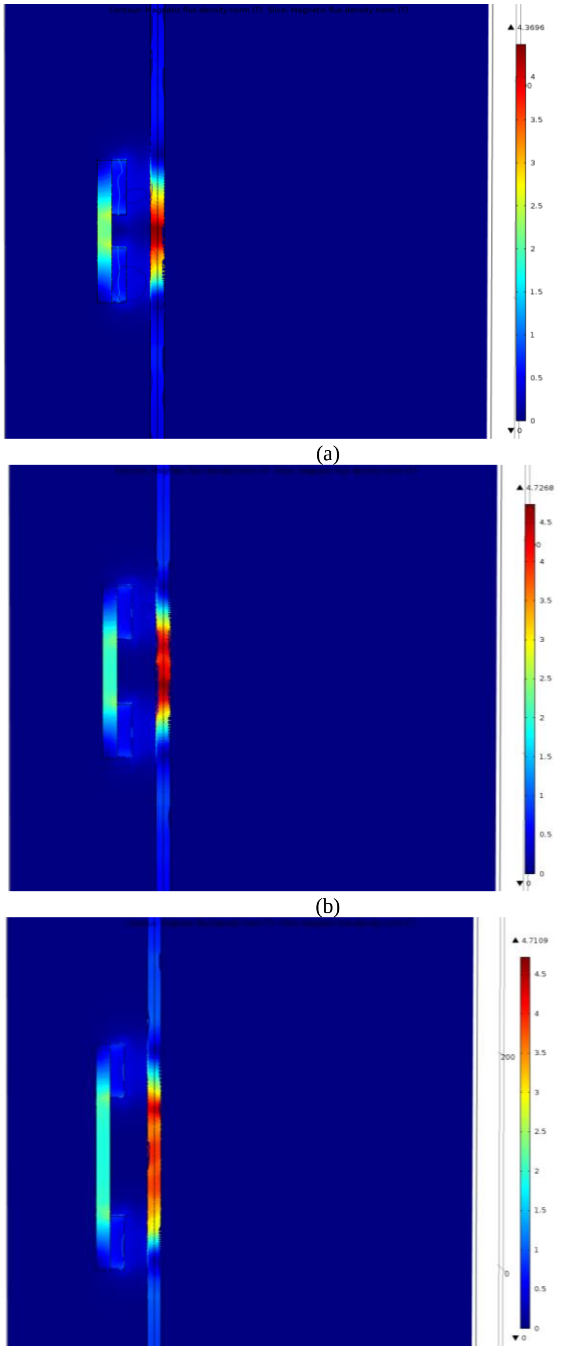


Fig. 7. Distribution of magnetic flux lines in the magnetic circuit travelling from North Pole of one magnet to South Pole of another magnet

The color graphs in fig. 8 represent the magnetic flux concentration norm along the rebar length. It is observed that as the distance between the magnets is increased from 10mm, the flux concentration area increases shown by the red hotspot on the rebar.

As two magnets are closer to each other, the resultant magnetic attraction area is small. But as the distance increases to an optimum level of 50-60 mm, magnetic flux leaving North Pole of one magnet have to travel longer to couple with South Pole of another magnet to complete the magnetic circuit between two magnets and as a result, total active area of attraction becomes bigger which in turns increases the adhesion force. Bigger red hotspot area in fig. 8(b) denotes the optimum distance between the magnets. At distance greater than 60mm, magnetic flux density reduces as represented by bigger but low intensity hotspot for 150 mm magnet distance in fig. 8(c).

According to simulation results in fig. 9, adhesion force increases gradually as the distance between the magnets increases. At 50 mm distance between magnets, adhesion force is maximum at 64.93 N. This is a 33% increase compared to adhesion force achieved at 10 mm distance between magnets. However, adhesion force falls sharply as the distance is increased above 50 mm. At 150 mm distance, adhesion force reaches as low as 51.6 N. These measurements support the magnetic flux density norm presented.



(a)

(b)

(c)

Fig. 8. Magnetic flux concentration norm at different distance between magnets; (a) Distance: 10mm, (b) Distance: 50mm, (c) Distance: 1500mm

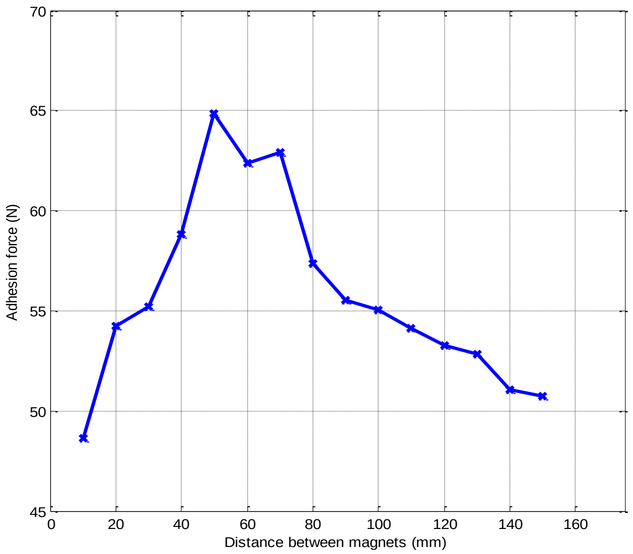
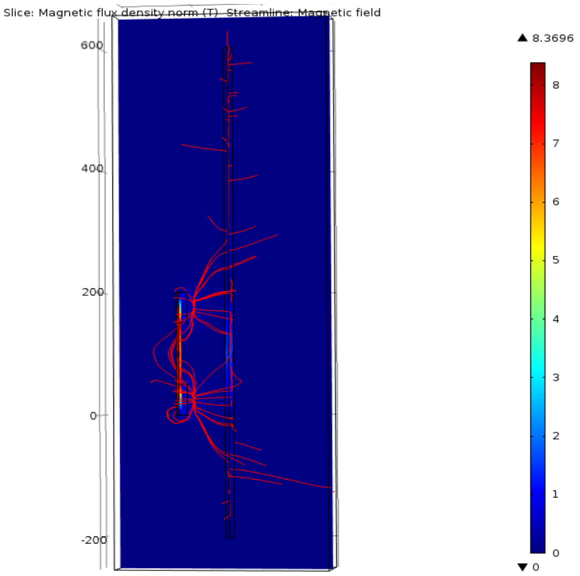


Fig. 9. Simulated adhesion force at different distances between magnets

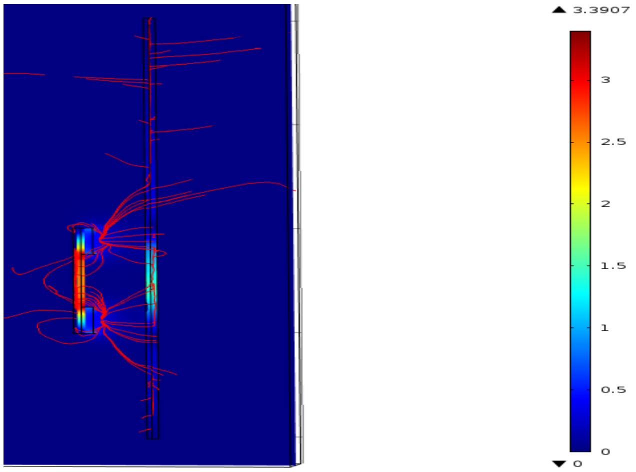
2) Effects of yoke thickness on adheison force

Previous simulations show that the use of yoke significantly increases the adhesion force. Using only two magnets placed at 50 mm distance between them without any yoke results in adhesion force of only 39.34 N for concrete cover of 30 mm. But implementing a yoke of 5 mm thickness increases the adhesion force significantly to 53.71 N. Therefore, thickness of yoke plays a critical part to maximize adhesion force. To investigate this parameter further, models with varied yoke thickness from 1 mm to 40 mm are simulated. All other parameters are kept the same as previous simulations.

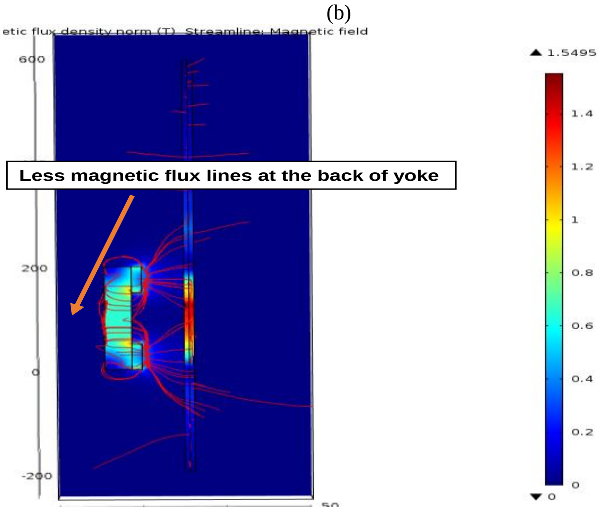
The adhesion force found to be increasing significantly when the yoke thickness is increased from 1 mm to 20 mm. A yoke of 1 mm thickness does not have any apparent effect on adhesion force as the flux leakage is higher in all direction as shown in magnetic flux density norm in fig. 10(a).


Slice: Magnetic flux density norm (T) Streamline: Magnetic field
600
400
200
0
-200
▲ 8.3696
8
7
6
5
4
3
2
1
0
▼ 0

(a)



(b)



Less magnetic flux lines at the back of yoke

(c)

Fig. 10. Magnetic flux density norm at yoke thickness; (a) 1 mm, (b) 5 mm, (c) 20 mm

Any influence of yoke thickness appeared when the thickness is increased to 5 mm, the system achieves magnetic coupling with rebar as presented in fig. 10(b). But at 20 mm thickness, the flux concentration is at maximum and reaches maximum adhesion force of 65.13 N. Moreover, the magnetic flux leakage is lower in the yoke’s opposite surface and more flux lines are concentrated toward the rebar shown as red hotspot in fig. 10(c). According to force against yoke thickness graph in fig. 11, the adhesion force comes to a saturation point of approximately 62 N at 25mm yoke thickness, where further increase in yoke thickness would not have any significant influence on adhesion force. Moreover, the force-to-weight ratio, Ƞ of the adhesion module falls sharply because of the added weight by increased yoke thickness. The primary reason for this is that, magnetic flux lines prefer to pass through medium of high permeability. When the yoke thickness exceeds an optimum point (in this case 25 mm), the magnetic flux lines prefer to pass through the yoke of high permeability rather than air as in figure 12. Thus, the magnetic coupling with the rebars is minimum to influence the adhesion force.

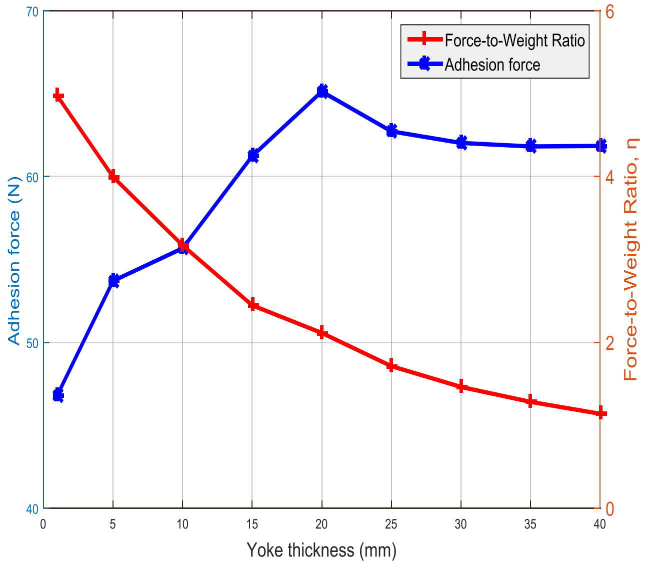


Fig. 11. Adhesion force for yokes with different thicknesses

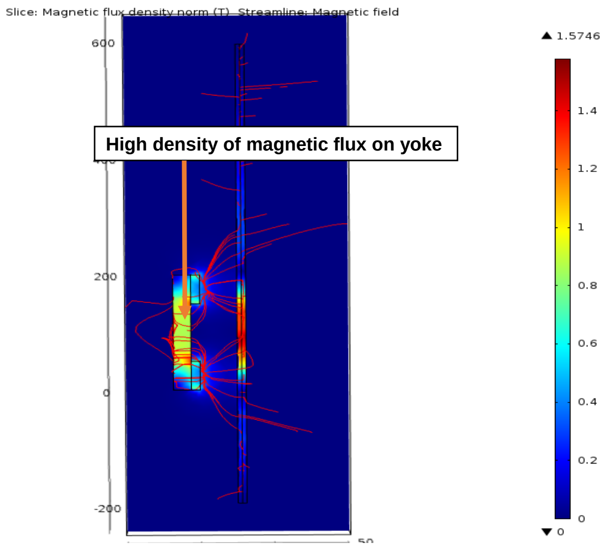

Slice: Magnetic flux density norm (T) Streamline: Magnetic field
High density of magnetic flux on yoke
600
200
0
-200
▲ 1.5746
1.4
1.2
1
0.8
0.6
0.4
0.2
0
▼ 0

Fig. 12. Density of magnetic flux on a 25 mm thick yoke

Therefore, the implementation of yoke ensures magnetic flux concentration between the yoke and steel rebars. As a result, attraction force is increased. It also acts as a magnetic shield and reduces the magnetic field strength at the back, so there is less magnetic interference with electronic control circuits.

3) Optimization analysis of magnet arrangement layout

The influence of the magnet and yoke dimensions on the performance of the adhesion module was analyzed in the previous sections. As the adhesion module is made of multiple magnets so, the coupling between them will influence the adhesion force. For the previous two sets of simulation, North-South configuration using two magnets was implemented. Such configuration essentially creates one magnetic circuit travelling from North Pole of one magnet to South Pole of the other magnet. Yet for deeper penetration of magnetic flux lines, a configuration of North-South-North using three magnets could be considered as fig. 13.

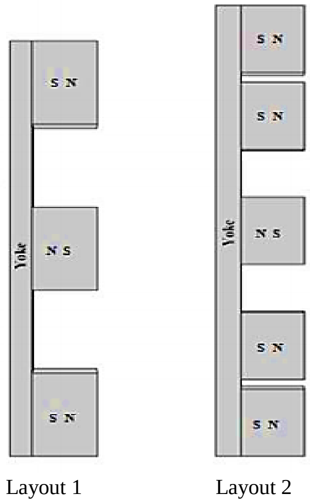


Fig. 13. Proposed N-S-N and N-N-S-N-N magnet arrangement layouts

In this case, adding one magnet on the system will create two separate magnetic circuits and thus might increase the overall adhesion force. Another configuration of North-North-South-North-North could be considered. This arrangement will also effectively create two magnetic circuits however, the magnetic flux lines have to travel further from North Pole to South due to the fact that two identical poles are located near to each other. As a consequent, a deeper dispersion of magnetic flux lines could be accomplished. For simulations, magnet dimension is kept similar to previous cases and yoke thickness of 15 mm is considered.

Simulation results show that adding one extra magnet into the system greatly increases the module's ability to concentrate the magnetic flux towards the buried rebar and thus can generate greater adhesion force. The maximum adhesion force is measured 113.19 N for N-S-N layout compared to 65.13 N in the previous case at 30 mm concrete cover. According to magnetic flux density norm from Fig. 14, magnetic flux for N-N-S-N-N does not concentrate as uniformly as N-S-N layout. Layout 2 essentially creates two magnetic circuits but as two North Poles are located in close proximity at both ends of the adhesion module, it actually distort the magnetic flux uniformity. Moreover, in case of layout 2, having two same poles closer to each other ensure that magnetic flux lines travel a longer distance to meet the opposite pole compared to layout 1.

However, adhesion force comparison result between two layouts shows that, layout 2 produces higher adhesion force than layout 1 for different concrete cover. At 10 mm concrete cover, resultant adhesion force is 312.57 N and 451.33 N respectively for layout 1 and 2. But as the concrete cover is increased to 20 mm, adhesion force reduced to 188.02 N and 248.17 N respectively. So this is a reduction of 39.84% for layout 1 and 45% for layout 2. The rate at which adhesion force falls as the rebar distance increases is marginally lower for layout 1 than layout 2. Moreover, the influence of weight of the additional magnets of N-N-S-N-N arrangement is cancelled out by generating higher adhesion force which means layout 2 will have higher Ƞ ratio compared to layout 1 as fig. 16.

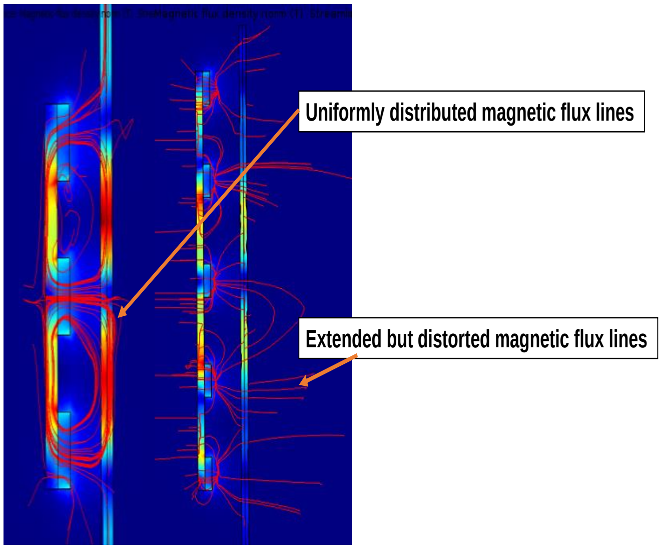

Uniformly distributed magnetic flux lines
Extended but distorted magnetic flux lines

Fig. 14. Comparison of magnetic flux line propagation behaviour of layout 1 and layout 2 from left to right

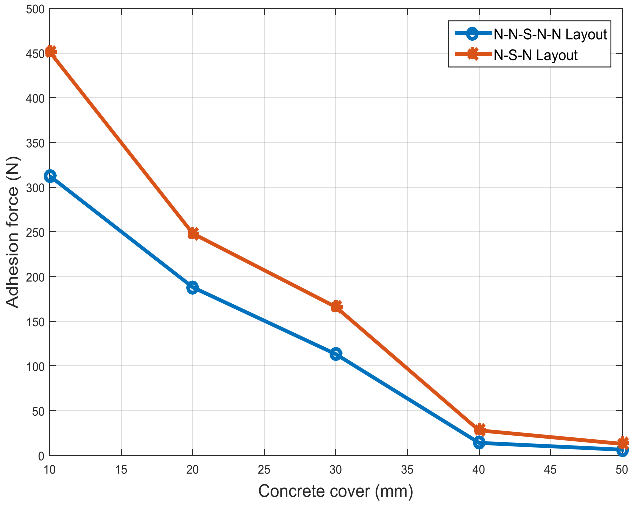


Fig. 15. Comparison of adhesion force between layout 1 and layout 2 for different concrete cover

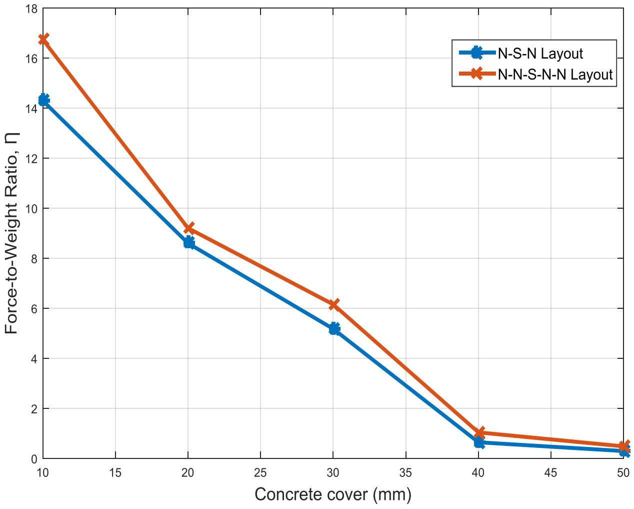


Fig. 16. Force to weight graph for two layouts

VI. Experiment Setup and Result Discussion

In order to validate the simulation results, a magnetic adhesion system consisting of three N42 grade neodymium magnets arranged in N-S-N orientation was built and attached to a prototype climbing robot. The measured dimensions of the whole system are: robot length = 360 mm, robot width = 210 mm, height of the center of gravity = 15 mm, magnet length and width = 50 mm, magnet thickness = 10 mm, yoke length = 350 mm. Two yokes with thickness of 5 mm and 15 mm as shown in figure 16 (a) were used for better comparison. The gap between the magnet surface and climbing surface is critical as a small gap increases the adhesion force significantly. Therefore, the gap was kept to a minimum of 2 mm. The ability to pass obstacles has been considered to be of secondary importance since the aimed climbing structures are closely uniform. Each wheel is independently driven and a differential drive system is adopted to realize the turn. The wheel diameter is 63mm. Rubber wheel with polyurethane layer is chosen to increase the traction. The output torque and rotation of the motor are 2.16 Nm and 30 rpm respectively. The maximum speed of the robot is 6 meter per minute. The robot's net weight is 2.23 kg and 3.68 kg when 5 mm and 15 mm thick yoke is used respectively. Therefore, the required force for sliding avoidance can be obtained as 46 N and 76 N for 5 mm and 15 mm yoked robot respectively by using equation 3 and 4 when acceleration, a = 0.5 m/s^{-2} and wheel friction coefficient, μ = 0.5.

Two FlexiForce A201 force sensors, capable of measuring force up to 445N was mounted underneath the base of the motors situated on the top right and bottom left corner of the robot. As the attraction area of the adhesion module is uniform, therefore by taking the average value of the two sensors will give the actual adhesion force at a given test point. A 10-bit analogue- to-digital AVR microcontroller based embedded system module as shown in fig. 17 was used to capture and serial transmission of the analogue data and MATLAB was used for data analysis. An operator can access the on board control module via Bluetooth wireless communication system. Furthermore, the on- board devices are powered by four 1.5 V Lithium-ion Polymer (Li-Po) batteries that supply enough operating voltage for the motors and force measurement devices. As a result, the robot is totally umbilical free. The adhesion force of the prototype robot was measured for different yoke thickness and different rebar distance. Measurements were taken at three different test points along the height of a concrete column and the average value was taken for consideration. Fig. 18 shows a working prototype of the robot climbing a concrete wall of a building. The concrete cover was measured to be 20 mm using a rebar detector. Meanwhile, experiments of the wall climbing robot carrying a magnetic field measurement sensor and climbing a concrete column of a multistoried building also carried out. Weight of the magnetic hall-effect sensor and control unit was measured to be 1.6 kg. The maximum adhesion force was measured 181.08 N using a 10 mm thick yoke.

Experiment vs simulation results shows when 5 mm thick yoke was used, the adhesion force reduced from 162.06 N to 4.0221 N as the concrete cover changed from 20 mm to 50 mm. Using a 10 mm thick yoke produced slightly higher adhesion force for concrete cover from 20 mm to 35 mm. However, if concrete cover is bigger than 35 mm, then adhesion force for 10 mm thick yoke decrease sharply compared to 5 mm thick yoke as fig. 19 (a) and (b).

Adhesion force for different yoke thickness at a constant concrete cover of 30 mm also followed the same results achieved in simulations. For 5 mm thick yoke, the adhesion force was 121.26 N which increased to 160.17 N as the thickness was increased to 40 mm. However, increased weight of the yoke had a reverse relationship to the force-to-weight ratio. Fig. 20 (b) shows a gradual fall of Ƞ from 7.413 to 0.747 as the yoke thickness increased from 20 mm to 50 mm. Therefore, yoke thickness of 20 mm should be an optimum design trade-off.

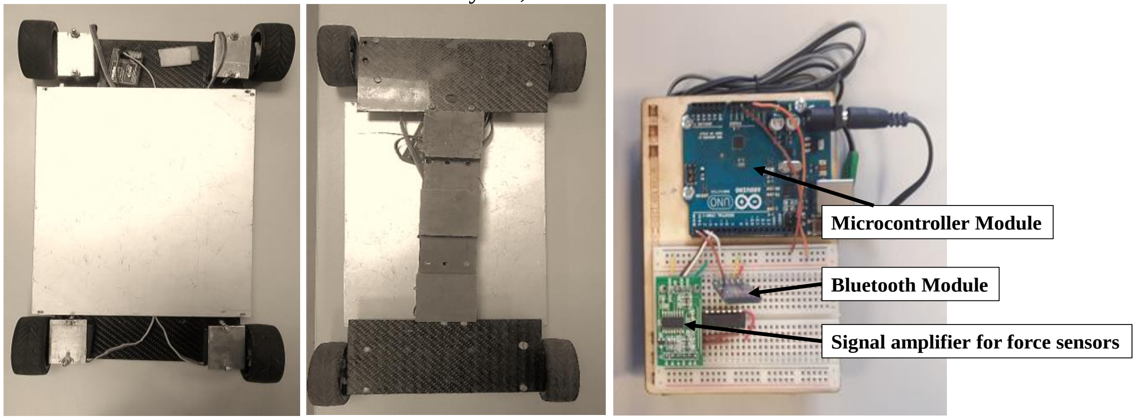


Fig. 17. Prototype of the concrete wall-climbing robot (top and bottom view) and adhesion force measurement unit circuit

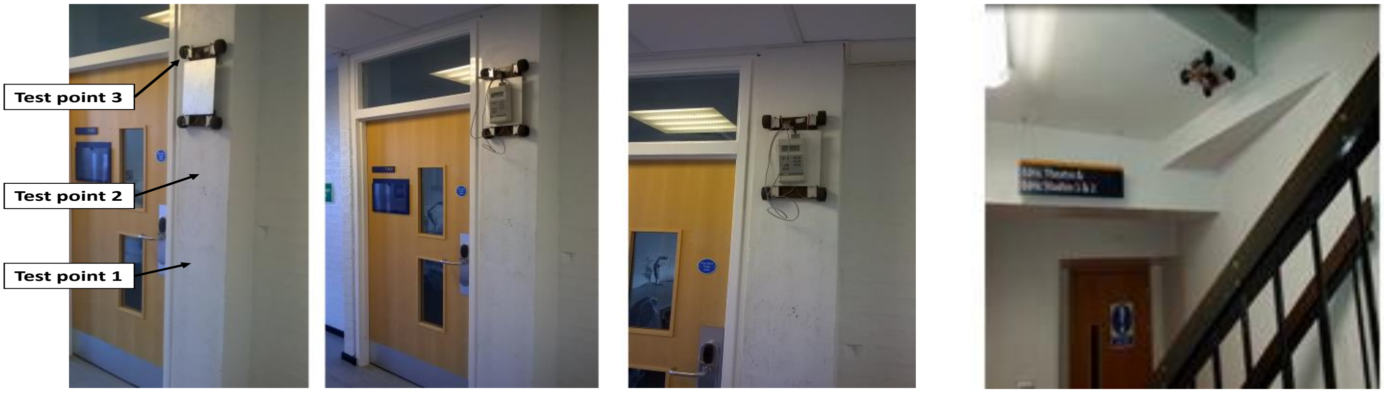


Fig. 18. Prototype robot climbing concrete column and ceiling with on board magnetic field sensor

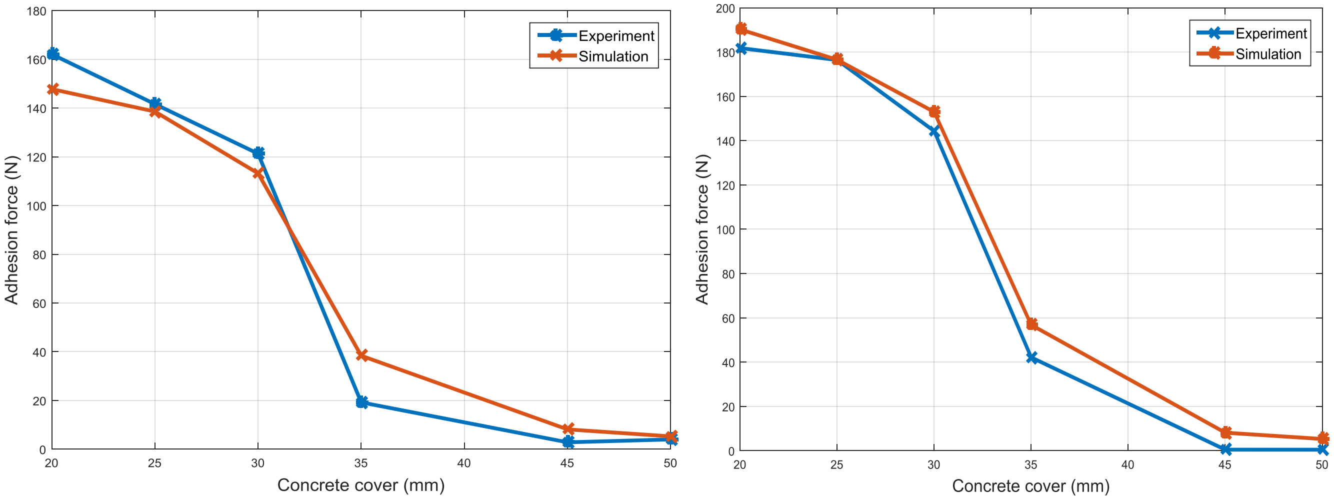


Fig. 19. Comparison of results at different concrete cover for (a) 5mm thick yoke; (b) 10mm thick yoke

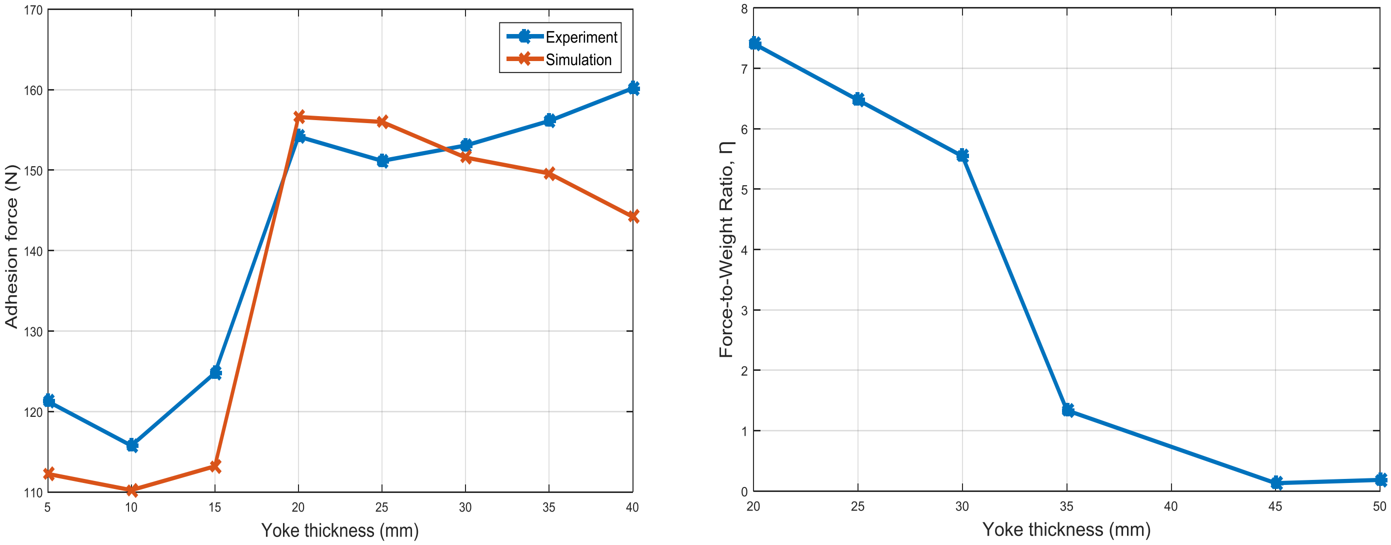


Fig. 20. (a) Experiment vs Simulation results for different thickness of yoke; (b) Experiment results for Force-to-Weight ratio, η for different yoke at 30mm concrete cover

VII. Comparison Study

This work is proposing permanent magnets based adhesion mechanism for concrete surface climbing robot. There are several literatures could be found that use permanent magnets mainly for ferrous wall climbing and other biological models such as adhesive legs or nano-spike gecko-bots for concrete wall climbing. However, those systems are not still suitable for real world application. Moreover, many of those robots are bulky, resulting in low force-to-weight ratio, Ƞ. A comparison study is of this proposed robot with other literature found is presented in table III.

TABLE III. Comparison of the Proposed Robotics Adhesion System with Other Sytems Found in the Literature

	Ref [7]	**Ref [8]**	**This work**
Robot weight (kg)	30	18	2.23
Maximum applicable adhesion force (N)	1400	667	121.26
Force-to-weight ratio	4.76	3.77	5.54
Adhesion mechanism	Magnetic	Magnetic	Magnetic
Locomotion	Wheel	Track	Wheel
Application medium	Ferrous wall	Ferrous wall	Concrete wall

VIII. Conclusion

A novel technique of magnetic flux concentration mechanism has been presented. In contrast to existing research findings, this research aims to manipulate physical and material constrains to apply such adhesion system on a concrete surface where the attraction medium for permanent magnets is limited. The simulation results shows that the distance between the magnet and buried rebar has the maximum impact on the total adhesion force. The prototype system exhibits good payload capacity and can generate adhesion force of 121.26 N for the robot to climb a concrete wall reinforced with only one rebar located as far as 30 mm. Considering the size and weight of the robot, a high payload capacity has been achieved by optimizing key design parameters of the adhesion mechanism. Overall, this mobile robot exhibits very good performance and force-to-weight ratio compared to other reported permanent magnet based robotic systems. Though 12 mm diameter rebar was used for simulations, rebars used is much thicker (55 mm in some cases) in its target application such as nuclear power plant and bridge column. Therefore, the adhesion force will be much higher in those conditions. Moreover, nearby rebars in a rebar mesh will also increase the adhesion force. To turn the robot into a practical application, a higher grade (N52) of neodymium magnets could be used that will increase adhesion while keeping the force to weight ratio as same as N42 grade magnets. Nonetheless, the proposed methodology will be the foundation of further research. Future research will involve investigating the effects of a rebar mesh on adhesion force as well as carrying out further testing in real life applications.

References

[1] GSSI Case study, “A Tall Order: Scanning a 300-feet Chimney”, Geophysical Survey Systems, Inc., Salem, USA, 2009.

[2] B. Zhiqiang, G. Yisheng, C. Shizhong, Z. Haifei, and Z. Hong, “A Miniature Biped Wall-Climbing Robot for Inspection of Magnetic Metal Surfaces”, IEEE International Conference on Robotics and Biomimetics, 2012, Guangzhou, China.

[3] T. Sattar, E. Leon-Rodriguez Hernando, S. Jianzhong, “Amphibious NDT Robots”, Climbing and Walking Robots, Towards New Applications, International Journal of Advanced Robotics Systems, Chapter 6, pp. 24, 2007.

[4] W. Fischer, F. Tache, G. Caprari, and R. Siegwart, “Magnetic wheeled robot with high mobility but only 2DOF to Control”, International Conference on Climbing and Walking Robots, 2008, Portugal.

[5] F. Tache, W. Fischer, G. Caprari, R. Moser, F. Mondada, R. Siegwart, “Magnebike: a magnetic wheeled robot with high mobility for inspecting complex shaped structures”, Journal of Field Robotics, Vol. 26, pp. 453-76, 2009.

[6] T. Kang, H. Kim, T. Son, H. Choi, “Design of quadruped walking and climbing robot”; IEEE/RSF International Conference on Intelligent Robots and System, 2003, Vol. 1, pp. 619-24.

[7] J. Shang, B. Bridge, T. Sattar, S. Mondal, and A. Brenner, “Development of a climbing robot for inspection of long weld lines”, Industrial Robot, Vol. 35, No. 3, pp. 217-223, 2008.

[8] P. Kalra, G. Jason, M. Max, “A wall climbing robot for oil tank inspection”, IEEE International Conference on Robotics and Biomimetics, China, 2006, pp. 1523-1528.

[9] L. Briones, P. Bustamante, M. Serna, “ROBICEN: a wall-climbing pneumatic robot for inspection in nuclear power plants”, Robotics and Computer-Integrated Manufacturing, pp. 287–292, 1994.

[10] A. Nagakubo, S. Hirose, “Walking and running of the quadruped wall-climbing robot”, IEEE International Conference on Robotics and Automation, pp.1005–1012, 1994.

[11] R. Pack, J. Christopher, K. Kawamura, “A Rubbertuator-based structure-climbing inspection Robot”, IEEE International Conference on Robotics and Automation, pp. 1869–1874, 1997.

[12] X. Chen, M. Wager, M. Nayyerloo, W. Wang, J. Chase, “A novel wall climbing Robot based on Bernoulli effect”, International Conference on Mechatronic and Embedded Systems and Applications, Beijing, China, 2008.

[13] S. Tso, T. Feng, “Robot Assisted Wall Inspection for Improved Maintenance of High-Rise Buildings”, International Symposium on Automation and Robotics in Construction, China, 2001, pp 63-71.

[14] D. Fanella, “Design of Low-Rise Reinforced Concrete Buildings Based on 2009 IBC”, ASCE/SEI 7–05, ACI 318–08. International Code Council, Washington, DC.

[15] European Standard. Eurocode 2, “Design of Concrete Structures, General Rules”.

[16] W. Xilin, J. QIU, T. Junyong, “Analytical dynamics and application in Electromechanical System”, Magazine of Science press, Beijing, pp. 212-213, 2003.

[17] P. Yao, D. Li, “The magnetic field analysis and optimization of permanent-magnetic adhesion device for a novel wall-climbing robot," Proc. of the IEEE Conference on International Technology and Innovation, pp. 2-6, 2009.

Facial Age Estimation based on Decision Level Fusion of AAM, LBP and Gabor Features

Asuman Günay and Vasif V. Nabiyev
Department of Computer Engineering
Karadeniz Technical University
Trabzon, TURKEY

***Abstract*—In this paper a new hierarchical age estimation method based on decision level fusion of global and local features is proposed. The shape and appearance information of human faces which are extracted with active appearance models (AAM) are used as global facial features. The local facial features are the wrinkle features extracted with Gabor filters and skin features extracted with local binary patterns (LBP). Then feature classification is performed using a hierarchical classifier which is the combination of an age group classification and detailed age estimation. In the age group classification phase, three distinct support vector machines (SVM) classifiers are trained using each feature vector. Then decision level fusion is performed to combine the results of these classifiers. The detailed age of the classified image is then estimated in that age group, using the aging functions modeled with global and local features, separately. Aging functions are modeled with multiple linear regressions. To make a final decision, the results of these aging functions are also fused in decision level. Experimental results on the FG-NET and PAL aging databases have shown that the age estimation accuracy of the proposed method is better than the previous methods.**

Keywords—AAM; LBP; Gabor filters; Regression; Fusion; Age estimation

I. INTRODUCTION

The researches on facial image processing have received considerable interest in recent decades because of the increasing need of automatic recognition systems. Face recognition, face detection, facial expression recognition and gender classification are the research topics that have been studied by many researchers in this area. Facial age estimation is a relatively new topic and the interest in this topic has significantly increased because it has many real world applications. For example, under ages can be prevented from accessing alcohol, cigarettes or obscene contents on websites using an age estimation system. In addition, age specific target advertising, face recognition and age prediction systems robust to age progression for finding the missing people and criminals are important age estimation applications.

Facial age estimation is a multi-class classification problem because an age label can be seen as an individual class. This makes age estimation much harder than other facial image processing problems such as gender classification, face detection, etc. Besides, real world age progression displayed on faces is varied and personalized as shown in Fig. 1. Aging process of a person is affected by the genetics, race, eating and drinking habits, living styles, climate, etc. [1]. Extent and frequency of facial expressions, emotional stress, exposure to sunlight, extreme weight loss, smoking, usage of anti-aging products, and plastic surgery also affect the person’s facial appearance [2]. Therefore, determining the type of facial features that represents the age directly is very difficult. Moreover, the accuracy of age estimation systems are insufficient, even the human skills about age estimation are limited. The lack of proper large data set including the chronological image series of individuals is another drawback in these systems.

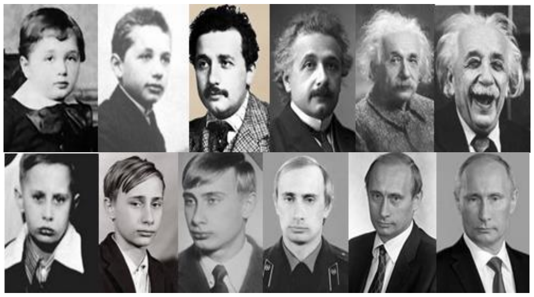


Fig. 1. Facial aging of different individuals

In this paper a new hierarchical age estimation method based on decision level fusion of global and local features of facial images both in age group classification and age estimation phases is proposed as shown in Fig. 2. The global facial features which contain both the shape and appearance information of human faces are extracted using active appearance models (AAM). The local facial features are extracted using Gabor filters and local binary patterns (LBP). A set of Gabor filters capable of extracting deep and fine wrinkles in different directions are used to extract wrinkle features and LBP is used to extract the detailed skin texture features of facial images. Then dimensionality reduction is performed using principal component analysis (PCA) for each feature vector separately. After finding the lower dimensional subspaces, three distinct support vector machine (SVM) classifiers are trained using global features, wrinkle features and skin features of facial images. Then the results of these classifiers are combined to find the age group of the subject. After that, age estimation is performed in that age group in a similar way, in which three aging functions are modeled with global and local features, separately using multiple linear regression. Finally the results of these aging functions are combined to estimate the age of the subject.

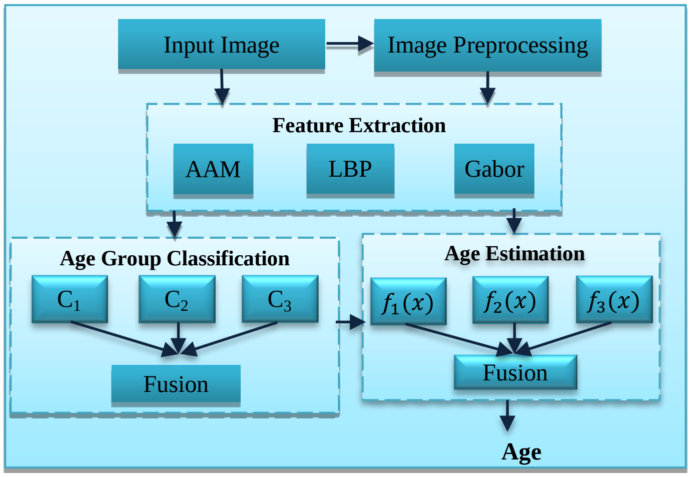


Fig. 2. System structure

This paper is organized as follows. A survey on age estimation methods is given in Section 2. In Section 3 the proposed method including image preprocessing, global and local feature extraction, dimensionality reduction, classification, regression and decision level fusion is introduced. The experimental results are given in Section 4 and Section 5 concludes the paper.

II. Age Estimation Methods

Over the years a great number of approaches have been proposed in the field of age estimation from facial images. These approaches typically consist of age image representation and age estimation techniques. Age image representation techniques rely often on shape and texture-based facial features. They can be grouped under the topics of Anthropometric Models, AAM, AGing pattErn Subspace (AGES), Age Manifold and Appearance Models. Then, age group classification or regression methods are performed for age estimation. Recently, hierarchical age estimation systems combining the classification and regression techniques are presented [3]. To build a universal human age estimator, robust multi-instance regressor learning algorithm based on facial information is also used [4].

In anthropometric models only the facial geometry is considered. Kwon and Lobo published the first work using the facial geometry in the age classification area [5]. They separated the babies from adults using a few ratios of distances on frontal images. However the shape of a human's head significantly changed in the childhood, but not in the adulthood [6]. For this reason their method can be successful in young ages. So the wrinkle information is used to classify the young and senior adults. In the experiments a small database including 45 images is used. Later on other age classification methods using geometrical features based on distance ratios and texture features based on wrinkles are also proposed [7, 8]

AAM [9, 10] based methods incorporate shape and appearance information together rather than just the facial geometry as in the anthropometric model based methods. An AAM uses a statistical shape and an appearance model to represent the images [10]. These models are generated by combining a model of shape variations with a model of the appearance variations in a shape-normalized frame. A statistical shape model can be generated with a training set of face images labeled with landmark points. The mean shape is produced with taking the mean of the landmark points in the training set. Then Principal Component Analysis (PCA) is applied to the data to extract the main principal components along which the training set varies from the mean shape. To build a statistical appearance model, each image has to be normalized, so that its control points match the mean shape. Then, PCA is applied to the gray-level intensities within a pre-specified image region for learning an appearance model. Using the AAMs for age estimation was initially proposed by Lanitis et al. [11]. The relationship between the age of individuals and the parametric description of face images was defined with an aging function $age = f(b)$. Kohli et al. [12] used ensemble of classifiers trained using AAM features to distinguish between child/teenhood and adulthood. Also different aging functions are used to estimate the age of the classified image. Chao et al. [13] proposed an age estimation approach based on label sensitive learning and age-oriented regression using AAM features.

Geng et al. [14, 15] proposed a method called AGES which uses the sequence of an individual's facial images arranged in chronological order to model the aging process. The features of face images are extracted with AAM. Then, PCA is used to learn a specific aging subspace for each individual. In AGES method missing age images of individuals can be synthesized with an expectation maximization-like iterative algorithm.

Age manifold methods intend to learn a common aging trend from the images of different individuals at different ages. The aging trend is learned in a low dimensional domain using manifold embedding techniques. The mapping from the image space to the low dimensional manifold space can be done either by linear or by nonlinear functions [16-21] such as Y=P(X, L). In this representation $X = [x_i : x_i \in R^D]_{i=1}^n$ is the image space, $L = [l_i : l_i \in N]_{i=1}^n$ is the vector contains the age labels associated with images and $Y = [y_i : y_i \in R^d]_{i=1}^n$ with $d \leq D$ is the low-dimensional representation of X in the embedded subspace. In age manifold methods all aging images of different individuals can be used together. But the size of the training data set should be large enough in order to learn the embedded manifold with statistical sufficiency.

Appearance models are mainly focused on the extraction of global and local aging-related facial features. Fukai et al. [22] extracted aging features from facial images using Fast Fourier Transform. Then the important features are selected using genetic algorithm. As the Local Binary Patterns (LBP) are efficient texture descriptors [23], they are used in age estimation systems. In Ju and Wang's study [24] regions which vary with aging are selected using Adaboost. Then LBP histograms are extracted from these regions and used for age estimation. Wrinkle information extracted with Gabor filters have also been used as effective texture features on age estimation tasks. [25-28].

III. Proposed Method

This paper proposes an innovative hierarchical age estimation method based on decision level fusion of global and local facial features. This method consists of the image preprocessing, global feature extraction with AAM, local feature extraction with Gabor filters and LBP, dimensionality reduction with PCA classification with SVM, aging function modeling with multiple linear regression and decision level fusion modules. These modules are explained in the following sections.

A. Image Preprocessing

The orientation and the size of original images are different from each other. Also they have unnecessary features such as background, cloth and hair which are not related to the face and can affect the performance of the algorithm. Therefore, image preprocessing step is performed to extract only the facial regions and to adjust the size and the orientation of the faces. In the this module, the facial images are cropped, scaled and transformed to the size of 88x88, based on the eye center locations as shown in Fig. 3.

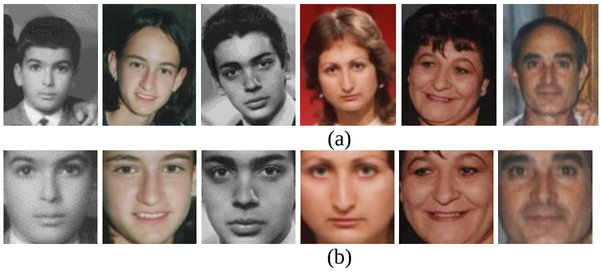
B. Feature Extraction

The feature extraction module consists of two modules: global feature extraction with AAM, local feature extraction with Gabor filters and LBP. These modules are explained in the following subsections.

1) Global Feature Extraction with AAM: AAM is a statistical shape and appearance model of facial images [10]. In AAM , a model of shape variations is combined with a model of the appearance variations in a shape-normalized frame

Training samples which are labeled with landmark points are used to generate a statistical shape model. The landmark points of various facial images are given in Fig. 4. Let $X = [x_i : x_i \in R^{D_s}]_{i=1}^{n}$ represents all the landmark points of training images and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ represents the mean shape of training images; the main principal components along which the training set varies from the mean shape is extracted with PCA. The projection is chosen to maximize the determinant of the total scatter matrix $S_s = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$of the projected samples as $\arg\max_{P_s}\left|{P_s}^T S_s P_s\right|$. P_s is the set of eigenvectors with d_s largest eigenvalues which provides a linear transformation from D_s dimensional shape space into a $d_s \le D_s$ dimensional parameter space. The shape parameters are defined by linear formulation as $b_s = {P_s}^T X$.

To build a statistical appearance model, each image has to be warped to mean shape as shown in Fig. 4. Then, the gray-level intensities within a pre-specified image region are used to train an intensity model. Let $G = [g_i : g_i \in R^{D_g}]_{i=1}^{n}$ represents all the gray level intensities of training images, the main principal components along which the training set varies from the mean appearance is also extracted with PCA. The projection is chosen to maximize the determinant of the total scatter matrix $S_g = \sum_{i=1}^{n}(g_i - \bar{g})(g_i - \bar{g})^T$of the projected samples as $\arg\max_{P_g}\left|{P_g}^T S_g P_g\right|$. P_g is the set of eigenvectors with d_g largest eigenvalues which provides a linear transformation from D_g dimensional appearance space into a $d_g \le D_g$ dimensional parameter space. The appearance parameters are defined by linear formulation as $b_g = {P_g}^T G$. b_s and b_g vectors can summarize the shape and appearance of any image. The combined shape-appearance parameters are obtained by concatenating b_s and b_g in a single vector and applying a further PCA in order to eliminate the correlations between them.


(a)

(b)

Fig. 3. Image preprocessing (a) Original images (b) Facial regions

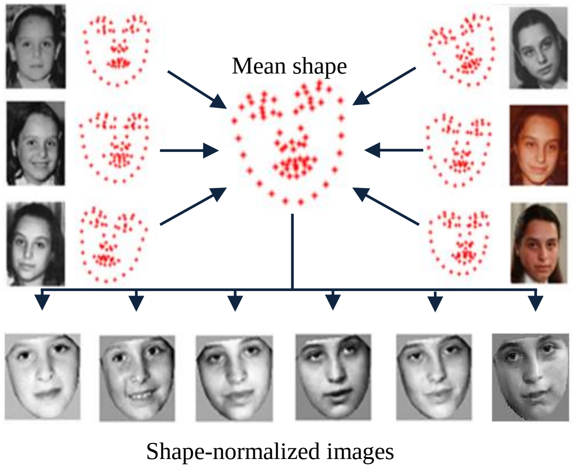


Fig. 4. Landmark points, mean shape and normalized images used in AAM

2) Local Feature Extracting with Gabor Filters: A Gabor filter is the modulation of a sinusoidal wave with a Gaussian function as shown Fig. 5. Therefore this filter will respond to the frequency which is in a localized part of the signal. 2 dimensional Gabor filters can be viewed as:

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right)\exp\left(i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right) \qquad (1)$$

Where$x' = x\cos(\theta) + y\sin(\theta)$,$y' = -x\sin(\theta) + y\cos(\theta)$. In (1), λ is the wavelength, θ is the orientation, ψ is the phase offset, σ is the standard deviation of the Gaussian kernel and γ is the spatial ratio of the Gabor function.

2D convolution operation is used to obtain the response of a Gabor filter to an image as follows:

$$G(x, y) = \iint I(p, q) g(x - p, y - q; \lambda, \theta, \psi, \sigma, \gamma) dpdq \qquad (2)$$

Where $I(p, q)$ is the image and $G(x, y)$ is the response of a Gabor filter to the image. In the study the fine and deep wrinkles at different orientations are extracted using a Gabor filter set with 4 scales and 6 orientations as shown in Fig. 6. The responses of these filters for an image are also given in the figure.

3) Local Feature Extracting with LBPs: LBPs are powerful descriptors of image texture [29]. LBP operator thresholds the center pixel with its neighbors and assigns a binary code for it. Then the occurrence histogram of these LBP codes is used as a texture feature. Every pixel of the image is labeled with the following equation,

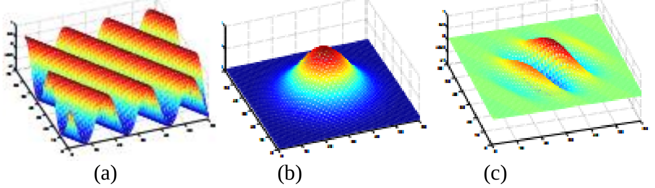


Fig. 5. Gabor filter composition (a) 2 dimensional sinusoid oriented at 30° (b) Gaussian kernel (c) the corresponding Gabor filter

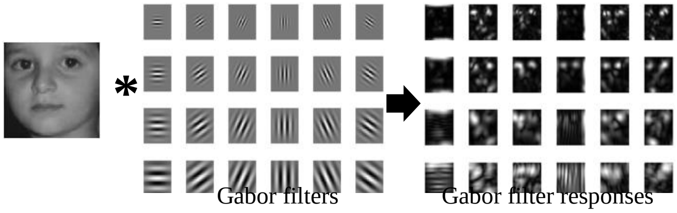


Fig. 6. Gabor filter process

$$LBP_{P,R}(x_c)=\sum_{p=0}^{P-1}u\left(x_p-x_c\right)2^p \quad u(y)=\begin{cases}1, & if\ y\geq 0\\ 0, & if\ y<0\end{cases} \quad (3)$$

Where x_c is center pixel, x_p represents one of his P neighbors and R is the radius. In this equation 2^P different LBP codes can be generated for the center pixel but all of them are not used. Generally the uniform patterns are used in texture description. Uniform patterns are the ones that contain at most two bitwise transitions from 0 to 1 or vice versa when the binary pattern is considered circular. These patterns account for a bit less than 90% of all patterns when using (8,1) neighborhood [29]. But holistic descriptions of facial images are not reasonable as the texture descriptors tend to average over the image area [30]. However it is important to retain the information of spatial relations for facial images. Furthermore local representations are more robust to illumination or pose variations than holistic representations. As a result, spatial LBP histograms are extracted for an efficient representation of facial images. For this purpose image is divided into m regions from which the spatial histograms are produced as follows,

$$H_{i,j}=\sum_{x_c\in R_j} f\left\{LBP_{8,1}(x_c)=U(i)\right\},\quad f(y)=\begin{cases}1, & y \text{ is true}\\ 0, & y \text{ is false}\end{cases}$$
$$i=0,1,\ldots,n-1\ \ j=0,1,\ldots,m-1 \qquad (4)$$

Where $H_{i,j}$ is the i^{th} value of the LBP histogram of j^{th} region and $U(i)$ is the vector keeps uniform patterns. To build a global description of the image, regional histograms are concatenated in a single vector. In this work the detailed skin textures of facial images are extracted using spatial LBP histograms as shown in Fig. 7. Spatial representation of a facial image is obtained by dividing the image into 8x8 regions, producing the LBP histograms of these regions and concatenating them into a single vector.

C. Dimensionality Reduction

After the feature extraction module, PCA is performed in order to find a lower dimensional subspace which carries significant information for age estimation. The PCA method finds the embedding that maximizes the projected variance given below.

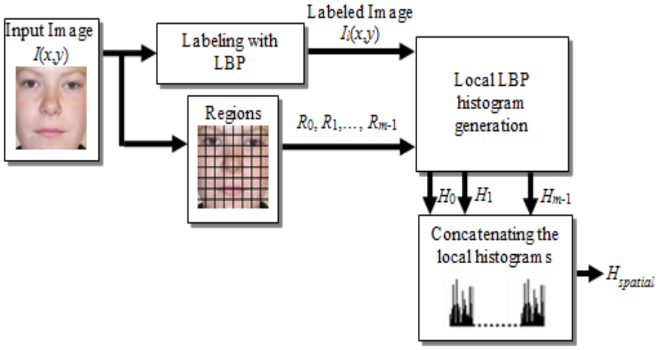


Fig. 7. Spatial LBP histogram generation

$$W_{opt}=\arg\max_{\|W\|=1} W^T SW \qquad (5)$$

In (5) $S=\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T$ is the scatter matrix, and $\bar{x}$ is the mean of feature vectors, $\{x_i: x_i\in R^D\}_{i=1}^{n}$. The solution of this problem is given by the set of $d\leq D$ eigenvectors associated with the largest eigenvalues of the scatter matrix. After determining the projection subspace, all the samples are projected on it using $y_i=W^T f_i$ allowing thus dimensionality reduction.

D. Classification

In the age group classification module, the subject is classified into one of the age groups using a SVM classifier. SVM is a supervised learning method which uses support vectors to build a classification or regression model [30]. SVM finds a linear and optimal hyperplane which can separate two classes, providing the lowest separation error and maximum margin between the classes. Consider a two class classification problem with M training points x_i and assigned labels y_i is defined as,

$$\left\{(x_i,y_i)\middle| x_i\in R^n, y_i\in\{-1,+1\}\right\}_{i=1}^{M} \qquad (6)$$

Linear SVM assumes that there exists a hyper plane separating two classes. The function of this hyper plane can be formulated as,

$$f(x)=w^T x+b \quad x\in R^n, w\in R^n, b\in R \qquad (7)$$

Where w is the normal vector and b is the distance from the origin. This function can be used as a decision rule for a data point x with label y is as follows:

$$y=\begin{cases}+1, & f(x)\geq 0\\ -1, & f(x)<0\end{cases} \qquad (8)$$

In the training phase the w and b are found such that this decision rule is valid for all training and test data points. In real world applications, data can rarely be separated by a linear hyper plane. Thus the basic version of SVM only allowing linear classification is changed by applying a so called kernel trick. The non-linear separable data is transformed into higher dimensional space using a kernel function $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$, which behaves like a scalar product and keeps the computational costs low. The conventional SVM assumes that there exists a linear hyperplane separating two classes. To extend this to nonlinear data, radial basis function kernel as shown in (9) is used in this paper.

$$K(x_i, x_j) = e^{\|x_i - x_j\|^2/2\sigma^2} \qquad (9)$$

SVM is originally designed for binary classification. To solve multi-class classification problems with SVM, different implementations like one-against-all and one-against-one are used. In the proposed approach one against-one method is used for multi-class classification. For this purpose, $k(k$-1) binary SVMs representing all possible pairs of k classes are constructed. Each of these classifiers is trained to discriminate only two of the k classes. Then majority voting strategy is used to predict the final output. The data point is assigned to the class that has maximum votes. The optimal parameters for SVM were selected experimentally from the training set.

E. Regression

After finding the age groups of facial images, the age estimation problem is recast as a multiple linear regression as follows:

$$L = YB + e, \;\; Var(e) = \sigma^2 I \qquad (10)$$

Where Y is the data matrix including a columns of 1s, B is the unknown parameter vector, L is the age label vector and e is the error vector with zero mean and common variance σ^2. During the learning stage the unknown parameters are estimated using least squares, or robust regression. The regression function used in this study is a quadratic function given by,

$$\hat{l} = \hat{\beta}_0 + \hat{\beta}_1{}^T y + \hat{\beta}_2{}^T y^2 \qquad (11)$$

Where $\hat{l}$ is the estimate of age, $\hat{\beta}_0$ is the offset, $\hat{\beta}_1$, $\hat{\beta}_2$ are the weight vectors and y is the low dimensional representation of the extracted feature vector.

F. Decision Level Fusion

In the proposed method decision level fusion is performed in both classification and regression modules. In the classification module, the age class labels are produced with three distinct classifiers which are trained with global and local features. Then the results of these classifiers are combined to determine the age group of the subject.

In the age estimation module three aging functions are modeled separately in that age group, using global and local features. The results of these aging functions are also combined to make a final decision for the age of the test sample as follows:

$$age = \sum_{i=1}^{N} age_i \Big/ N \qquad (12)$$

Where *age* is the final age of the test sample, *age*$_i$ is the age estimated by the i^{th} aging function and N is the total number of aging functions.

IV. Experiments and Results

In this paper, the performance of the proposed method is evaluated using FG-NET [31] and PAL [32] aging databases. FG-NET database comprises of 1,002 images in the age range of 0-69 years.

The images were retrieved from real-life albums of 82 subjects, so the dataset includes uncontrolled variations of occlusion, facial expressions, head pose, illumination, etc. The data distribution of the FG-NET database according to age is shown in Fig. 8-(a). It can be seen from the figure that the image distribution is not uniform which can adversely affect the system performance.

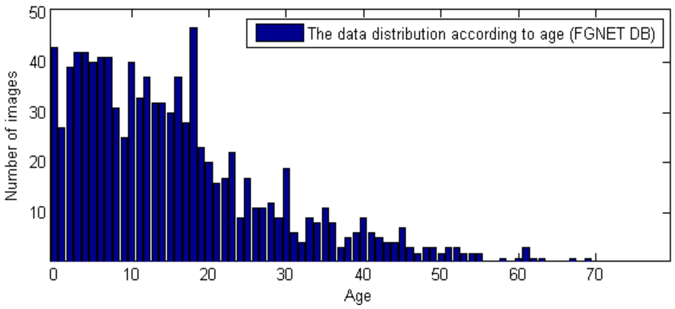
The PAL aging database contains 580 images of different individual in the age range of 18-93 years. The images were captured under natural lighting conditions using a digital camera. This database includes various expressions such as neutral faces, anger, sadness or smiling. The distribution of images in this database according to age is shown in Fig. 8-(b).

Performance evaluation is done using leave-one-person-out (LOPO) for FG-NET database. In this method the images of a person are used as test set and all the other images are used as training set in each fold. This procedure is iterated for 82 folds, which is the number of subjects in the database. After 82 folds final estimation is calculated by taking the mean of all estimations. In the experiments 3-fold cross validation mode is used for PAL database in which the 1/3 of the images are selected randomly as test set and the rest are used as training set. After 3 folds the mean of all estimations is determined as estimation performance of the system.

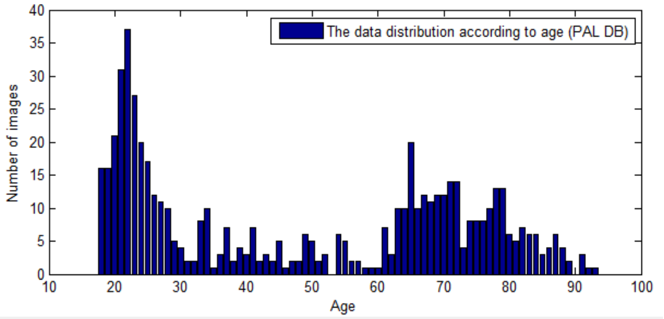
The Mean Absolute Error (MAE) and Cumulative Score (CS) metrics are used for performance comparison in the study. MAE is defined as the average of the absolute error between the recognized labels and the ground truth labels as follows:

$$MAE = \sum_{i=1}^{N} \left| \hat{l}_i - l_i \right| / N_t \qquad (13)$$

Where $\hat{l}_i$ is the estimated age for i^{th} test sample, l_i is the corresponding ground truth, and N_t is the total number of the test



(a)



(b)

Fig. 8. The data distributions of (a) FG-NET and (b) PAL databases according to age

samples. CS enables performance comparison at different absolute error levels. It is the ratio of the number of images, whose absolute errors are lower than a threshold value to the total number of images. It is expressed by,

$$CS(th) = \frac{N_{e\leq th}}{N} \times 100(\%) \qquad (14)$$

Where $N_{e\leq th}$is the number of images with the absolute estimation error is less than *th*, and *N* is the number of test images.

In the global feature extraction step, the coordinates of 68 landmark points on the training samples are used to train the shape model. Also the mean shape is determined from these points. Next, affine transformation is used in the warping process of all images to the mean shape. Then approximately 7000 gray-level intensities in the facial region of the corresponding shape-normalized images are used to train the appearance model. Finally, 277 AAM model parameters are used as global features to represent the images.

In the local feature extraction, wrinkle information of the facial images is extracted using Gabor filters. The fine and deep wrinkles at different orientations are extracted with Gabor filters applied in 4 scales and 6 orientations. The responses of these filters are concatenated into a single vector and dimensionality reduction is performed using. Furthermore the detailed skin textures of facial images are extracted using spatial LBP histograms. For this purpose LBP histograms are produced from 8x8 sub-regions of facial images and concatenated into a single vector, resulted a spatial representation of the facial image. Also PCA is applied to learn a low dimensional representation of this feature vector.

In the spatial LBP histogram generation phase, the number of sub-regions is determined experimentally. For this purpose, the age estimation performances of the spatial LBP histograms produced with different number of sub-regions are calculated and the results are shown in Fig. 9. It can be seen from the figure that using the 8x8 sub-regions gives better results for age estimation.

After the feature extraction and dimensionality reduction phase, age group classification is performed using three SVM classifiers. The age ranges of age groups are selected as: 0-12 childhood, 13-19 adolescence, 20-39 young adulthood, 40-64 middle adulthood and ≥65 late adulthood. Then age estimation is performed in the specified age group using multiple linear regression. For this purpose, three aging functions are modeled separately using global features, wrinkle features and skin features for age estimation. Then the results of these aging functions are combined and a final decision is made for the test sample. In the experiments, first the age estimation performances of the global, local and fused features are determined using a single level age estimation scheme. In this scheme all the images are used to train the aging functions and the decision level fusion is performed for a final decision for the age. The experimental results on FG-NET and PAL databases are listed in Table 1. It can be seen from the table that age estimation performance of the AAM features is better than Gabor and LBP features on FG-NET and PAL databases as the AAM features both include the shape and appearance information of facial images. But when they are fused with local features including wrinkle and skin texture information at decision level, age estimation performance noticeably increased. As a result MAE of 4.87 years on FG-NET database and MAE of 5.38 years on PAL database is achieved when using single level age estimation based on decision level fusion of facial features.

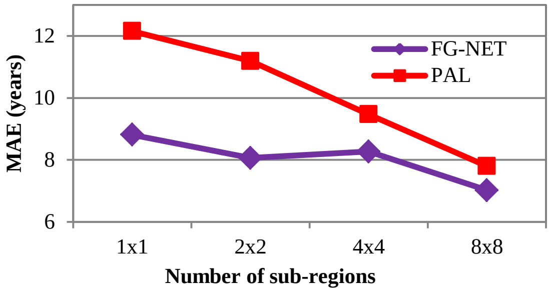


Fig. 9. MAE's of different number of sub-regions used in spatial LBP histogram generation

TABLE I. SINGLE LEVEL AGE ESTIMATION RESULTS (MAE) BASED ON DECISION LEVEL FUSION OF FEATURE VECTORS

Feature	**FG-NET**	**PAL**
AAM	6.02	6.61
LBP	7.02	7.80
Gabor	6.55	7.37
AAM+LBP	5.36	5.96
AAM+Gabor	5.23	5.70
AAM+LBP+Gabor	**4.87**	**5.38**

The Cumulative Scores of the single level age estimation scheme on FG-NET and PAL databases at error levels from 0 to 15 years are shown in Fig. 10. Age of approximately 8.08% of the subjects in the FG-NET database and 5.12% of the subjects in PAL database can be estimated with zero error level. As the error level increases the estimation accuracy also increases for all feature extraction methods. This single level age estimation approach is able to achieve cumulative scores of 89.92% and 85.86% for an absolute error of 10 years for FG-NET and PAL databases, respectively.

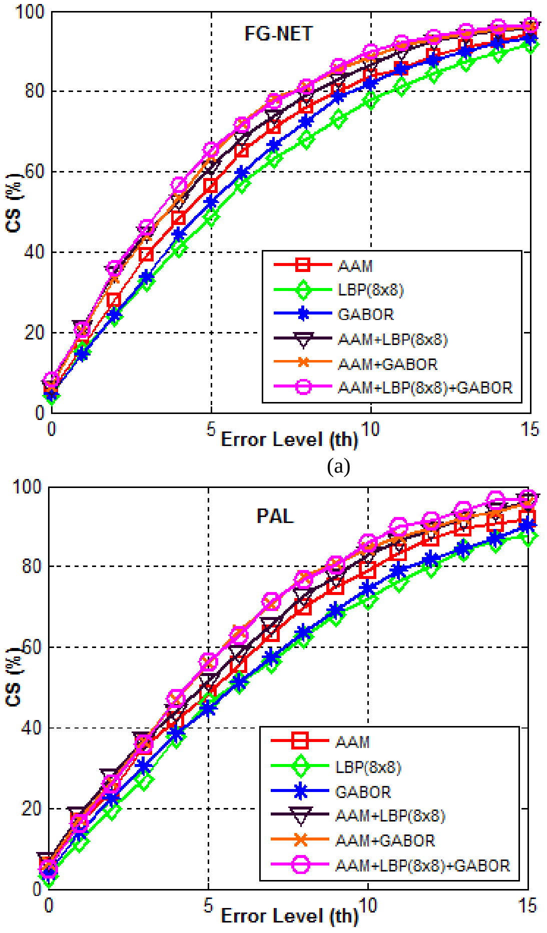
Age estimation performance of the proposed hierarchical age estimation approach based on decision level fusion of global and local features both in classification and detailed age estimation phases are given in Table 2. One can see from the table that proposed method achieves the MAE of 4.13 on FG-

NET database and MAE of 4.67 on PAL database. Cumulative scores for FG-NET and PAL databases for an absolute error of 10 years are increased to 90.88% and 92.17%, respectively as shown in Fig. 11.

As the FG-NET database is the most common database used in age estimation works, the performance of the proposed method and the previous works on FG-NET aging database are compared in Table 3. One can see from Table 3 that proposed method has an MAE of 4.13 years which is lower than the previous methods. This result also shows that decision level fusion of global and local features in a hierarchical system improves the age estimation performance. Gabor filters and block-based LBP histograms encode the texture information of the facial images. Global features are extracted with AAMs which encodes the shape and appearance information of facial images. These feature sets capture differential complementary information. The decision level fusion of these features estimates the age better when compared with the age estimation accuracies obtained using these features alone.

TABLE II. HIERARCHICAL AGE ESTIMATION RESULTS BASED ON DECISION LEVEL FUSION OF FEATURE VECTORS

Database	FG-NET	PAL
MAE	**4.13**	**4.67**



(a)

(b)

Fig. 10. Cumulative scores of global, local and fused features using single level age estimation scheme on (a) FG-NET and (b) PAL databases

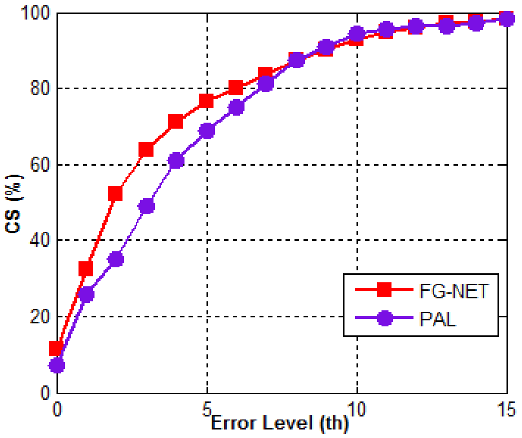


Fig. 11. Cumulative scores of the proposed method on FG-NET and PAL databases

TABLE III. THE COMPARISON OF ESTIMATION RESULTS ON FG-NET DATABASE

Methods	MAEs
AAS [14]	14.83
WAS [15]	8.06
AGES [15]	6.77
AGES_{lda} [15]	6.22
LARR[20]	5.16
RMIR[4]	8.37
Ju and Wang[24]	6.85
Lu and Tan [21]	5.75
Choi et al. [3]	4.32
Proposed	**4.13**

V. CONCLUSION

In this paper, a hierarchical age estimation method relying on decision level fusion of AAM, Gabor and LBP features of facial images is proposed. Its main contribution is decision level fusion of global texture features and local texture features of facial images. Locality is preserved by regional LBP histograms and Gabor filters. Furthermore, these local features are combined with global features of images extracted with AAMs. Experimental results using the FG-NET and PAL aging databases have shown that the proposed method is better than previous methods.

REFERENCES

[1] M. Albert, K. Ricanek and E. Patterson, A review of the literature on the aging adult skull and face: implications for forensic science research and applications, Forensic Science International 172 (1) (2007) 1-9.

[2] M. Gonzalez-Ulloa and E. S. Flores, Senility of the face-Basic study to understand its causes and effects. Plastics & Reconstructive Surgery 36 (2) (1965) 239-246.

[3] S. E. Choi, Y. J. Lee, S. J. Lee and K. R. Park, “Age estimation using a hierarchical classifier based on global and local facial features”, Pattern Recognition, vol. 44, no. 6, pp. 1262-1281, June 2011.

[4] B. Ni, Z. Song and S. Yan, “Web image and video mining towards universal and robust age estimator”, IEEE Transactions on Multimedia, vol. 13, no. 6, pp. 1217-1229, December 2011.

[5] Y. H. Kwon and N. V. Lobo, “Age classification from facial images”, Computer Vision and Image Understanding, vol. 74, no. 1, pp. 1-21, April 1999.

[6] T. R. Alley, Social and Applied Aspects of Perceiving Faces, Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

[7] W.-B. Horng, C.-P. Lee and C.-W. Chen, "Classification of Age Groups Based on Facial Features", Tamkang Journal of Science and Engineering vol. 4, no.3, pp. 183-192, 2001.

[8] M. M. Dehshibi and A. Bastanfard, "A new algorithm for age recognition from facial images", Signal Processing, vol. 90, no.8, pp. 2431-2444, 2010.

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

[10] T. Cootes, G. Edwards and C. Taylor, "Active appearance models", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 681-685, Jun 2001.

[11] A. Lanitis, C. Taylor and T. Cootes, "Toward automatic simulation of aging effects on face images", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 4, pp. 442-455, April 2002.

[12] S. Kohli, S. Prakash and P. Gupta, "Hierarchical age estimation with dissimilarity-based classification", Neurocomputing, vol. 120 pp. 164-176, 2013.

[13] W. -L. Chao, J. -Z. Liu and J. -J. Ding, "Facial age estimation based on label-sebsitive learning and age oriented regression", Pattern Recognition, vol. 43, pp. 628-641, 2013.

[14] X. Geng, Z. H. Zhou, Y. Zhang, G. Li and H. Dai, "Learning from facial aging patterns for automatic age estimation", Proc. of ACM Conference on Multimedia, pp. 307-316, 2006.

[15] X. Geng, Z. H. Zhou and K. S. Miles, "Automatic age estimation based on facial aging patterns", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp. 2234-2240, December 2007.

[16] Y. Fu, Y. Xu and T. S. Huang, "Estimating human age by manifold analysis of face pictures and regression on aging feature"s, Proc. IEEE International Conference on Multimedia and Expo, pp. 1383-1386, 2-5 July 2007.

[17] D. Cai, X. He, J. Han and H.-J. Zhang, "Orthogonal laplacianfaces for face recognition", IEEE Transactions on Image Processing, vol. 15, no. 11, pp. 3608-3614, 2006.

[18] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold", IEEE Transactions on Multimedia, vol. 10, no. 4, pp. 578-584, June 2008.

[19] G. Guo, Y. Fu, T. S. Huang and C. R. Dyer, "Locally adjusted robust regression for human age estimation", IEEE Workshop on Applications of Computer Vision (WACV'08), pp. 1-6, 7-9 Jan 2008.

[20] G. Guo, Y. Fu, C. R. Dyer and T. S. Huang, "Image-based human age estimation by manifold learning and locally adjusted robust regression", IEEE Transactions on Image Processing, vol. 17, no. 7, pp. 1178-1188, July 2008.

[21] J. Lu and Y. -P. Tan, "Ordinary Preserving Manifold Anaysis for Human Age and Head Pose Estimation", IEEE Transactions on Human-Machine Systems, vol.43, no.2, pp. 249-258, 2013.

[22] H. Fukai, H. Takimoto, Y. Mitsukura and M. Fukumi, "Apparent age estimation system based on age perception", Proc. SICE 2007 Annual Conference, pp. 2808-2812, , 17-20 Sept 2007.

[23] T. Ahonen, A. Hadid and M. Pietikainen, "Face description with local binary patterns: Application to face recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no.12, pp. 2037-2041, December 2006.

[24] C. H. Ju and Y. H. Wang, "Automatic age estimation based on local feature of face image and regression", 2009 International Conference on Machine Learning and Cybernetics, pp. 885-888, 12-15 July 2009.

[25] F. Gao and H. Ai, "Face age classification on consumer images with gabor feature and fuzzy LDA method", Proc. of 3rd International Conference on Advances in Biometrics (LNCS'5558), pp. 132-141, 2-5 June 2009.

[26] G. Guo, G. Mu, Y. Fu and T. S. Huang, "Human Age Estimation Using Bio-Inspired Features", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 112-119, 2009.

[27] H. Han, C. Otto, X. Liu and A. Jain, K., "Demographic Estimation from Face Images: Human vs. Machine Performance", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014.

[28] C. Li, Q. Liu, W. Dong, X. Zhu, J. Liu and H. Lu, "Human Age Estimation Based on Locality and Ordinal Information", IEEE Transactions on Cybernetics, 2014.

[29] T. Ojala, M. Pietikainen and T. Maenpaa, "Multi-resolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971-987, 2002.

[30] C. Cortes and V. Vapnik, "Support-vector network", Machine Learning, vol. 20, no.3, pp. 273-297, 1995.

[31] FG-Net aging database. Available: http://sting.cycollege.ac.cy /~alanitis/fgnetaging.

[32] M. Minear and D. C. Park, "A lifespan database of adult stimuli", Behavior Research Methods, Instruments and Computers, vol.36, no.4, pp.630-633, 2004.

Flow-Based Specification of Time Design Requirements

Sabah Al-Fedaghi
Computer Engineering Department
Kuwait University
Kuwait

***Abstract*—This paper focuses on design requirements in real-time systems where information is processed to produce a response within a specified time. Nowadays, computer control applications embedded in chips have grown in significance in many aspects of human life. These systems need a high level of reliability to gain the trust of users. Ensuring correctness in the early stages of the design process is especially a major challenge in these systems. Faulty requirements lead to errors in the final product that have to be fixed later, often at a high cost. A crucial step in this process is modeling the intended system. This paper explores the potential of flow-based modeling in expressing design requirements in real-time systems that include time constraints and synchronization. The main emphasized problem is how to represent time. The objective is to assist real-time system requirement engineers, in an early state of the development, to express the timing behavior of the developed system. Several known examples are modeled and the results point to the viability of the flow-based representation in comparison with such time specifications as state-based and line-based modeling.**

***Keywords*—design requirements; conceptual model; time constraints; model-based systems engineering; requirements specification**

I. Introduction

The product development life cycle in the engineering domain aims at achieving, among other goals, a design process with complete and precise specifications that satisfy all requirements. Requirements are descriptions of functions, features, and goals of the product. The requirements describe ‘what’ the intended product should do, but the ‘how’ is specified as design requirements during the design phase, where measurability and verifiability are of utmost importance. Design requirements (the focus of this paper) include the specifications that the intended product must meet in order to pass the acceptance test. Specifications consist of information that controls the creation of the intended product.

Early assurance of the correctness of design requirements is a major challenge in any system. Faulty design requirements lead to errors in the final product that have to be fixed later, often at a high cost. Reoccurring causes of failures include:

- Inadequate definitions and modifications of specifications
- Faulty interpretation and understanding
- Not meeting customer requirements
- Design not meeting manufacturing requirements
- Difficulties in specifying technical requirements
- Difficulties in interpreting and understanding specifications [1]

There are various methods for specifying real-time systems. For example, prototyping tools can be used by the designer and user to view the product in the development stage [2]. However, prototyping is a phase that comes after the specifications. If prototyping has produced unsatisfiable results, then the designer may have to re-specify the requirements. There are also formal specifications of real-time systems that should enable the system designer to verify mathematically that a system meets timing constraints. However, formal methods are still limited as a verification tool, especially for software systems, not to mention the complexity introduced by timing. Various specification languages for real-time systems with timing constraints can be expressed within the specifications (e.g., [3]), “but at the cost of restricting other features” [4].

The specifications of design requirements are usually formulated in a mixed of English, tables, graphs, screen shots, and unified modeling language (UML) diagrams. According to Palshikar [5], design requirements are examined in terms of:

- accurate reflection of the users’ requirements
- clarity, unambiguity, and understandability
- flexibility and feasibility for the engineers
- easily defined acceptance test cases
- an abstract and high-level manner of writing, away from design, implementation, and technology platforms

“Despite some help from modeling tools such as UML, the problem of ensuring the quality of requirements remains. The process is heavily manual and time-consuming, involving reviews and sometimes-partial prototyping. Using multiple notations (such as those in UML) introduces additional problems” [5].

Additionally, this paper is concerned with *design requirements* in *real-time systems* where information is processed to produce a response within a specified time. A real-time system interacts with the environment within certain timing constraints and the requirement specifications for such a system must include representation of timing which can guarantee meeting these constraints. The notion of *time* is an important element in such systems, especially if critical features (e.g., safety) are functionally required. The problems here are how to represent time, how to capture causality behavior, and how to integrate functional and timing activities [6].

Embedded systems where the software is completely encapsulated by the hardware that it controls are often real-time systems. An embedded system is a system that interacts continuously with its physical sphere via sensors and actuators. Nowadays, computer control applications embedded in chips have grown in significance in many aspects of human life (e.g., medicine, mobile phones, and vending machines). These embedded systems need a high level of reliability to gain the trust of users. Ensuring correctness in the early stage of the design process is especially a major challenge in these systems.

A crucial step in this process is modeling the intended system. Model-based design has been introduced as the method to deal with the design process where the requirements are specified in a systematic way before continuing with the design and implementation phases. A great deal of attention has focused on this, such as interest in the unified modeling language with its graphical notation, which is used for documentation, communication, and requirement capture, as well as being an abstraction base for implementation details. This paper explores the potential of the flow-based modeling [7–12] in expressing design requirements in real-time systems that include time constraints and synchronization.

This paper focuses on the *representation* of timing constraints. The objective is to assist real-time system requirement engineers, at an early state of the development, to express the timing behavior of the developed system. *Representation* here refers to humans' and machines' representation of knowledge for the purpose of communication and understanding and analyzing the embedded semantics (e.g., diagrams, formal notations). Representation is usually associated with reasoning (e.g., the computational understanding of human-level cognitive abilities). This concentrates on the representation aspect that can be used for manual or computation analysis, as in problem solving in artificial intelligence.

In preparation to recast the representation of several known design problems in terms of flow-based modeling, and to make this paper self-contained, the next section briefly reviews published materials describing the flow-based model. Several features of the model will be further illustrated.

II. Flowthing Model

The flowthing model (FM) is a uniform method for representing "things that flow," called *flowthings.* Flow in FM refers to the exclusive (i.e., being in one and only one) transformation among six *states* (also called stages): transfer (input/output), process, creation, release, arrival, and acceptance, as shown in Fig. 1. We will use *receive* as a combined stage of *arrive* and *accept* whenever arriving flowthings are always accepted.

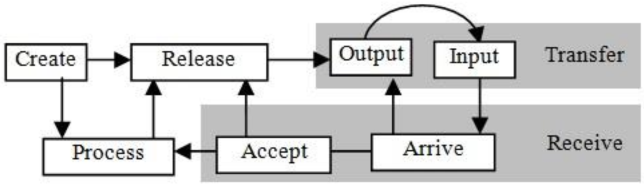


Fig. 1. Flowsystem

A flowthing has the capability of being created, released, transferred, arrived, accepted, and processed while flowing within and between "units" called *spheres*. A flow system (referred to as *flowsystem*) is a system with six stages and transformations (edges) between them. In FM, flows can be controlled by the progress (sequence) of the stream of events (creation, release, transfer, transfer within the next sphere, release, reception, …) or by a triggering (denoted by a dashed arrow) that can initiate a new flow. Spheres and subspheres are the environments of the flowthing, such as a company, a computer, and a person. A sphere can include the sphere of a flowsystem that includes the transfer stage. Triggering is the transformation from one flow to another, e.g., a flow of electricity triggers a flow of air.

Example: In studying a "successful" model checking for verifying requirements, Palshikar [5] used a simple pumping control system that transfers water from a source tank A into sink tank B using a pump P as shown in Fig. 2. Each tank has two level-meters to detect whether their levels are empty or full. The tank level is ok if it is neither empty nor full.

Initially, both tanks are empty. The pump is to be switched on as soon as the water level in tank A reaches ok (from empty), provided that tank B is not full. The pump remains turned on as long as tank A is not empty and as long as tank B is not full. The pump is to be switched off as soon as either tank A becomes empty or tank B becomes full. The system should not attempt to switch the pump off (on) if it is already off (on). [5]

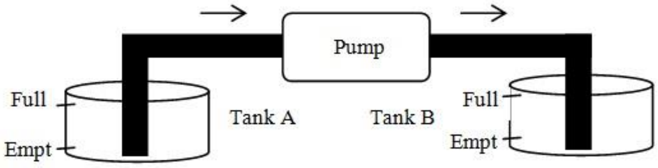
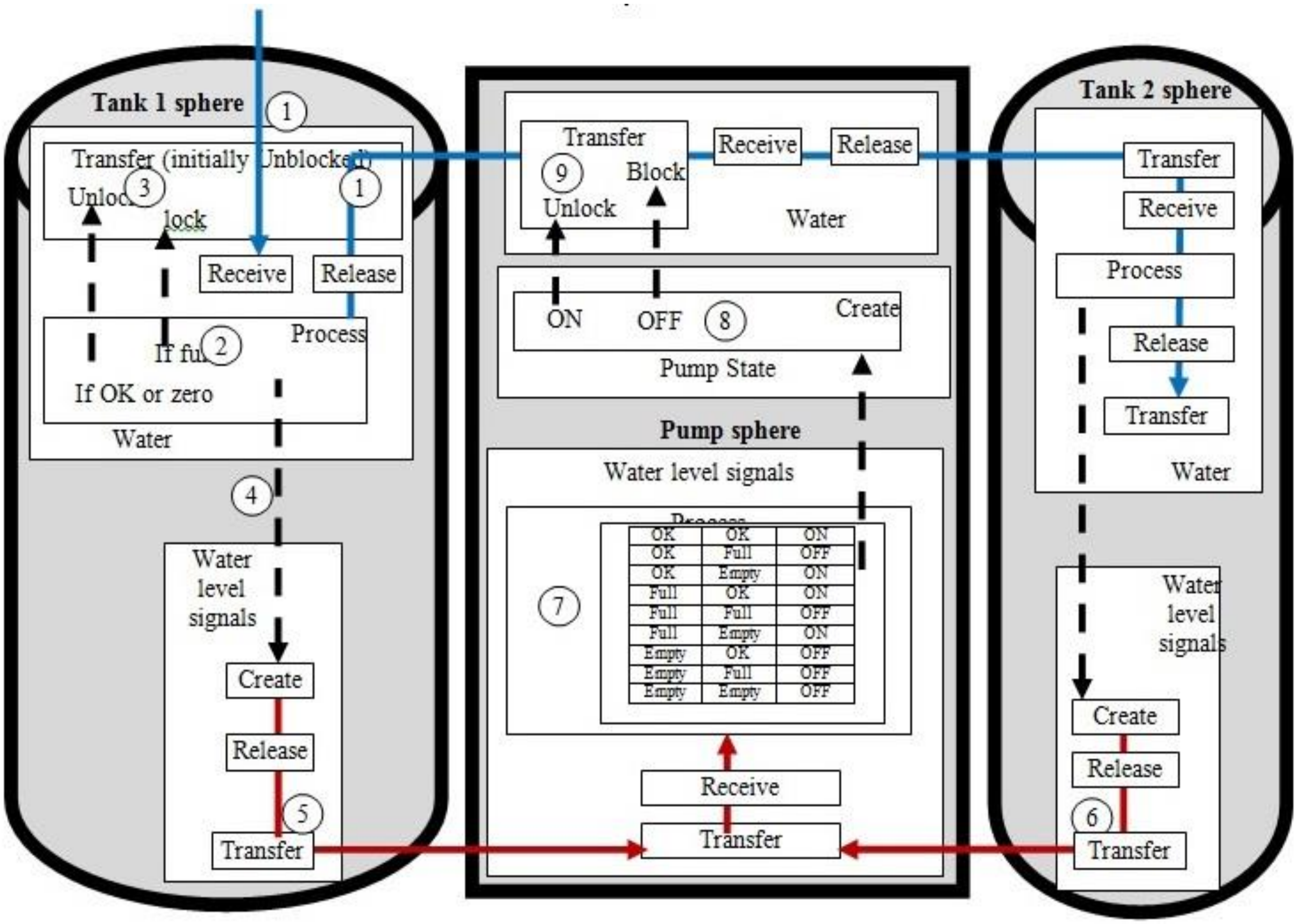


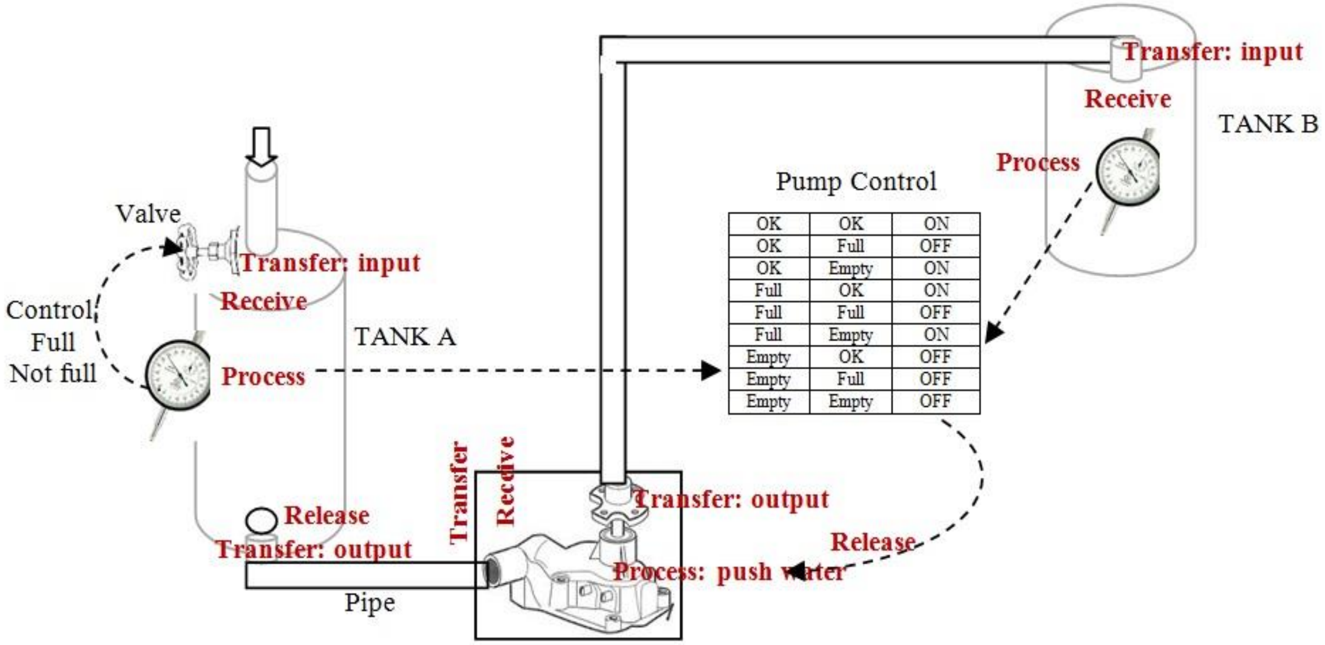
Fig. 2. A simple pumping control system (redrawn from [5])

Finite state machine (FSM) approach is utilized as an abstract notation for defining requirements and design. Fig. 3 shows the FM representation of this pumping control system. It impedes some assumptions which are illustrated in Fig. 4.



OK	OK	ON
OK	Full	OFF
OK	Empty	ON
Full	OK	ON
Full	Full	OFF
Full	Empty	ON
Empty	OK	OFF
Empty	Full	OFF
Empty	Empty	OFF

Fig. 3. FM representation of the pumping control system



OK	OK	ON
OK	Full	OFF
OK	Empty	ON
Full	OK	ON
Full	Full	OFF
Full	Empty	ON
Empty	OK	OFF
Empty	Full	OFF
Empty	Empty	OFF

Fig. 4. Illustration of assumptions in FM representation

In Fig. 4, it is assumed that water flows in tank A with the transfer of this flow controlled within that tank system. The water does not flow to tank B (and therefore tank B is drawn above tank A). Accordingly, the pump is installed between the two tanks to push the water toward tank B.

Tank A is a flowsystem with transfer, receive, process, and release stages. The transfer stage is drawn twice to simplify the drawing. The gate valve controls the *transfer* to tank A. As soon as the valve is opened the water is *received* in the tank. The *process* in tank A involves measuring the amount of water and, accordingly, the valve is opened or closed. At the bottom of tank A there is no control, hence *release* and *transfer* to the pump is immediate. This is analogous to passengers that proceed immediately to a waiting airplane after finishing passport processing. Imagine that this passport checking is on one end of the boarding bridge while the airplane is at the other end of the bridge. In this case, the bridge would correlate to the *release* stage as part of the airport system. Moving from the bridge end to the airplane door would be a flow between two *transfer* stages. Accordingly, in tank A's control system, the flow from the tank to the pipe (see the figure) leading to the pump is a flow between two *transfer* stages. It is possible to include each pipe in Fig. 3 as a flowsystem with transfer, release, and transfer stages. However, this is not shown in the Fig. 3.

The pump is similarly a flowsystem. The process stage involves pushing/not-pushing the water toward tank B. There is no need for valves because the water cannot flow to tank B without pushing.

There seems to be incompleteness in Palshikar [5]'s original description of this system in a case where both tanks are full. In this case, the valve to tank A is closed and the pump is off forever. Accordingly, an outlet has been added in the flowsystem in tank B.

Switching the description to Fig. 3, the water flows in (circle 1 in the figure) to be processed (circle 2, measuring its water level) and accordingly opens or closes the valve (3). Also, the processing triggers (4) the control flowsystem of tank A to send a signal (5) about the *current* level of water: empty, okay, or full to the pump control system. On the other hand, tank B also sends (6) such a signal. These signals are processed in the pump control flowsystem (7) to turn on/off the pump (8), which results in the stoppage or flow of the water to tank B (8).

The next section applies FM to the method known time representations in order to compare the two methods side by side.

III. Time and FM

Time requirements play a central role in understanding and designing systems. Timing is typically incorporated after tasks and software architectures are defined, when holistic scheduling algorithms and expected worst-case execution times are analyzed [13]. This paper does not involve such a detailed level of description; rather, it is concerned with a very high level of requirements specifications, e.g., the level of UML use-case, sequence, and activity diagrams. Accordingly, this section relates time to its representation in FM.

Philosophically, time can be conceptualized as a fourth-dimensional phenomenon. Such a conceptualization is inspired by Edwin Abbott's *Flatland*:

> Dr. Abbott pictures intelligent beings whose whole experience is confined to a plane, or other space of two dimensions, who have no faculties by which they can become conscious of anything outside that space and no means of moving off the surface on which they live. He then asks the reader, who has consciousness of the third dimension, to imagine a sphere descending upon the plane of Flatland and passing through it. How will the inhabitants regard this phenomenon? […]
>
> Their experience will be that of a circular obstacle gradually expanding or growing, and then contracting, and they will attribute to *growth in time* what the external observer in three dimensions assigns to motion in the third dimension. If there is motion of our three-dimensional space relative to the fourth dimension, all the changes we experience and assign to the *flow of time* will be due simply to this movement, the whole of the future as well as the past always existing in the fourth dimension. (Italics added.) [14]

The sphere (ball) is seen as constantly changing, and the whole change from birth to disappearance is the "lifetime" of the sphere. Applying the 3-dimensional world, this time must then be a 4th dimension.

Strachan [15]'s conceptualization of the same phenomenon is as follows:

> Let's imagine a miniature world which is a cube. Now suppose that one of the faces of the cube—say the bottom face—is a little 2-dimensional world, a Flatland, inhabited by creatures called 'Toodies' (2-D) . . .
>
> Since the Toodies' Flatland is infinitely thin . . . , then an infinite number of Flatlands could be stacked into the cube . . .
>
> But let us now suppose that a Toody is subjected to some force which can lift him up the 3rd (up and down) dimension of the cube. So he is propelled out of his own paper-thin world, the bottom face of the cube, right up through the cube to its top face. As he does so, he will pass through all the 2-dimensional 'paper' Flatlands which lie in between. Since the whole cube exists, then all of these Flatlands exist, even though they won't exist for Toody until he reaches them. So they lie in Toody's future.
>
> But change occurs, and can only occur, in time. So his movement in this 3rd (up/down) space dimension will seem like *the passage of time* to Toody: it is his time dimension. (Italics added.)

A. Time as spheres

Accordingly, from the FM point of view, these "flatlands" are *flowthings that flow through times spheres*: past1, past2, . . . , now, future1, future 2, . . . In this case, time is modeled as spheres. All of these spheres are projections of different times on flatlands. UML representation of this modeling of time is shown in Fig. 5, which includes slices of time with processes happening in them. Fig. 6 shows the corresponding FM representation.

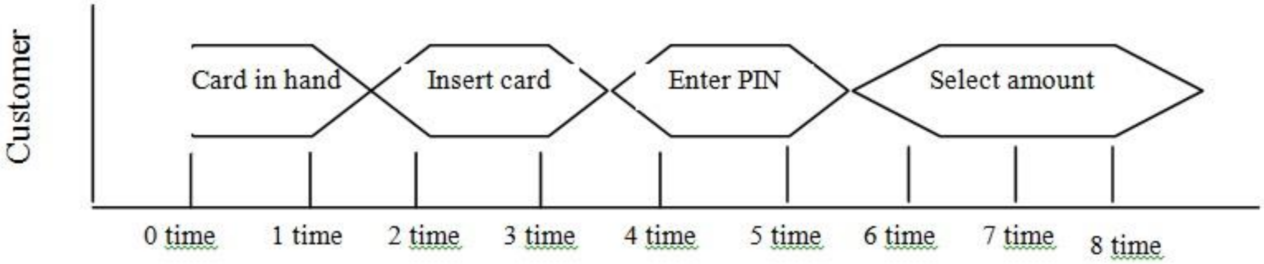


Fig. 5. Sample of UML representation of time (from [16])

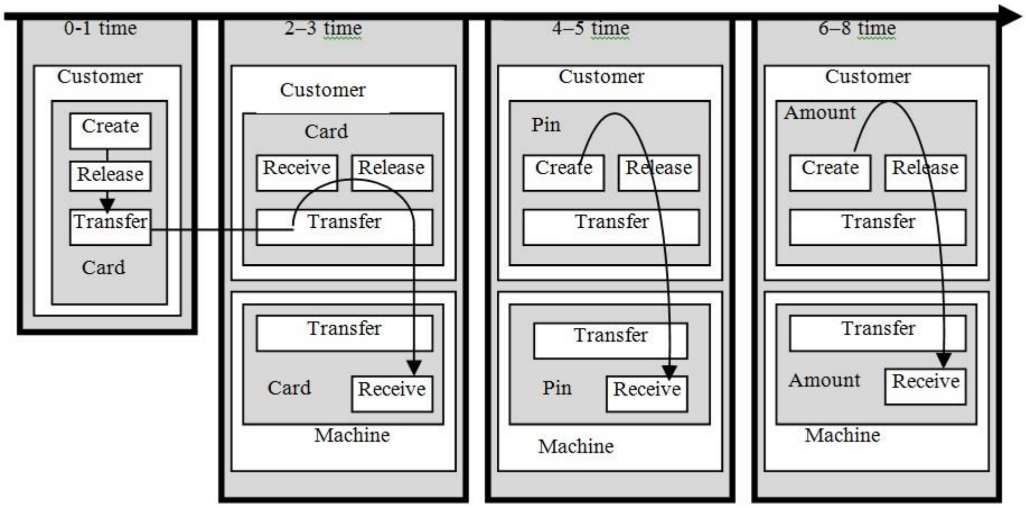


Fig. 6. Time spheres with “flatlands” flow through them

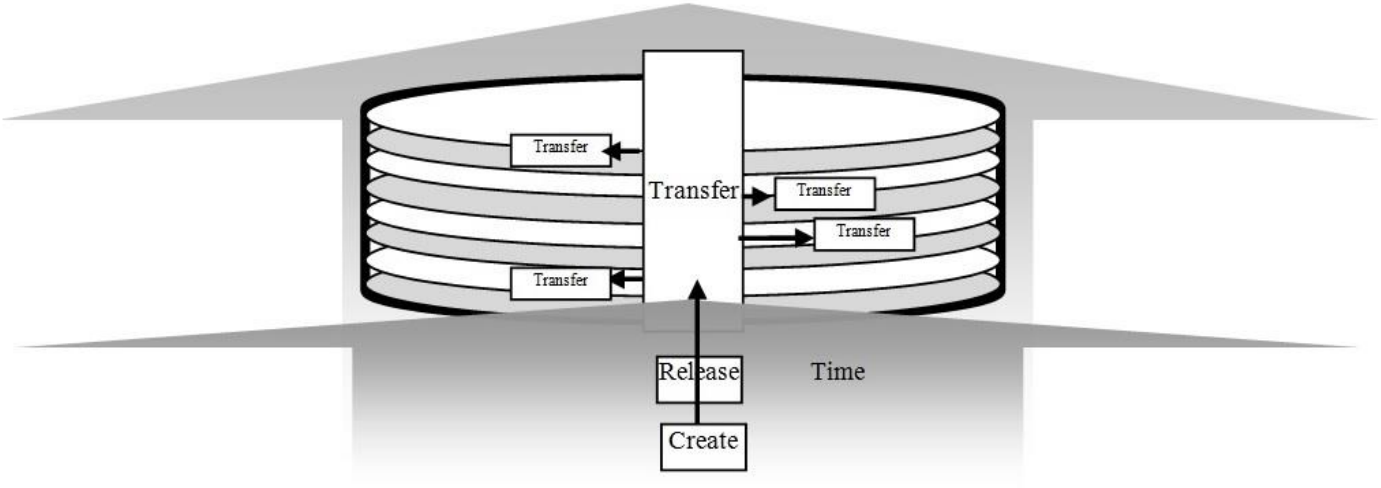


Fig. 7. Cylinder is used instead of a cube to illustrate time flows through “Flatlands”

B. Time as flowthings

Alternatively, time can be conceptualized as a *flowthing* that flows through “flatlands.” In this case, Strachan [15]’ s cube (though we prefer to use a cylinder instead of a cube; see Fig. 7) passes through all the 2-dimensional Flatlands, accomplishing the same result.In this case, time in FM is a flowthing that can be released, transferred, received, and processed. For each flowsystem, it is processed to count its passing though counting, as will be illustrated in the next section. In FM, time is something that flows contiguously from a fourth-dimension sphere to any other sphere, as shown in Fig. 8. If this is of relevance to flows or triggering in that sphere, it is represented by a flowsystem. This conceptualization of time as a flowthing will be utilized in the discussions in the next sections.

IV. Linear Time Diagrams

Timing diagrams "focus on conditions changing within and among lifelines along a linear time axis … on time of events causing changes in the modeled conditions of the lifelines" [17]. They utilize the notions of lifeline, state or condition timeline, destruction event, duration constraint, and time constraint. Timelines are one of the simplest means of representing the flow of events. In UML 2, timing diagrams are a special form of sequence diagrams where the axes are reversed and the lifelines are shown in separate compartments arranged vertically. These diagrams "aren't the most popular" [18].

According to the Web site [17], time duration constraint refers to the duration used to determine whether the constraint is satisfied. It is an association between a duration interval and the constructs that it constrains. For example, that ice should melt into water in 1 to 6 minutes can be represented as shown in Fig. 9. From the conceptual point of view, lining (putting in one category) ice, melting, and water is a categorical mix. Ice and water can be categorized as "states" of H_2O, but melting is certainly not. Also, it seems that H_2O is another name for water. Fig. 10 shows the FM representation.

There are three subspheres: time, ice, and water. The units of time are continuously received (1) and ignored. They are processed (2) as soon as the melting (a kind of process (3)) starts in the ice sphere until "counting" 6 units of time. When the ice starts melting (3), it triggers (4) the counting (processing (2)) of time. When the melting ends (5), the time is ignored again (6) and water is generated (7). The model reflects that time *always* flows through systems, and thus time constraint is awareness of this flow and alignment of events with the flowing time.

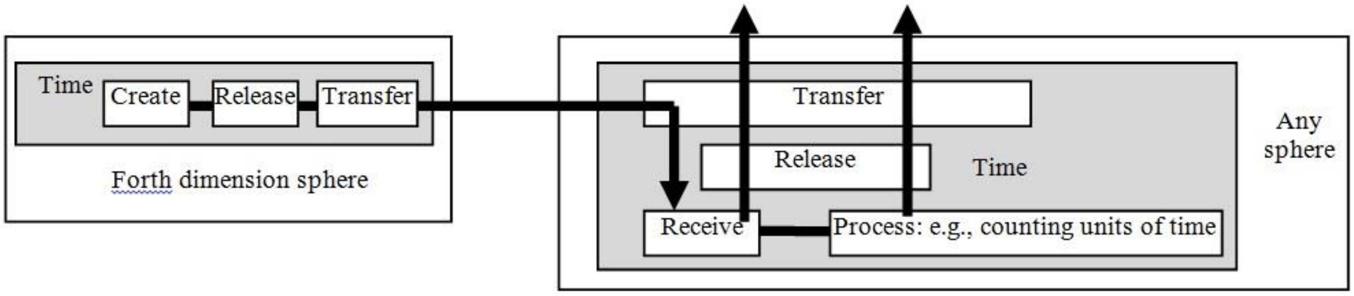


Fig. 8. Time conceptualization FM representation

In addition, a time constraint is time expression used to determine whether the constraint is satisfied. "All traces where the constraints are violated are negative traces, i.e., if they occur, the system is considered as failed" [17]. Fig. 11 is given as a representation of this constraint. It involves two states: sleep and awake. At the change from sleep to awake, the time period {5:40 a.m., 6 a.m.} passes to accomplish this change. The state (sleep or awake) is represented by a horizontal line: no line, no state. The change from a state to another is represented by a vertical line that connects the horizontal lines. The delay that corresponds to the change is represented by the diagonal line and the text {5:40 a.m., 6 a.m.} at the point of beginning the awake state.

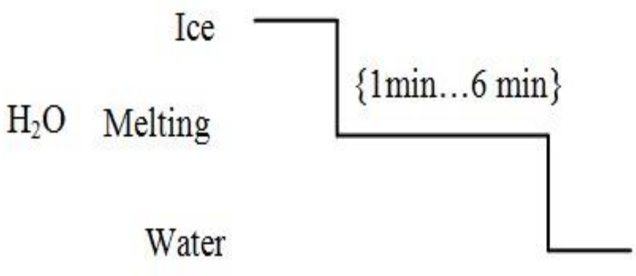


Fig. 9. Representation of how ice should melt into water in 1 to 6 minutes (from [17])

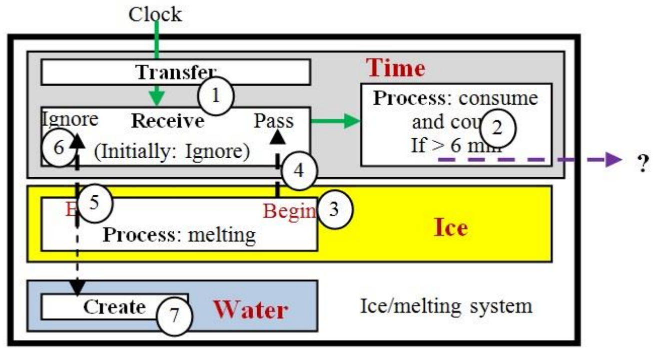


Fig. 10. FM representation

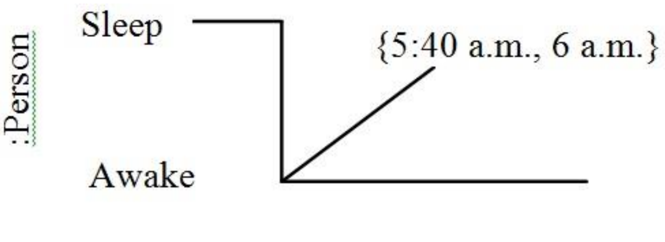


Fig. 11. Person should wake up between 5:40 a.m. and 6 a.m

Semantically, {5:40 a.m., 6 a.m.} is the "length" of sleep. Accordingly, the diagonal line and {5:40 a.m., 6 a.m.} look like a comment and not a modeling of the situation. If it is not a comment, then the representation is misleading because it gives the impression of a three-dimensional representation. Also, there is no indication of "failure" as mentioned in the given constraint. This example shows the limitations of the line representation of time.

Fig. 12 shows the corresponding FM representation. These are the spheres: time, sleep, awake, and the logical join. The clock performs the following:

- At 5:40 a.m., it triggers sleeping
- At 6:00 a.m., it triggers awaking
- At 6:00 a.m. it also triggers checking whether the awaking occurs

Time is generated by the clock and received by the sphere of the time in the system (circle 1). This sphere is the part of the total system that deals with time. The clock sends continuous signals, say 12:00, 12:01, 12:02, . . . , and these data arrive and are received and processed (2). This processing involves the recognition of 5:40 a.m. and 6:00 a.m. If it is 5:40 a.m. then this triggers the person to enter into sleeping (3). He/she is processed (absorbed) into sleeping (4). If it is 6:00 a.m., then this triggers:

- The release (5) of the person from sleeping to awaking (6)
- The checking of whether the person has arrived to the awaking state (7). If this is the case then this triggers success (For simplification sake, success is reported instead of failure; accordingly, the recipient of the report assumes failure if success does not arrive.)

Note that the horizontal joint bar can be represented in FM as shown in Fig. 13.

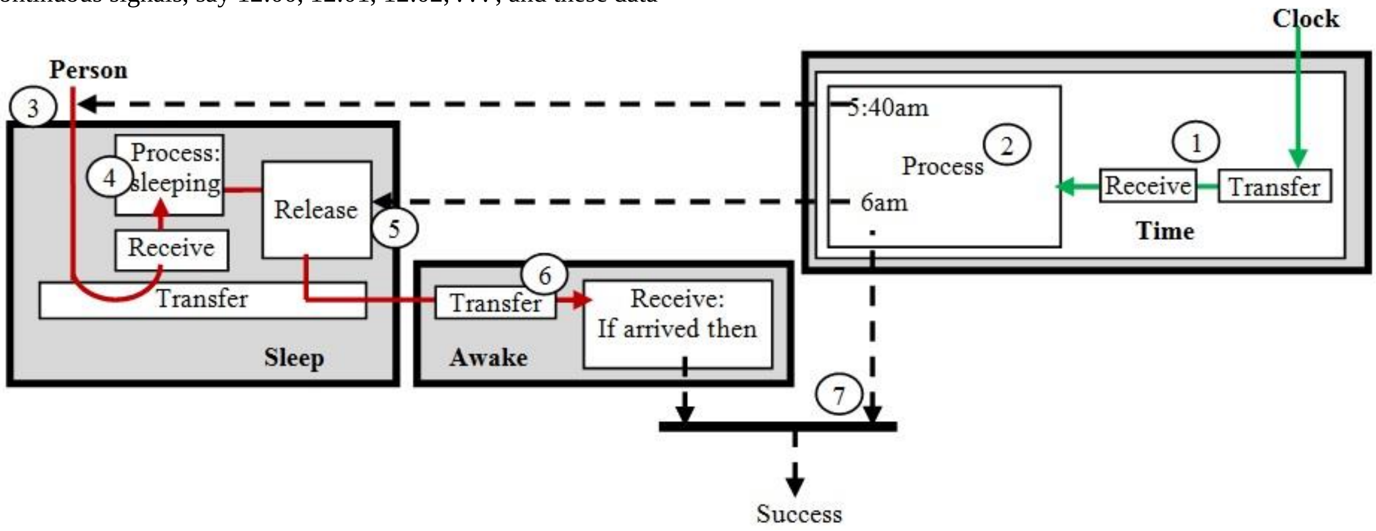


Fig. 12. FM representation

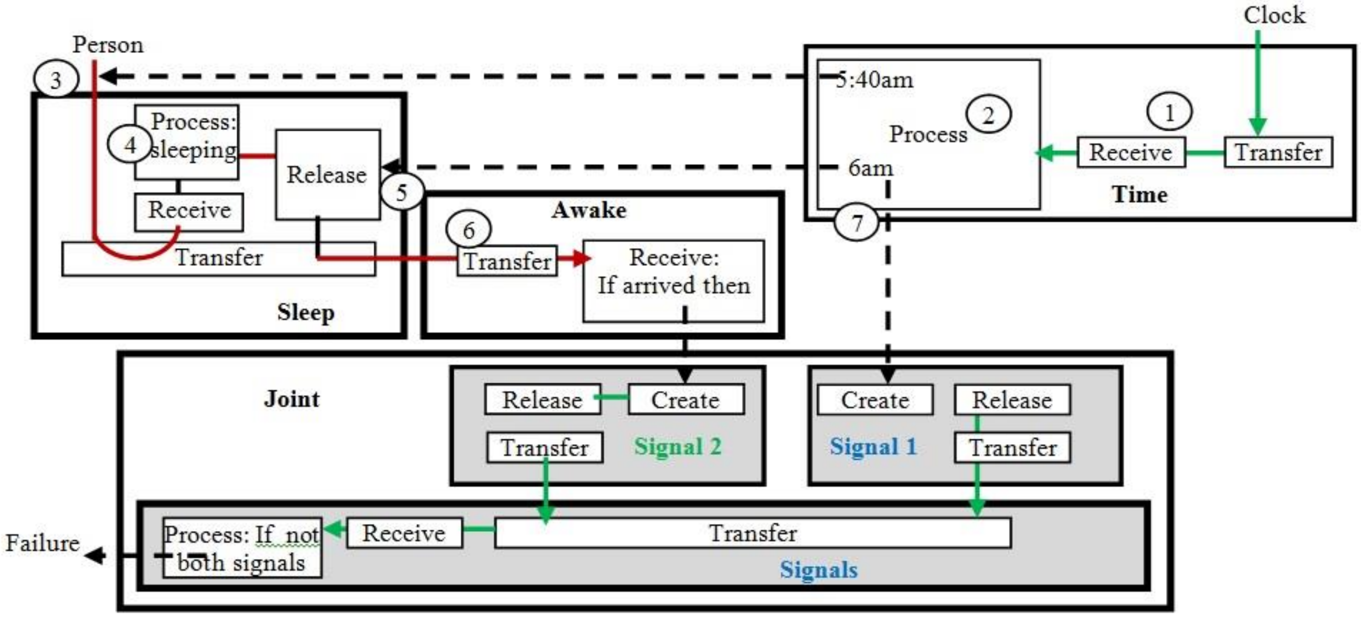


Fig. 13. FM representation without the joint bar

V. Timing and Real-Time Systems

Coffee machines have been used as a well-known example of modeling real-time systems using such languages as Uppaal and UML (e.g., [19–23]) In this section, we investigate the specification of design requirements for the coffee machine problem in the context of Uppaal, as it is described in many publications and course materials.

The coffee machine problem involves modeling the behavior of a system with three elements: a *machine*, *person*, and an *observer*. The person repeatedly inserts coins to receive coffee, after which he/she produces a publication. There is time delay after each such action. The machine takes some time for brewing the coffee.

It also takes a timeout if the brewed coffee is not taken before a certain upper time limit. The observer complains if more than 8 time units elapse between two consecutive publications.

In modeling the coffee machine in FM, we find that to complete the conceptual picture and flows, we need additional items (spheres) in addition to person, machine, and observer. One interesting aspect of FM description is the systematic application of the same generic stages for entities, subentities, and spheres. This repeatability of application creates specifications that are more complete. It is also possible to simplify the depiction by reducing the level of description in several ways. As an introduction, before giving the complete FM representation, Fig. 14 shows a brief description of the "waves" of flow and the new additional spheres.

In the figure, coin flow (A) triggers the coffee (B) and cup (C—a new sphere with an important role that will be explained later) flows as well as the flow of "counted time units" (D). As was mentioned previously, time flows continuously, but it is ignored until certain events (e.g., the arrival of coins) trigger counting units of time. Accordingly in the figure with the passing of the coffee preparation period, the coffee and the cup flow to the "filled cup compartment" (E) and start the "fill cup" flow (F). In this case, time is also counted (G), and if the filled cup does not flow (i.e., it is taken from the compartment), then this triggers dispensing. The flow of the filled cup outside the compartment (H) is supposed to trigger the flow of the coffee to the person (I—e.g., being drank). This in turn triggers producing publications (J) that flow to the observer (K). Upon the arrival of publications the observer starts counting time (L) and complains start to flow out (M) if time reaches its maximum without receiving new publications.

The completeness and continuity of events (technical and physical) are grounds for the validity of the model. Take the state-based modeling of the machine as given by Anderson [19] and represented in Fig. 15, according to Anderson [19–20]:

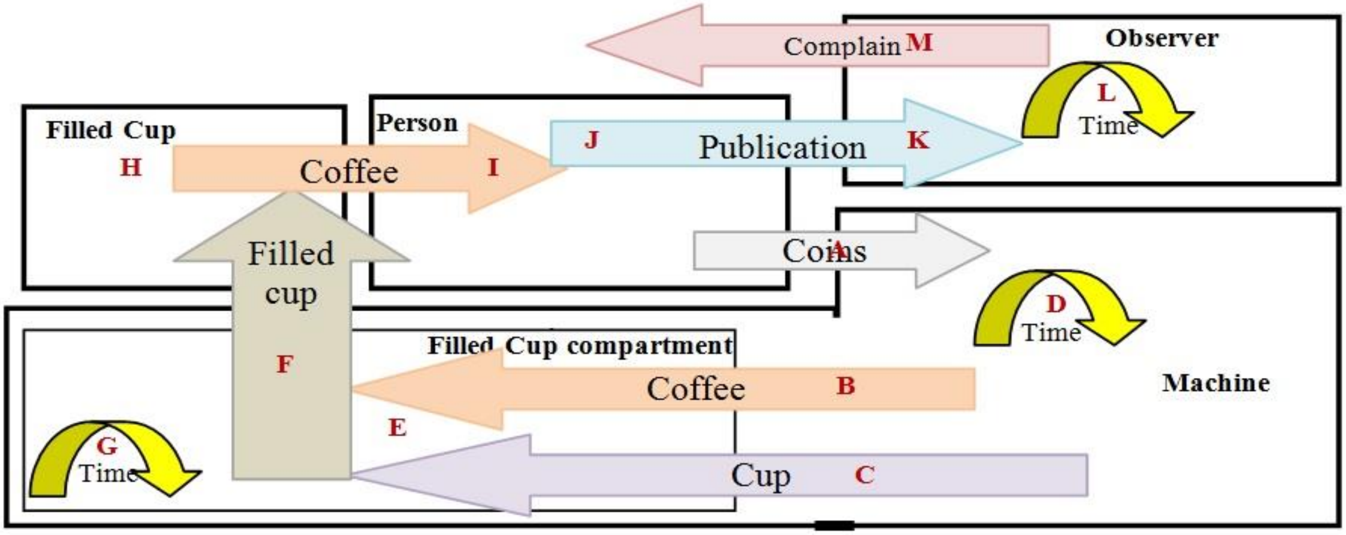


Fig. 14. Flows in the coffee machine problem

Coffee machine accepts a coin and then delays for some time (above it is 6 time units). It then sets a timeout timer, and either (to the right) dispenses coffee, or (to the left) times out and then dispenses coffee. The extra state on the left is because Uppaal does not allow both guards and synchronizing elements to appear on the same transition.

Note that this model assumes that the coffee *flows* outside the machine immediately, after the brewing process, just as water flows outside a pipe. This means the coffee does not *wait* to be taken outside the machine. The flow-based FM representation (see Fig. 16) forces introduction of a container for the coffee because there is waiting time. Thus, the items of *cup* and *filled cup* (cup+coffee) are necessary to convert the flowthing coffee from the state of liquidity (which makes its flow outside the machine compulsory) to the state of "handle-ability" (a thing that stands by itself waiting to be picked up). From the "state" perspective, Fig. 17 shows the two methods of conceptualization. On the left, the model is not based on flows, hence the conceptualization is represented by conceptual jumps from one state to another. On the right side, the FM model is casted in state jumps. In the figure, the two triggering arrows that come from outside the machine sphere change the waiting state.

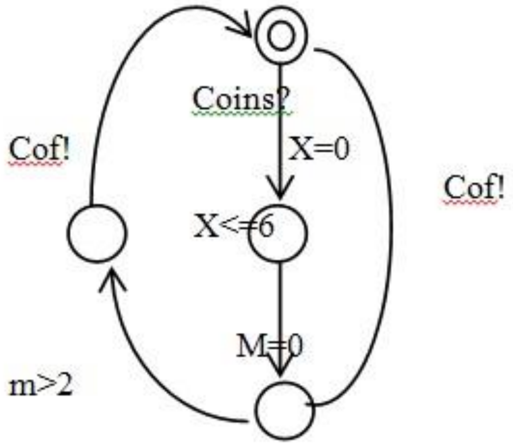


Fig. 15. Automata for machine (redrawn from [19])

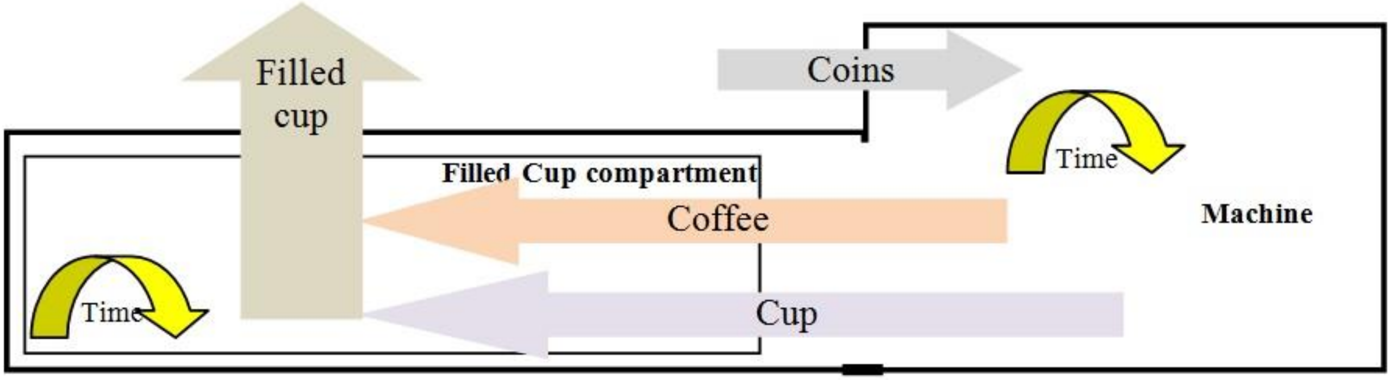


Fig. 16. Flows in the coffee machine problem

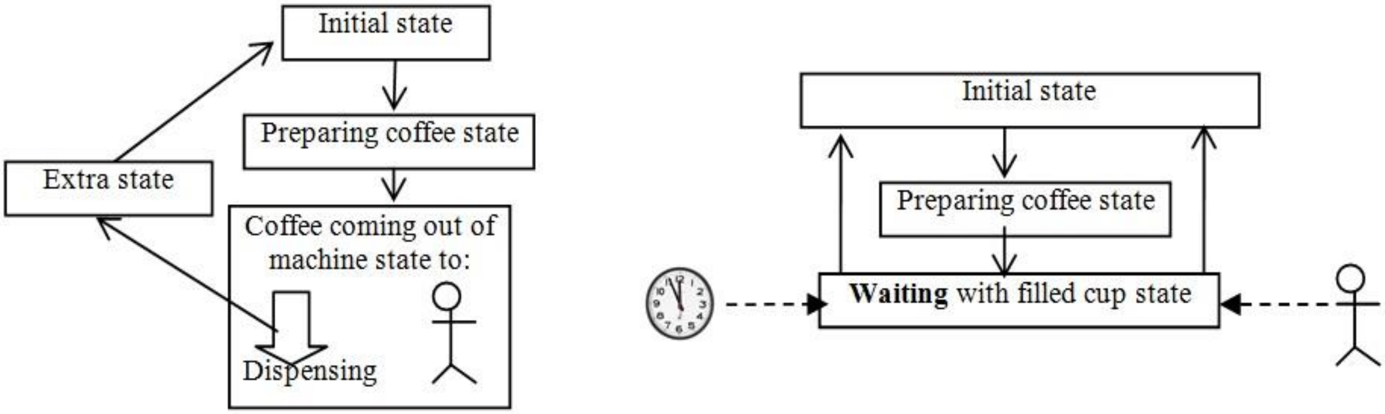


Fig. 17. The coffee problem described in terms of states

The point here is that the flow-based conceptualization “forces” continuity and completeness of the narration of events, thus identifying items (e.g., cups) and processes (e.g., waiting liquid).

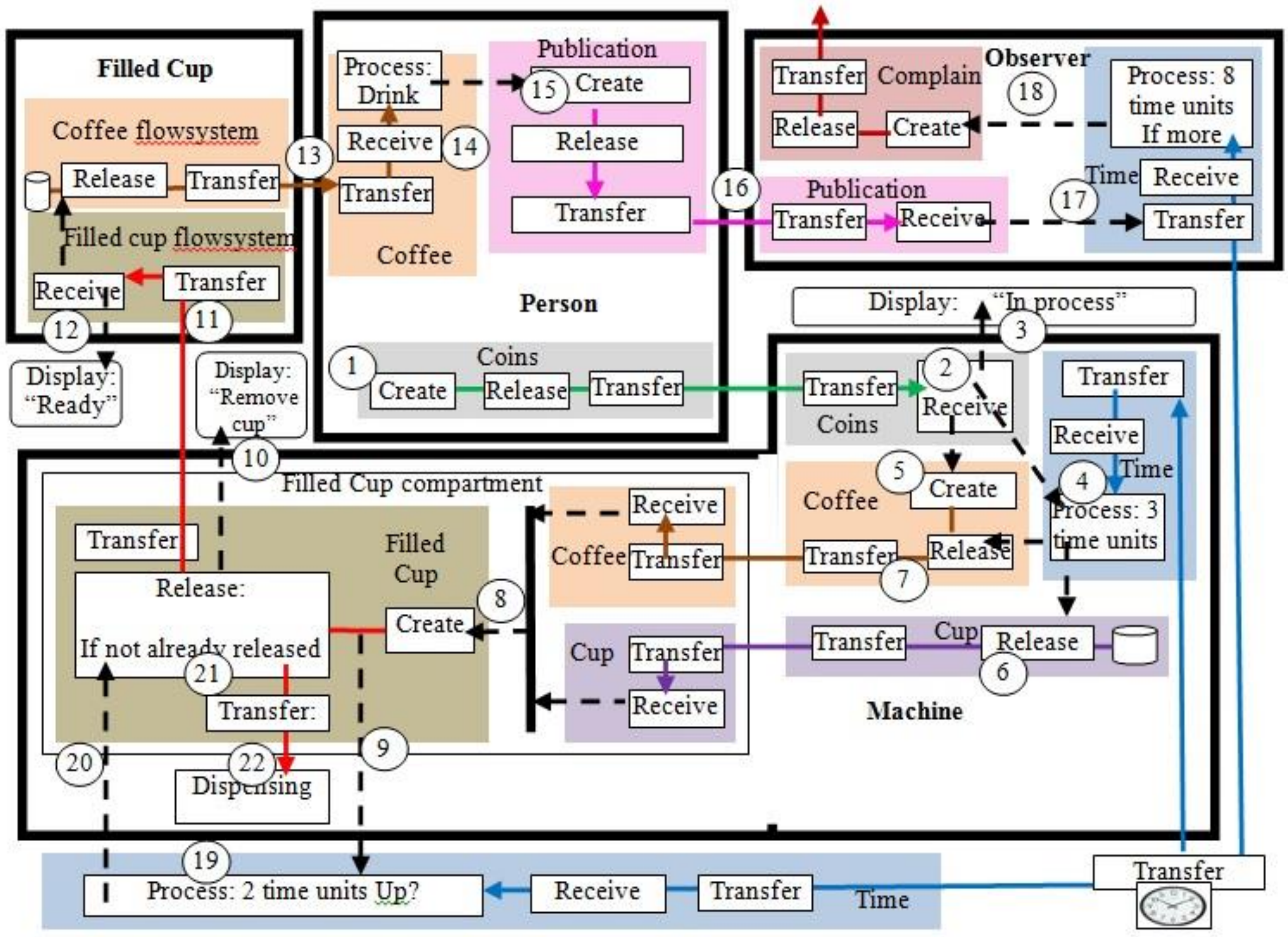


Fig. 18. The problem described in terms of states

Fig. 18 shows the complete FM representation of the coffee machine problem. We start with inserting the coins (circle 1). The *creation* here (in the figure) means the appearance of coins in the episode, just as the first appearance of a new character in a play in theater.

The coins flow to the machine (2) where they are received and trigger three events:

- Displaying "in process" to the user. Initially, we assume it displays "ready" (3).
- Triggering the time counter (4)
- Triggering preparing the coffee (5)

The machine is continuously receiving time units; however, the triggering makes it "pay attention" and count these time units. Note that the time sphere is represented by a clock picture for illustrative purposes, but it is really the flowsystem that creates time units. Also, it is possible to detail the coffee sphere by drawing flowsystems for the coffee powder and water separately to be processed and make coffee.

At the end of the coffee preparation time, the cup is dropped (6) and the coffee is released (7); this happens in the machine compartment subsphere to create the filled cup (8). Creating the filled cup and releasing it trigger waiting time (9) to pick up the cup and display that (10). If the person takes out the filled cup (11), this triggers displaying "ready" and triggers (12) the flow of coffee to the person (13).

Note that, in general, the filled cup sphere includes three subspheres: the filled cup (cup+coffee), coffee, and cup (see Fig. 19). In any sphere, we can focus on any of its subspheres. Accordingly, when the person removes the filled cup outside the compartment the "attention" (matters of interest) is on the filled cup and the coffee subspheres (flowsystems (11 and 12)), but the cup by itself is of no interest.

Continuing the flows, when the coffee is received by the person (14), he/she drinks it to trigger creation of publications (15) that flow to the observer (16), which in turn triggers initializing a waiting time period for the next publication (17). If no publication arrives, this triggers creation of a complaint (18). We assume that initially the waiting timing here is set to zero.

Returning to releasing the filled cup that triggers waiting time (9); if the waiting time is over (19), then this triggers (20) checking whether the filled cup has been already removed (21); if not, the filled cup is dispensed with (22).

VI. Conclusion

Methodologies of time representation can be based on states, UML, Petri nets, and other types of diagrams. Each has its own advantages and weaknesses, especially with regard to having the features of understandability and simplicity. This paper proposed a flow-based representation that is based on the notion of flow with a focus on exploring the representation of time. The new methodology was demonstrated through sample timing-related problems.

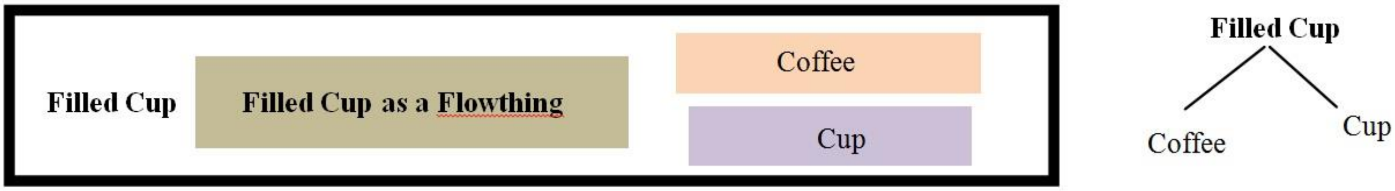


Fig. 19. The filled cup as a flowthing and its two flowthing components

FM can serve as an early system understanding and communication among stakeholders, including those without technical knowledge, and facilitate agreement between clients/users and designers. Additionally, it can be used as a base for system development and the design phase. The resultant FM representation avoids ambiguous textual language and heterogeneous diagramming. Of course, FM is still not well developed in comparison with such well diagram-oriented modeling methodology such as UML. Its weaknesses in terms of expressivity and complexity have to be studied more in different applications. Nevertheless, comparing FM diagrams side by side with other types of modeling techniques reveals it is a promising viable modeling tool.

We are currently exploring further time representation in FM, especially its relation to the actual design phase.

References

[1] H. Personnier, M.-A. le Dain, and R. Calvi. Failures in Collaborative Design with Suppliers: Literature Review and Future Research Avenues. 21st Annual IPSERA Conference **(2012)** Italy.

[2] Luqi and V. Berzims, Knowlede-Based Support for Rapid software Prototyping, IEEE Express, pp. 9–18 **(1988)**

[3] E. Klingerman and A. D. Stoyenko. Real-Time Euclid: A Language for Reliable Real-Time Systems, IEEE Trans. Softw. Engng., SE-12, 9, **(1986)** , pp. 941-949.

[4] S. Berryman and I. Sommerville. Modelling Real-Time Constraints. Proceedings of the 3rd International Conference on Software Engineering for Real-Time Systems, **(1991)**; Cirencester, UK

[5] G. K. Palshikar. An Introduction to Model Checking. Available from www.eetasia.com/ARTICLES/2005FEB/B/2005FEB16_EMS_ST_TA.pdf. **(2005)**

[6] T. A. Henzinger and J. Sifakis. The Embedded Systems Design Challenge. Proceedings of the 14th International Symposium on Formal Methods (FM), **(2006)**

[7] S. Al-Fedaghi, Pure Conceptualization of Computer Programming Instructions. International Journal of Advancements in Computing Technology, 3, 9 (2011)

[8] S. Al-Fedaghi and A. Alrashed, Threat Risk Modeling. International Conference on Communication Software and Networks (ICCSN), (2010) February 26–28; Singapore

[9] S. Al-Fedaghi. Flow-Based Enterprise Process Modeling. International Journal of Database Theory and Application Compendex, 6, 3 (2013)

[10] S. Al-Fedaghi. Schematizing Proofs Based on Flow of Truth Values in Logic. IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC), (2013) October 13–16; Manchester, UK

[11] S. Al-Fedaghi. An Alternative Approach to Multiple Models: Application to Control of a Production Cell. International Journal of Control and Automation, 7, 4 (2014)

[12] S. Al-Fedaghi. System for a Passenger-Friendly Airport: An Alternative Approach to High-Level Requirements Specification. International Journal of Control and Automation, 7, 2 (2014)

[13] K. Tindell and J. Clark. Holistic Schedulability Analysis for Distributed Hard Real-Time Systems. Microprocessing and Microprogramming, 40 **(1994)**

[14] W. Garnett. Letter to the Editor. Nature **(1920)**

[15] B. Strachan. The Skirts of Alpha, APPENDIX I: SOME THOUGHTS ABOUT DIMENSIONS. http://panpsychic-philosophy.org.uk/index.php

[16] Lucid Software Inc. 2, UML–Timing Diagram Tutorial. **(2014)** https://www.lucidchart.com/pages/UML-timing-diagram-tutorial

[17] webmaster@uml-diagrams.org., Timing Diagrams, **(2014)** http://www.uml-diagrams.org/timing-diagrams.html

[18] Lucid Software Inc, Timing Diagram, **(2014)** https://www.lucidchart.com/pages/uml/timing-diagram

[19] H. Anderson. CS5270 Verification of Real Time Systems, 6.2.2 Coffee Machine Example in Uppaal. http://www.comp.nus.edu.sg/~cs5270/Notes/chapt6a.pdf

[20] H. Anderson. Verification of Real-Time Systems. **(2007)** http://www.comp.nus.edu.sg/~cs5270/2006-semesterII/foils11.colour.pdf

[21] K. G. Larsen. Quantitative Model Checking: Real-Time Systems, Exercises. http://people.cs.aau.dk/~kgl/QMC2010/exercises/#coffee

[22] H. S. Hong, J. H. Kim, S. D. Cha, and Y. R. Kwon. Static Semantics and Priority Schemes for Statecharts. Computer Software and Applications Conference (COMPSAC), **(1995)**

[23] S. Van Langenhove. Towards the Correctness of Software Behavior in UML: A Model Checking Approach based on Slicing, Ph.D. thesis, Faculteit Wetenschappen — Universiteit Gent, **(2006)** http://lib.ugent.be/fulltxt/RUG01/000/970/662/RUG01-000970662_2010_0001_AC.pdf

Analysis of Heart Rate Variability by Applying Nonlinear Methods with Different Approaches for Graphical Representation of Results

Evgeniya Gospodinova, Mitko Gospodinov
Institute of Systems Engineering and Robotics
Bulgarian Academy of Sciences, Sofia, BULGARIA

Nilianjan Dey
Department of Computer Science and Engineering,
Bengal College of Engineering and Technology
Durgapur, West Bengal, INDIA

Ivan Domuschiev
II Department of Internal Diseases
Multiprofile Hospital for Active Treatment
“St.Panteleimon”, Plovdiv, BULGARIA

Amira S. Ashour
Department of Electronics and Electrical Communications
Engineering, Faculty of Engineering, Tanta University,
EGYPT. College of CIT, Taif University, KSA

Dimitra Sifaki-Pistolla
GIS expert in the Clinic of Social and Family Medicine, Faculty of Medicine,
University of Crete, GREECE

***Abstract*—There is an open discussion over nonlinear properties of the Heart Rate Variability (HRV) which takes place in most scientific studies nowadays. The HRV analysis is a non-invasive and effective tool that manages to reflect the autonomic nervous system regulation of the heart. The current study presents the results of HRV analysis based on 24-hour Holter ECG signals of healthy and unhealthy subjects. Analysis of heart intervals is performed with the use of original algorithms and software, developed by the authors, to quantify the irregularity of the heart rate. The main aim is the formation of the parametric estimate of patients' health status, based on mathematical methods that are applied on cardiac physiology. The obtained results show that the analysis of Holter recordings by nonlinear methods may be appropriate for diagnostic, forecast and prevention of the pathological cardiac statuses. Different approaches of graphical representation and visualization of these results are used in order to verify this.**

Keywords—Heart Rate Variability (HRV); ECG signal; Holter signal; nonlinear graphical methods

I. Introduction

Cardiovascular disease is among the top 10 causes of death worldwide. Ischemic heart disease, stroke and chronic obstructive lung have remained the top major causes of death for the past decade. Latest scientific research has shown that cardiovascular diseases can be mathematically analyzed and predicted by cardiac screening for early detection of possible complications to prevent disease or improve quality of life. The diagnostic parameter defined by the electrocardiograms (ECG) is the heart rate variability (HRV) which is calculated by measuring the time between heartbeats. In 1996, the European Society of Cardiology and North American Society of Pacing and Electrophysiology gave recommendations on the clinical usage of the HRV method for evaluating cardiology disease risks as: myocardial infarction (heart attacks), sudden cardiac death and essential hypertension. In addition, diabetes mellitus is associated with a reduced vagal tone and elevated sympathetic activity [1]. HRV is reduced in patients with diabetes mellitus, suggesting dysfunction of cardiac autonomic regulation which has been associated with increased risk for pathological cardiac events [2]. The main advantage of HRV signals is that it can be calculated real time in non-invasive manner, while all current biomarkers used in clinical practice are discrete and imply blood sample analysis [3].

Methods of HRV analysis are divided in two groups: linear and non-linear methods. Linear methods can be used in time- or frequency- domain for HRV analysis [4]. Time-domain parameters are statistical calculations of consecutive RR time intervals, and they are correlated with each other (SDNN, SDANN, pNN50, etc.). Frequency-domain parameters are based on spectral analysis. They are used to evaluate the contribution on HRV of autonomic nervous system (VLF, LF, HF, HF/LF ratio) [5- 7]. The last years, the interest for non-linear analysis is constantly increasing, because the HRV measurements are non-linear and non-stationary and a considerable part of information is coded in the dynamics of their fluctuation in different time periods. Through implementation of conventional (linear) mathematical methods part of the important characteristics of signal dynamics are missed. Therefore, development and implementation of new non-linear mathematical methods and knowledge, based on the fractal, multifractal and wavelet theory will allow scientists to identify new reasons for HRV fluctuations. This research is conducted on selected control groups of patients with cardiovascular diseases, forming part of an information database of 300 patients of a cardiological clinic; each with 24-

hour Holter recordings. Each record contains information of around 100,000 heartbeats. The first problem is based on the fact that fluctuation of the physiological signals possesses hidden information in the form of self-similarity, scale structure, monofractality and multifractality, through the application of these methods [8-13]. The fractal, multifractal and wavelet-based multifractal analysis of the fluctuations is useful not only for getting the comprehensive information for physiological signals of patients, but also provide a possibility for foresight, prognosis and prevention of the pathological statuses. The second problem is due to the large volume of research information is extremely important the correct determination the trend of the disease of the patients for relatively large period of time - several years, for instance. Such tests are performed periodically to compare the graphic characteristics of images from clinical studies undertaken as a result of treatment and give an idea of the patient's condition and treatment quality. Nonlinear analysis of cardiology data is a relatively new scientific approach to assess the dynamics of heart activity. The above described problems are solved by setting of the following interesting and important objectives:

- Presentation of some interesting results that depict the fractal and multifractal structure of RR intervals time series, extracted from 24-hour Holter ECG records of healthy and unhealthy subjects by applying the following non-linear methods: Detrended Fluctuation Analysis (DFA), Rescaled adjusted the range statistics plot (R/S), Wavelet Transform Modulus Maxima (WTMM) and Poincaré plot.
- Demonstration of different approaches of graphical representation of results from clinical studies of HRV and assessment of health status of patients with cardiovascular disease with multiple and periodic 24-hour Holter studies of treatment. Trends in the evolution of patients' health status are not possible to ascertain in short, five minute HRV records, described most often in specialized articles.

The key components of the proposed approaches are directed at the analysis of HRV by applying nonlinear methods for graphical representation of results, as new direction of the scientific investigations in parallel with research works with the linear methods.

II. Subject and Methods

A. Subjects

In this article two groups of signals are analyzed: RR time series of 16 normal subjects and 16 congestive heart failure (CHF) patients. These signals are consisting of around 100 000 data points, corresponding to 24-hour Holter ECG recordings.

B. Methods

Detrended Fluctuation Analysis (DFA)

DFA is a technique for detecting correlations in time series [14]. These functions are able to estimate several scaling exponents from the RR time series being analyzed. The scaling exponents characterize short or long-term fluctuations. The main steps of the DFA algorithm [15] are as follows:

Step 1: The RR interval time series is integrated using (1):

$$y(k) = \sum_{j=1}^{k}(RR_j - \overline{RR}),\ k = 1,\ldots,N. \qquad (1)$$

Where:

- $y(k)$ is the kth value of the integrated series;
- RR_j is the jth inter beat interval;
- $\overline{RR}$ is the average inter beat interval over the entire series.

Step 2: The integrated time series is divided into boxes of equal length n.

Step 3: In each box of length n, a least square line is fitted to the RR interval data and $y_n(k)$ denote these regression lines.

Step 4: The integrated series y(k) is detrended by subtracting the local trend in each box. The root-mean-square fluctuation of this integrated and detrended series is calculated using (2):

$$F(n) = \sqrt{1/N \sum_{k=1}^{N}(y(k) - y_n(k))^2}. \qquad (2)$$

Where: F(n) is a fluctuation function of box size n.

Step 5: The procedure is repeated for different time scales. The relationship on a double-log graph between fluctuations F(n) and the time scales n can be approximately evaluated by a linear model that provides the scaling exponent α given in (3):

$$F(n) \approx n^{\alpha}. \qquad (3)$$

The parameter α is depended by the correlation properties of the signal. By changing the parameter n can be studied how to change the fluctuations of the signal. Linear behavior of the dependence F (n) is an indicator of the presence of a scalable behavior of the signal. The slope of the straight line is used to determine the value of the parameter α. For uncorrelated signals, the value of this parameter is within the range (0, 0.5), where $\alpha > 0.5$-it is an indication for the presence of correlation. When $\alpha = 1$, the signal is 1 / f – noise, while $\alpha = 1.5$ – usually Brownian motion.

In the case of RR time series, DFA shows typically two ranges of scale invariance, which are quantified by two separate scaling exponents, α_1 and $\alpha_{2,}$ reflecting the short-term and long-term correlation [14]. The short-term fluctuation is characterized by the slope α_1 obtained from the (log n, log F(n)) graph within range $4 \leq n \geq 11$ and the slope α_2 obtained from the range $12 \leq n \geq 64$.

Rescaled adjusted range Statistics plot (R/S)

The rescaled range is a statistical measure of the variability of a time series introduced by British hydrologist Harold Hurst [16]. The Hurst exponent is one closely associated method with the R/S . This exponent is a measure that has been widely used to evaluate the self-similarity and correlation properties of fractional Brownian noise, the time series produced by a fractional Gaussian process. The self-similarity means that the statistical properties (all moments) of a stochastic process do not change for all aggregation levels. The main properties of self-similar processes include:

- Slowly decaying variance – the variance of the sample is decreased more slowly than the reciprocal of the sample size.
- Long-range dependence - the process is called a stationary process with longrange dependence if its autocorrelation function is non-summable. The speed of decay of autocorrelations is more hyperbolic than exponential.
- Hurst effect – it expresses the degree of self-similarity.

Based on the Hurst exponent value, the following classifications of time series can be realized:

- H=0.5 indicates a random series;
- 0<H<0.5 – the data in the signal are anti-correlated;
- 0.5<H<1 – the data in the signal are long-range correlated.

The R/S method for the time series X(n) is asymptotically given by a power law:

$$R(n)/S(n) \infty n^{H} . \tag{4}$$

Where:

- R(n) is the range which is the difference between the minimum and maximum accumulated values;
- S(n) is the standard deviation estimated from the observed data X(n);
- H is the Hurst exponent.

To estimate the Hurst exponent is plotted R(n)/S(n) versus n in log-log axes. The slope of the regression line approximates the Hurst exponent [17].

Wavelet Transform Modulus Maxima

One of the most popular tools in wavelet-based multifractal analysis is the Wavelet Transform Modulus Maxima (WTMM) method. This method is based on wavelet analysis that is called "mathematical microscope" due to its ability to maintain good resolution at different scales [18]. The WTMM method is a powerful tool for statistical description of non-stationary signals, because the wavelet functions are localized in time and frequency.

Physiological signals, such as RR time series, can be efficiently represented by decomposition at different frequencies. The conventional method for this approach is a Fourier analysis. This analysis works well for stationary time series, but not for non-stationary signals, when the frequency content changes over time. Scalograms is based on the wavelet transform simultaneously to provide both time and frequency information and is important for investigating the signal. The WTMM method is used for analysis the multifractal scaling properties of fractal signals. This method uses continuous wavelet transform to detect singularities of a signal. WTMM is based on Wavelet analysis (continuous wavelet transform, skeleton construction) and Multifractal formalism (partition function calculation, scaling exponential function estimation, multifractal spectrum estimation). The method consists in the following basic steps [19, 20].

Step 1: Calculation of the Continuous Wavelet Transform (CWT)

In many previous researches, the wavelet decomposition has been used in the medical domain [21- 27] and in other domains [28- 30]. A wavelet is simply a finite energy function with a zero mean value. The wavelet transform is defined by the continuous time correlation between the time series and the particular wavelet of scaling parameter τ and shift parameter α:

$$W(\tau(\alpha) = \int_{-\infty}^{+\infty} f(t)\,\Psi_{\tau,\alpha}(t)dt . \tag{5}$$

Where the analyzing wavelet $\psi_{\tau,\alpha}(t)$ is a zero average function with local support, centered around zero. The family of wavelet vectors is obtained by the translations and dilatations of the "mother" wavelet:

$$\psi_{\tau,\alpha}(t) = \frac{1}{\sqrt{\alpha}}\,\psi\left(\frac{t-\tau}{\alpha}\right) \tag{6}$$

The wavelet transform has a time frequency resolution which depends on the scale α.

Step 2: Calculation of the local maxima of the modulus of the CWT

The modulus maxima (largest wavelet transform coefficients) are found at each scale α as the supreme of the computed wavelet transforms such that:

$$\frac{\partial W(\tau(\alpha)}{\partial \tau} = 0. \tag{7}$$

Step 3: Calculation of the partition function Z(q,α) based on wavelets, where α is the dilatation and q is a scale factor

The originality of the WTMM method is in the calculation of the partition function Z(q,α) from the maxima lines. The space-scale partitioning given by the wavelet tiling or skeleton defines the particular partition function:

$$Z(q,\alpha) = \sum_{\tau,\alpha} \sup_{\alpha} \left|W(\tau(\alpha)\right|^{q} . \tag{8}$$

Where α is the dilatation and q is a scale factor. This partition function effectively computes the moments of the absolute values of the wavelet resonance coefficients W(τ,α).

Step 4: Calculation of the decay scaling exponent τ(q)

The scaling exponent τ(q) is the Legendre Transform of the multifractal spectrum f(α) for self-similar time series and relates the fractal dimensions to the order q of the partition function Z(q,α). The slope of the double-logarithmic plot allows the computation of the decay scaling exponent. This slope can be obtained using linear regression:

$$\log_2 Z(q,\alpha) \approx \tau(q)\log_2 \alpha + C(q) . \tag{9}$$

If the scaling exponent is everywhere convex, that indicates multifractal behaviour of the signal. In case of monofractal behaviour, the scaling function is a line.

Step 5: Estimation of the spectrum of singularities

Multifractal formalism uses multifractal spectrum for the detailed fractal analysis of the signal. The Multifractal spectrum function shows the scope of all fractal measures. The Multifractal spectrum function is calculated from scaling function via Legendre transform by formulas:

$$\alpha(q)=\frac{d\tau(q)}{dq} \quad \text{and} \quad f(\alpha)=\min_{\alpha}(\alpha q\text{-}\ \tau(q)). \qquad (10)$$

The spectrum width on multifractality degree is $\Delta\alpha=\alpha_{max}-\alpha_{min}$, this quantity is a measure of the range of fractal exponents in the time series, so if Δα is large, the signal is multifractal.

Poincaré plot

The Poincaré plot analysis is a graphical nonlinear method to assess the dynamic of HRV [31, 32]. The method provides summary information as well as detailed beat-to-beat information on the behaviour of the heart. It is a graphical representation of temporal correlations within the RR intervals derived from ECG signal. The Poincaré plot is known as a return maps or scatter plots, where each RR interval from time series RR = {RR1, RR2, …, RRn, RRn+1} is plotted against next RR interval. The Poincaré plot parameters used in this paper are SD1, SD2 and SD1/SD2 ratio. SD2 is defined as the standard deviation of the projection of the Poincaré plot on the line of identify (y=x) and SD1 is the standard deviation of projection of the Poincaré plot on the line perpendicular to the line of identify. These parameters can be defined by (11), (12) and (13).

$$x=\{x1, x2, \dots, xn\}=\{RR1, RR2, \dots, RRn\} \qquad (11)$$

$$y=\{y1, y2, \dots, yn\}=\{RR2, RR3, \dots, RRn+1\} \qquad (12)$$

$$SD1=\sqrt{\mathrm{var}(d_1)}\ ;\ SD2=\sqrt{\mathrm{var}(d_2)}\ ;\ Ratio=\frac{SD1}{SD2}\ . \qquad (13)$$

Where:

- i = 1, 2, 3, …, *n* and *n* is the number of points in the Poincaré plot;
- var(d) is the variance of d;
- $d_1=\frac{x-y}{\sqrt{2}}$; $d_2=\frac{x+y}{\sqrt{2}}$.

Parameter SD1 has been correlated with high frequency power, while SD2 has been correlated with both low and high frequency powers. The ratio SD1/SD2 is associated with the randomness of the HRV signal. It has been suggested that the ratio SD1/SD2, which is a measure of the randomness in HRV time series, has the strongest association with mortality in adults.

III. Results and Discussion

The analyzed data for describing two types of signals are combined in two groups: records of 16 CHF patients and 16 normal subjects.

Fig. 1(a) and Fig. 1(b) distribute the RR interval signals of normal subject and CHF patient. The graphs of RR time data are highly nonstationary (the mean, variance and other variation in the time statistical parameters). The fluctuations of heart-beat time series are larger in healthy subject compared to CHF patient.

Fig. 2(a) and Fig. 2(b) illustrate the values of scaling exponents and the slope of the line F(n) on double logarithmic plot obtained by using the DFA method for the investigation of signals. The results indicate a significant difference between patients with CHF and healthy controls in short- and long- time scales. Healthy subjects typically show the fractal behavior of heartbeat dynamics while patients with CHF show an alteration in fractal correlation properties.

The results of the R/S method applied to the studied signals to determine the value of the Hurst exponent are shown in Fig.3 (a) and Fig. 3 (b). The obtained results show that RR time series are correlated, i.e. they are fractal time series. For normal subject, the value of Hurst exponent is high due to the variation being chaotic, and for CHF patient this value decreases because the RR variation is low.

The wavelet decomposition of RR sig.nals can be used to provide a visual representation of the fractal structure of the investigated signals (Fig. 4 (a), Fig. 4 (b)). The wavelet coefficients are presented in their absolute values and coloured in accordance with colour bar. Dark colours correspond to lower absolute wavelet coefficient values. Light colours indicate higher absolute wavelet coefficient values corresponding to large heartbeat fluctuations. The wavelet analysis uncovers hierarchical scale invariance and reveals a self-similar fractal structure in the healthy subjects and a loss of this fractal structure in the unhealthy (CHF) patients.

On Fig.5 (a) and Fig. 5 (b) show the variations of scaling exponents τ(q) over all the values of q for the normal subject and CHF patient. The constantly changing curvature of the τ(q) curves for the normal subject suggests multifractal behaviour. In contrast, τ(q) is a straight line for CHF patient, indicating monofractal behaviour.

Fig. 6 (a) illustrates the multifractal spectrum of RR time series of normal subject and Fig. 6 (b) for CHF patient. The curve of normal subject has multifractal behaviour due to the wide range of local values of the Hölder exponent α ($\Delta\alpha=\alpha_{max}-\alpha_{min}$). The range of values of the Hölder exponent α for RR time series of healthy subject is Δα=0.79177, and for RR time series of CHF patient is Δα=0.30351. The interval *Δα*, corresponding to the RR signal for CHF patient is more than two times lower than the signal for a normal subject.

The results of the Poincaré plot analysis of RR time series of healthy subject and CHF patient are presented in Fig.7 (a) and Fig.7 (b). The Poincaré plot for healthy subject is a cloud of points in the shape of an ellipse (‘comet’ shaped plot). On the other hand, points for CHF patient are a cloud of points in the shape of a circle (‘complex’ shaped plot). The geometry of these plots has been shown to distinguish between healthy and unhealthy subjects. The obtained Poincaré plot parameters are directly related to the physiology of the heart. The parameter SD1 is the length of the semi-minor axis of the ellipse and it is the reflection of short term variability of heart rate. The parameter SD2, the length of the semi-major axis, is the measure of long term variability. Since the values of the SD1 and SD2 parameters depend on the RR intervals, the ratio of SD1 to SD2 is used to make comparison among Poincaré plots from different subjects. The values of SD1 and SD2 are higher in normal subjects comparatively to the congestive heart failure subjects.

The modern measuring devices (ECG Holters) and computational methods for data analysis allow the researcher and medical expert derive various summary parameters, analyze recurring patterns, transform the data and much more [33-37].

The values of the investigated parameters (mean ± standard deviation) are reported in Table 1. The described methods are realized by the Matlab software developed in the research project for nonlinear analysis of ECG signals.

There are some limitations in the using of the described analytical system. First, in the study are include restricted number of investigated subjects for analysis of RR time series (16 normal subjects and 16 CHF patients). Secondly, the study is oriented only at nonlinear methods for analysis of RR time series. These limitations are included because this paper is prepared not for showing the medical treatment results, but for demonstration the conceptual developments of the software environment and database as a new and modern tools to analyze, quantify and display the results of nonlinear analysis of HRV signals.

Concerning the practical application of the above described nonlinear methods of graphical representation, the graphical user interface of the software environment, database and appropriate signal processing algorithms were developed and implemented in MATLAB 6.0. The database includes background parameters, for example subject's gender, age and level of physical activity as well as disease history. The variations are presented by measurement the time of day and describing the circadian rhythm of the subject. The graphical representation of the circadian rhythm for every subject is stored in database, which enable evaluation of the results in a relatively long period of time for the post-analysis of the prescribed treatment, for instance. Moreover, the circadian rhythm of the physiological states can be assessed with this database's results.

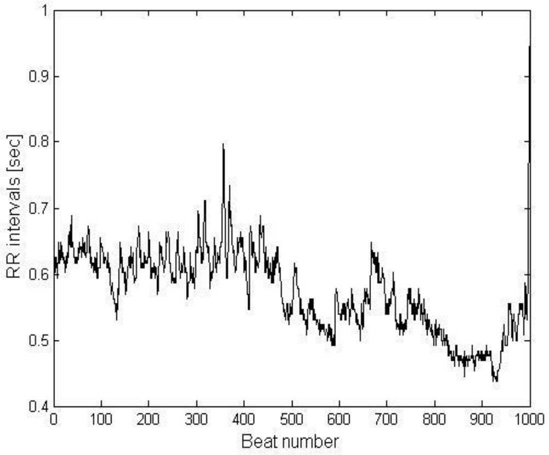


Fig. 1. (a). RR interval series of healthy subject

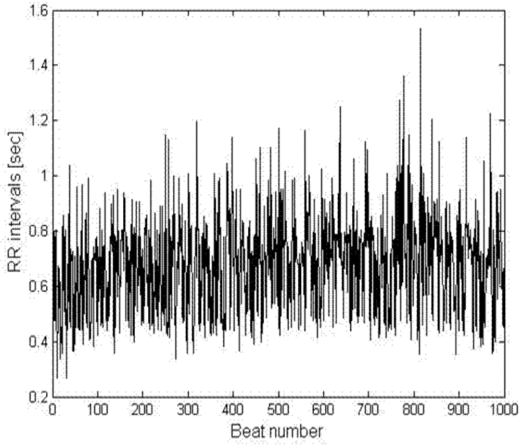


Fig. 1. (b). RR interval series of CHF patient

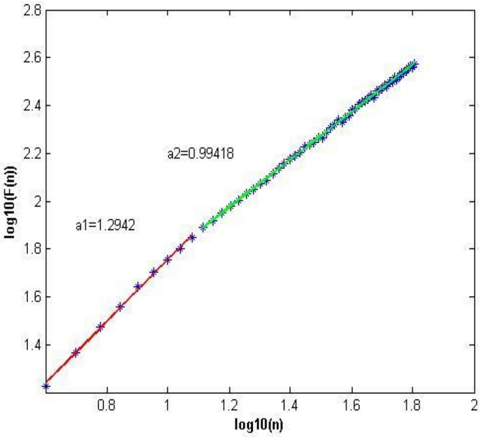


Fig. 2. (a). DFA analysis of healthy subject

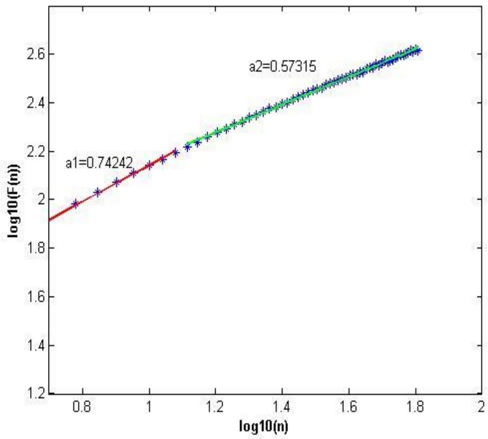


Fig. 2. (b). DFA analysis of CHF patient

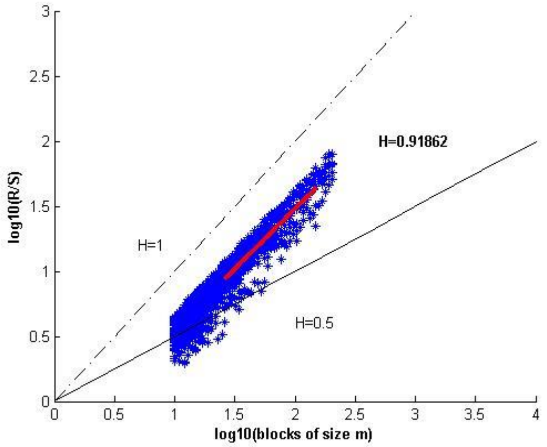


Fig. 3. (a). R/S analysis of healthy subject

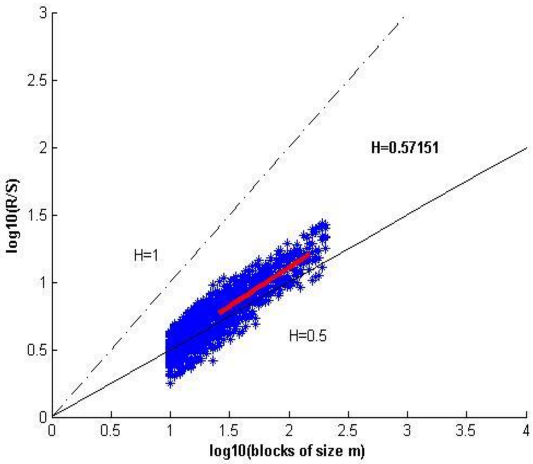


Fig. 3. (b). R/S analysis of CHF patient

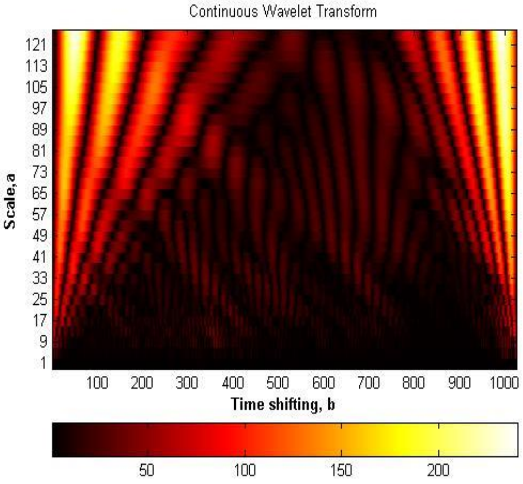


Fig. 4. (a). Continuous wavelet transform of healthy subject

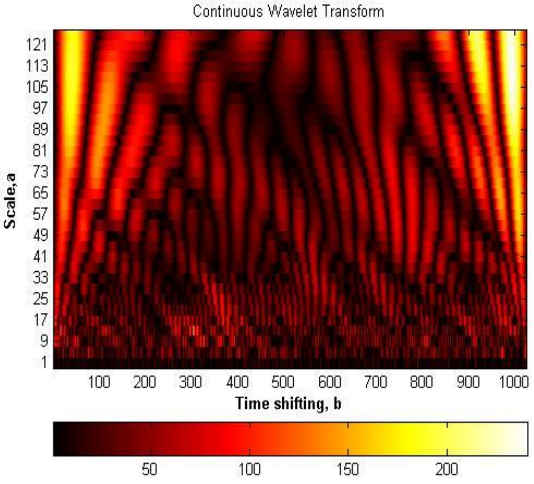


Fig. 4. (b). Continuous wavelet transform of CHF patient

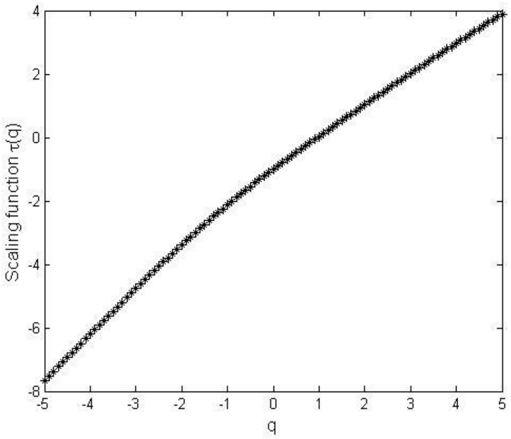


Fig. 5. (a). Scaling exponent τ(q) for RR series of healthy subject

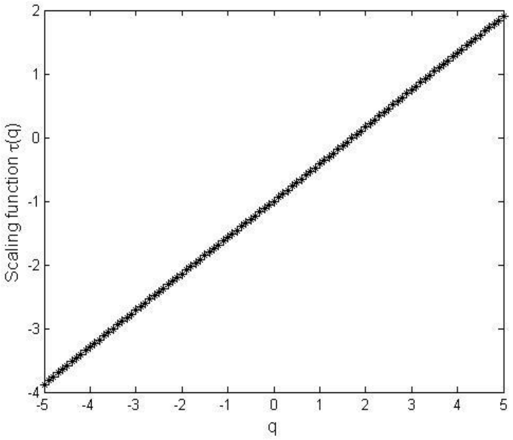


Fig. 5. (b). Scaling exponent τ(q) for RR series of CHF patient

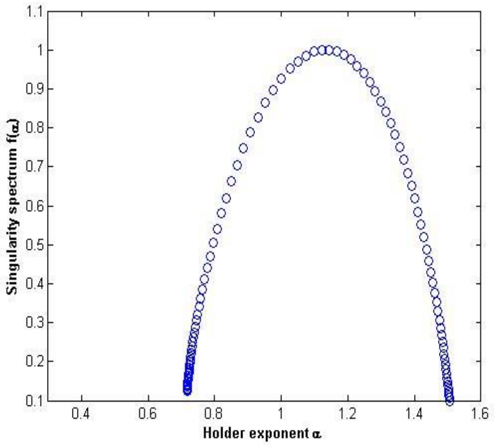


Fig. 6. (a). Multifractal spectrum for RR series of healthy subject

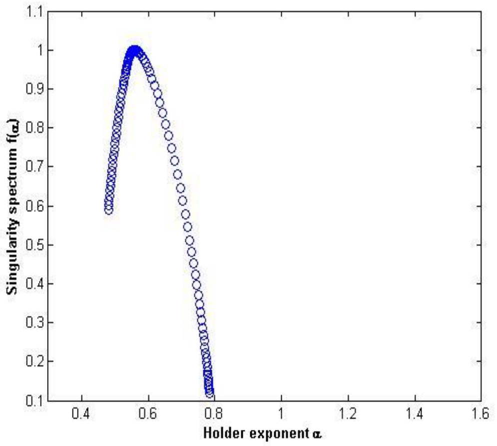


Fig. 6. (b). Multifractal spectrum for RR series of CHF patient

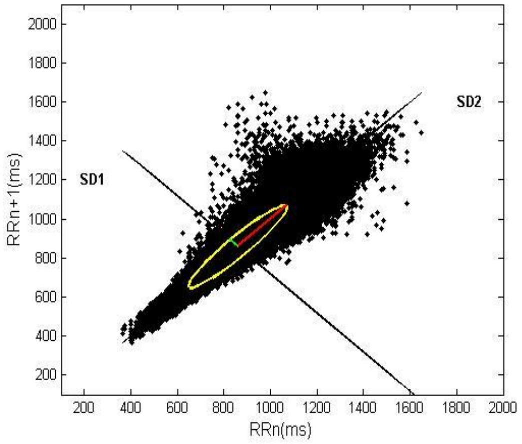


Fig. 7. (a). Poincaré plots for RR series of healthy subject

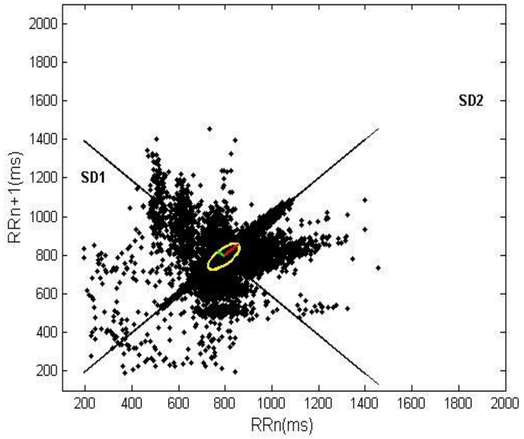


Fig. 7. (b). Poincaré plots for RR series of CHF patient

TABLE I. PARAMETERS FOR HEALTHY SUBJECTS AND CHF PATIENTS

Parameter	**Healthy subjects**	**CHF patients**
alpha 1 (DFA)	1.2942±0.205	0.7424±0.311
alpha 2 (DFA)	0.9942±0.319	0.5731±0.376
alpha (DFA)	1.0578±0.198	0.6192±0.231
Hurst (R/S)	0.9186±0.018	0.5715±0.219
q=0, **α** (WTMM)	1.1214±0.9	0.55752±0.4
q=10, **α** (WTMM)	0.7167±0.6	0.48008±0.3
q=-10, **α** (WTMM)	1.5085±0.8	0.78359±0.5
q=0, **f(α)** (WTMM)	1.0	1.0
q=10, **f(α)** (WTMM)	0.12414±0.08	0.59004±0.26
q=-10, **f(α)** (WTMM)	0.06879±0.07	0.12029±0.16
Δ**α** (WTMM)	0.79177±0.15	0.30351±0.09
SD1 [ms] (Poincare plot)	46.393±22.02	35.123±18.23
SD2 [ms] (Poincare plot)	295.295±35.9	87.5673±29.12
SD1/SD2 (Poincare plot)	0.15711±0.32	0.40121±0.41

IV. CONCLUSION

The paper presents a part of the latest scientific investigations and the results of application of nonlinear mathematical methods, based on the fractal theory for selected group of patients. The used nonlinear graphical methods for HRV analysis are an effective tool to visualize HRV fluctuations. The results of the cardiac data and the selected methods of analysis managed to give detailed information about the physiological status of the patients and to develop a work towards a framework for prognosis and prevention of pathology status in case of cardiovascular disease. This proves the importance of the proposed study.

The use of graphic-oriented methods for analysis and visualization of cardiac diseases facilitates decision-making by health professionals, especially in case of large amounts of information such as the analysis of diurnal heart rhythms.

Furthermore, analysis of trends of patients with cardiovascular diseases is more easily run by periodically comparing the plots of about 100.000 entries heart rate, rather than to make comparisons of numerical values or ECG data.

Graphic images of the clinical studies for each patient are automatically stored in a database with easy access and the ability to compare the periodic 24-hour recordings. The study of patient data jointly with the application and comparison of graphics obtained from several nonlinear mathematical methods is recommended.

REFERENCES

[1] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *Heart rate*

variability: standards of measurement, physiological interpretation, and clinical use. Circulation, 1996; 93:1043-1065.

[2] M. P. Tarvainen, D.J. Cornforth, P. Kuoppa, J. A. Lipponen, H. F. Jelinek, "*Complexity of Heart Rate Variability in Type 2 Diabetes-Effect of Hyperglycemia"*, 35th Annual International Conference of the IEEE EMBS, Osaka, Japan, 3-7 July, 2013.

[3] F. Buccelletti, M. G. Bocci, E. Gilardi, V. Fiore, S. Calcinaro, C. Fragnoly, R. Maviglia and V. Franceschi, "*Linear and Nonlinear HRV Indexes in Clinical Practice"*, Computational and Mathematical Methods in Medicine, 2012; vol. 2012.

[4] H.Kadat, V. Akkava, A. B. Sozen, S. Salman, S. Demirel, M. Ozcan, D. Atilqan, M. T. Yilmaz, O. Guven, "*Heart rate variability in diabetes patients"*. J Int Med Res., 2006; 34(3): 291-296.

[5] M. Mirza, A. N. K. Lakshmi, "*A comparative study of Heart Rate Variability in diabetic subjects and normal subjects"*, International Journal of Biomedical and Advance Research, 2012; 3(8): 640-644.

[6] G. Ernst, *Heart Rate Variability*, Springer-Verlag London, 2014.

[7] U. R. Acharya, J. S. Suri, J. A. E. Spaan, S. M. Krishnan, "*Advances in Cardiac Signal Processing"*, Springer-Verlag Berlin Heidelberg , 2007.

[8] D. M. Kumar, S.C. Prasannakumar, B. G. Sudarshan, D. Jayadevappa, "*Heart Rate Variability Analysis: A Review"*. International Journal of Advanced Technology in Engineering and Science, 2013; 1(6): 9-24.

[9] P. Ivanov, L. Amaral, A. Goldberger, S. Halvin, M. Rosenblum, H. Stanley, Z. Struzik, "*Multifractality in human heartbeat dynamics"*. Macmillan Magazines Ltd, 1999; 461-465.

[10] P. Ivanov, Z. Chen, K. Hu, H. Stanley, "*Multiscale aspects of cardiac control"*. Physica A 344 (3–4), 2004; 685–704.

[11] P. Ivanov, L. Amaral, A. Goldberger, S. Halvin, M. Rosenblum, H. Stanley, Z. Struzik, "*From 1/f noise to multifractal cascades in heartbeat dynamics"*. Chaos, 2001; 11(3): 641-652.

[12] H. E. Stanley, L. A. N. Amaral, A. L. Goldberger, S. Havlin, P. Ch. Ivanov, C. -K. Peng, "*Statistical physics and physiology: Monofractal and multifractal approaches"*. Elsevier, Physica A 270, 1999; 309-324.

[13] J. Wang, Y. Ning, Y. Chen, "Multifractal analysis of electronic cardiogram taken from healthy and unhealthy adult subjects". Physica A 323, 2003; 561-568.

[14] C.K. Peng, S. Halvin, H. E. Stanley, A. L. Goldberger, "*Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series*". Chaos, 1995; 5: 82-87.

[15] K. Rawal and I. Sain, *"Comparative Analysis of Measuring Heart Rate Variability during Different Phases of Menstrual Cycle in Young Healthy Women"*, International Journal of Information and Electronics Engineering, 2014; 4(1): 62-66.

[16] H. E. Hurst, Black and Y. M. Sinaika, *"Long-term Storage in Reservoirs: An experimental Stud"*, Constable, London, 1965.

[17] M. Gospodinov, E. Gospodinova, *"The graphical methods for estimating Hurst parameter of self-similar network traffic"*. International Conference on Computer Systems and Technologies, 2005; IIIB.19-1—IIIB.19-6.

[18] A. Puckovs, A. Matvejevs, "*Wavelet Transform Modulus Maxima Approach for World Stock Index Multifractal Analysis"*. University, Information Technology and Management Science, 2012; 76-86.

[19] C. Los, R. Yalamova, "*Multifractal Spectral Analysis of the 1987 Stock Market Crash"*. Social Science Research Network, 2004; 1-21.

[20] C. TAŞ, Dr. G. Ünal, "*Multifractal Behavior in Natural Gas Prices by using Mf-DFA and WTMM Methods"*. Global Journal of Management and Business Research Finance, USA, 2013; 13(11: 1-5.

[21] N. Dey, A. Das, S. S. Chaudhuri " *Wavelet Based Normal and Abnormal Heart Sound Identification Using Spectrogram Analysis*", International Journal of Computer Science & Engineering Technology (IJCSET), 2012; 3(6).

[22] N. Dey, P. Das, A. Das, S.i S. Chaudhuri, **"***DWT-DCT-SVD Based Intravascular Ultrasound Video Watermarking*", Second World Congress on Information and Communication Technologies (WICT 2012), Trivandrum, India, 2012.

[23] N. Dey, P. Das, A. Das, S. S. Chaudhuri, "Nilanjan Dey, Poulami Das, Achintya Das, Sheli Sinha Chaudhuri, **"***DWT-DCT-SVD Based Blind Watermarking Technique of Gray Scale Image in Electrooculogram Signal***",** International Conference on Intelligent Systems Design and Applications (ISDA-2012), Kochi, 2012 .

[24] N. Dey, S. Biswas , A. Roy, A. Das, S. S. Chaudhuri, *"Analysis Of Photoplethysmographic Signals Modified by Reversible Watermarking Technique using Prediction-Error in Wireless Telecardiology"*, International Conference of Intelligent Infrastructure, 47th Annual National Convention of CSI –Kol, 2012.

[25] N. Dey, S. Biswas , P. Das, A. Das, S. S. Chaudhuri, *"Lifting Wavelet Transformation Based Blind Watermarking Technique of Photoplethysmographic Signals in Wireless Telecardiology"*, Second World Congress on Information and Communication Technologies (WICT 2012), Trivandrum, India, 2012.

[26] N. Dey, G. Mishra, B. Nandi , M. Pal, A. Das, S. S. Chaudhuri*,"Wavelet Based Watermarked Normal and Abnormal Heart Sound Identification using Spectrogram Analysis"*, 2012 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC),Tamilnadu College of Engineering,Coimbatore,India, 2012.

[27] N. Dey, P. Maji , P. Das, A. Das, S. S. Chaudhuri, *"An Edge Based Watermarking Technique of Medical Images without Devalorizing Diagnostic Parameters"*, International Conference on Advances in technology and Engineering, NMIMS University, Mumbai, India, 2013.

[28] T. Bhattacharya, N. Dey, S. R. B. Chaudhuri, "*A Novel Session Based Dual Image Encoding and Hiding Technique Using DWT and Spread Spectrum"*, International Journal on Computer Science and Engineering, 2012; 3(11):3510-3517.

[29] N. Dey, A. Roy, S. Dey*,"A Novel Approach of Color Image Hiding using RGB Color planes a*nd DWT", International Journal of Computer Applications, 2011; 36(5).

[30] T. Bhattacharya, N. Dey, S. R. B. Chaudhuri, "*A Session based Multiple Image Hiding Technique using DWT and DCT*", International Journal of Computer Applications, 2012; 38(5).

[31] R. L. Smith, E. R. Wathen, P. C. Abaci, N. H. V. Bergen, I. H. Law, M. D. Dick II, C. Connor, E. L. Dove, "*Analyzing Heart Rate Variability in Infants Using Non-Linear Poincare Techniques*". Computer in Cardiology, 2009; 36:673-876.

[32] Md. Rhaman, A. H. M. Karim, M. Hasan, J. Sultana, "Successive RR Interval Analysis of PVC with Sinus Rhythm Using Fractal Dimension, Poincaré plot and Sample Entropy Method". I.J.Image, Graphics and Signal Processing, 2013; 2:17-24.

[33] N. Dey, S. Samanta, XS. Yang, SS Chaudhri , A Das, "*Optimisation of Scaling Factors in Electrocardiogram Signal Watermarking using Cuckoo Search*", International Journal of Bio-Inspired Computation (IJBIC), 2013; 5(5):315-326.

[34] N. Dey, S Mukhopadhyay, A Das, SS Chaudhuri, "*Analysis* of *P-QRS-T Components Modified* by *Blind Watermarking Technique Within* the *Electrocardiogram Signal* for *Authentication* in *Wireless Telecardiology using DWT*", I.J. Image, Graphics and Signal Processing (IJIGSP), ISSN:2074-9074, 2012; 7:33-46.

[35] N. Dey, AB. Roy, A. Das, SS. Chaudhuri,"*Stationary Wavelet Transformation Based Self-Recovery of Blind-Watermark from Electrocardiogram Signal in the Wireless Telecardiology*", International Workshop on Intelligence and Security Informatics for International Security (IIS'12) , Trivandrum, India, 2012.

[36] N. Dey, B. Nandi, M. Dey, A. Das , SS. Chaudhuri, "*BioHash Code Generation from Electrocardiogram Features*", 3rd IEEE International Advance Computing Conference (IACC-2013) Ghaziabad, India, 2013.

N. Dey, B. Nandi, P. Das, A. Das, SS. Chaudhuri, **"***Retention Of Electrocardiogram Features Insignificantly Devalorized As An Effect Of Watermarking For A Multi-Modal Biometric Authentication System***"**, published by "Advances in Biometrics for Secure Human Authentication and Recognition", CRC Press, Taylor and Francis, 2013, 450.

Diagnosis of Wind Energy System Faults Part I : Modeling of the Squirrel Cage Induction Generator

Lahcène Noureddine
Department of Electrical Engineering
Amar Thlidji University
Laghouat, Algeria

Omar Touhami
Research Laboratory of Electrotechnics
High National Polytechnic College
Algiers, Algeria

***Abstract*—Generating electrical power from wind energy is becoming increasingly important throughout the world. This fast development has attracted many researchers and electrical engineers to work on this field. The authors develop a dynamic model of the squirrel cage induction generator exists usually on wind energy systems, for the diagnosis of broken rotor bars defects from an approach of magnetically coupled multiple circuits. The generalized model is established on the base of mathematical recurrences. The winding function theory is used for determining the rotor resistances and the inductances in the case of n- broken bars. Simulation results, in Part. II of this paper, confirm the validity of the proposed model.**

Keywords—Induction Generator; Rotor Broken Bars; Faults Diagnosis; MCSA

I. Introduction

In a general way, the problem of fault diagnosis consists in the determination of the fault type with as much as possible details as the fault size, location and time of detection. From these important details, various approaches using mathematical models are developed for more than three decades, and it has a rich literature of hundreds of papers including many surveys [1, 2].

For the electrical machines, the faults, and their diagnosis techniques are reported in a great number of papers and categorized in three important sections:

Coupled magnetic circuit theory. This theory uses the finite element method associated the state space models, based on the Maxwell's equations by taking account of the machine geometry and the magnetic properties of materials. This method offers considerable advantages, such as the modeling of the topology of the coupled magnetic circuits and the capacity to model magnetic saturation on a high level of precision. However, the finite element method was not the privilege of all the researchers for the simulation of the rotor defects of the induction machines and remains today, little used because of the high computing time and the absence of on line knowledge of the faults [3, 4].

Multiple coupled circuit approach. The approach of the multiple coupled electric circuits is based on the Kirchhoff's laws. If in this method, magnetic saturation is neglected, it is however possible to observe the defects to the rotor and the stator, thanks to the design of a detailed model of an induction machine having m-stator circuits and n-rotor bars, [5-7]. After having drawn up the equivalent scheme under operation normal conditions, it is possible then to write the equations of the meshed network and to simulate the faults in the rotor by the breaks of the rotor bars and/or an end ring segment.

Spectral analysis and parameter estimation. The procedure is based on the spectral analysis of the stator current signature and on the parameter estimation respectively, to identify the rotor bar defects. It is commonly called "Motor Current Signature Analysis.". The stator current spectra shows that the broken rotor bars give rise to a sequence of sidebands around the supply frequency *f* [8]. These sidebands are at $f_m = f(1 \pm 2ks)$, $k=1, 2\ldots$ The lower sideband is specifically due to a broken bar, and the upper sideband is due to consequent speed oscillation. Another signal that has been proposed for motor fault analysis is an air-gap torque. Frequency components are observing at characteristic frequencies $f_m = f(2ks)$. The fault spectra are dependent on the slip, which changes with the load.

The air-gap torque spectrum and the current stator spectrum are a potential signature that can be used by on line fault diagnosis. The analytical knowledge in the process is also used to produce quantifiable, analytical information. The first characteristic values are generated by the process analysis using mathematical process models together with parameter estimation, state estimation and parity equation methods. Characteristic values are parameters, state variables or residuals. As an example: Broken bar detection on an induction machine using state and parameter estimation techniques have been reported in [9, 10].

This paper presents a generalized induction machine model for the diagnosis of rotor defects by an approach of magnetically coupled multiple circuits (using ideas *A.R. Muñoz and T. A. Lip*, [11]). The developed inductances (self and mutual inductances) were given for rotors having an unspecified number of broken rotor bars, by using of mathematical recurrences. Considering the configuration of the rotor, it is more realistic to model the cage as m-identical magnetically coupled circuits. Moreover, one advantage of this approach is that it is applicable for all numbers of broken bars per pole pair. From where the need for determining inductances and resistances to the rotor meshes affected by the defect. Experimental results and simulation in Part. II of this paper confirm the validity of the proposed approach.

II. MODELING OF SCIG BY A MULTIPLE COUPLED CIRCUIT APPROACH

The three-phase model does not make it possible to know the actual values of the currents circulating in the rotor bars and thus does not allow determining the defects with the rotor. This is why the model multi-windings were developing. One will represent then, a model of the rotor as being made up with as many phases as bars; what results in regarding the currents circulating in the end ring segments as rotor phase currents [12, 13].

By taking account of the following assumptions:

- ✓ negligible saturation,
- ✓ sinusoidal distribution of the EMF,
- ✓ uniform air-gap,
- ✓ negligible inter-bar currents

A. Stator equations

The stator equations are written by:

$$v_s = R_s . i_s + \frac{d\phi_s}{dt} \tag{1}$$

With: $v_s = \begin{bmatrix} v_a & v_b & v_c \end{bmatrix}^T$, $i_s = \begin{bmatrix} i_a & i_b & i_c \end{bmatrix}^T$

The resistance matrix (diagonal (3x3)) contains resistance of each winding:

$$R_s = \begin{bmatrix} r_s & 0 & 0 \\ 0 & r_s & 0 \\ 0 & 0 & r_s \end{bmatrix}$$

And the total stator flux is:

$$\phi_s = \phi_{ss} + \phi_{sr} \tag{2}$$

$$\phi_s = L_s i_s + L_{sr} . i_r \tag{3}$$

Where: ϕ_{ss} is stator flux due to the stator currents and ϕ_{sr} is stator flux due to the rotor currents.

With:

$$L_s = \begin{bmatrix} L_{ls} + L_{ms} & -\frac{L_{ms}}{2} & -\frac{L_{ms}}{2} \\ -\frac{L_{ms}}{2} & L_{ls} + L_{ms} & -\frac{L_{ms}}{2} \\ -\frac{L_{ms}}{2} & -\frac{L_{ms}}{2} & L_{ls} + L_{ms} \end{bmatrix}$$

Where L_{ls} and L_{ms} are respectively leakage and magnetizing inductances of stator windings. Magnetizing inductance L_{ms} for a winding having N_s number of stator phase turns in series, is given by:

$$L_{ms} = \frac{\mu_0 lr}{g} N_s^2 \left(\frac{\pi}{4} \right) \tag{4}$$

Leakage inductance is computed by [11].

Each rotor mesh is made of two adjacent bars and the end- ring segment, which connect them and is magnetically coupled with all the other rotor meshes and the three-phase stator. Mutual inductance matrix (3xN_r) stator-rotor is:

$$L_{sr} = \begin{bmatrix} L_{a1} & L_{a2} & \cdots & L_{aN_r} \\ L_{b1} & L_{b2} & \cdots & L_{bN_r} \\ L_{c1} & L_{c2} & \cdots & L_{cN_r} \end{bmatrix}$$

The inductances are computed by using winding function theory “WFT”. This method supposes that there is no symmetry according to the winding function theory [6]. Mutual inductance between two unspecified reels "*i*" and "*j*" of the machine can be calculated from Eq.(5), by supposing that iron permeance is infinite [14].

$$L_{ij}(\theta) = \mu_0 lr \int_0^{2\pi} g^{-1}(\varphi,\theta) N_i(\varphi,\theta) N_j(\varphi,\theta) d\varphi \tag{5}$$

Where θ represents the rotor position compared to a given reference (related to the stator). φ is a particular angular position along the interior surface of the stator. $g^{-1}(\varphi, \theta)$ is the inverse function of the air-gap, if we then suppose that the air-gap is constant and small compared to rotor ray, the function is equal to $1/g$. The term $N_i(\varphi,\theta)$ is known as winding function and represent the spatial distribution of the MMF along the air-gap for a unit current circulating in winding "*i*". Leakage inductances are calculated by only setting $i=j$. From winding function theory, one can determine the MMF of air-gap produced by a current i_a circulating in winding "*a*" for that it position of the air-gap, as follows:

$$F_a = N_a(\theta) . i_a \tag{6}$$

The created flux in a reel, having N_b whorls, by the current i_a is given by:

$$\phi = F . \lambda \tag{7}$$

Where λ and F are respectively the permeance and air-gap magneto-motive force. The differential flux across the air-gap from rotor to stator through the cross section *(r.dθ.l)* is:

$$d\phi = F_a\,(\theta) . \mu_0 . l . r . d\theta / g \tag{8}$$

The differential flux created in the reel *B* is:

$$d\Lambda_{ba} = N_a\,(\theta) . N_b\,(\theta) i_a\,(\theta) . \mu_0\,. l . r . d\theta / g \tag{9}$$

From where the total fluxes expression:

$$\Lambda_{ba} = \mu_0 . l . r . i_a . g^{-1} \int_{\theta_{b1}}^{\theta_{b2}} N_a(\theta) . N_b(\theta) . d\theta \tag{10}$$

From total flux, it is then possible to determine inductance between reel *a* and reel *b*:

$$L_{ba} = \frac{\mu_0 .l.r}{g} \int_{\theta_{b1}}^{\theta_{b2}} N_a(\theta).N_b(\theta).d\theta \qquad (11)$$

By considering the distribution of the stator windings and the rotor bars, the winding function for the stator phases is:

$$N_a = \frac{N_s}{2} cos(\theta)$$

$$N_b = \frac{N_s}{2} cos\left(\theta - \frac{2\pi}{3}\right)$$

$$N_c = \frac{N_s}{2} cos\left(\theta + \frac{2\pi}{3}\right)$$

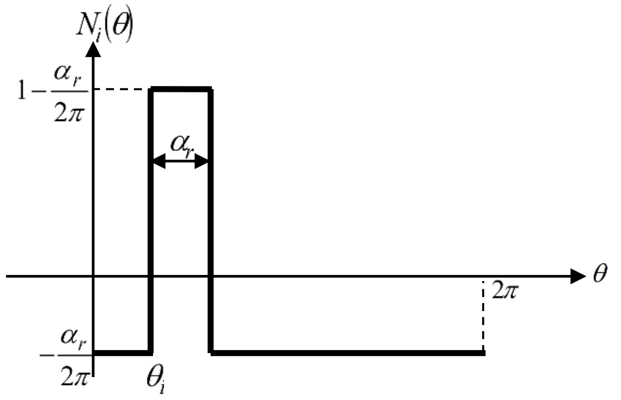


Fig. 1. Rotor loop winding function of healthy generator

The winding function for the i^{th} mesh of the rotor is:

$$N_i = \begin{cases} -\alpha_r / 2\pi & 0 < \theta \le \theta_i \\ 1 - \alpha_r / 2\pi & \theta_i < \theta \le \theta_{i+1} \\ -\alpha_r / 2\pi & \theta_{i+1} < \theta \le 2\pi \end{cases}$$

Where α_r is the angle between two adjacent bars and θ_i the angle of the bar i^{th}. By replacing $g^{-1}(\varphi,\theta)$ and $N(\varphi,\theta)$ previously defined in the expression (7), for the phase "*a*", one will determine following inductances:

$$L_{ai} = \frac{\mu_0 lr}{g} \int_0^{2\pi} N_a(\theta) N_i(\theta) d\theta =$$

$$\frac{\mu_0 lrN_s}{2g} \left(-\int_0^{\theta_i} \frac{\alpha_r}{2\pi} cos\,\theta d\theta + \int_{\theta_i}^{\theta_{i+1}} \left(1 - \frac{\alpha_r}{2\pi}\right) cos\,\theta d\theta - \int_{\theta_{i+1}}^{2\pi} \frac{\alpha_r}{2\pi} cos\,\theta d\theta \right) \qquad (12)$$

After integration this expression yields

$$L_{ai} = \frac{\mu_0 lr}{g} \frac{N_s}{2} (sin\theta_{i+1} - sin\theta_i) \qquad (13)$$

θ_i can be set in the following form: $\theta_i=\theta_r+(i-1)\alpha_r$, with θ_r is an arbitrary angle.

so :

$$L_{ai} = \frac{\mu_0 lr}{g} \frac{N_s}{2} \left[sin(\theta_r + i\alpha_r) - sin(\theta_r + (i-1)\alpha_r)\right] \qquad (14)$$

After transformation by trigonometric relations:

$$L_{ai} = \frac{\mu_0 lrN_s}{g} sin\frac{\alpha_r}{2} cos\left(\theta_r + \left(i - \frac{1}{2}\right)\alpha_r\right) \qquad (15)$$

Let:

$$L_m = \frac{\mu_0 lrN_s}{g} \qquad \text{with : } \delta = \frac{\alpha_r}{2}$$

We obtain:

$$L_{ai} = L_m\, sin\delta\, cos(\theta_i + \delta) \qquad (16)$$

It is then easy to find mutual inductances for two other phases.

B. Rotor equations

For needs for simple comprehension, the rotor cage is modeled by an equivalent circuit containing *k+1* magnetically meshes. The rotor bars are numbered from 1 to *k+1*.

It is to be considered that each mesh is defined by two adjacent bars of the rotor and connected between them by end-ring segment. Moreover, each rotor bar and end ring segment are replaced by an equivalent circuit represented by a resistance and an inductance, as the Fig.2 shows it.

From these equivalent circuits of a squirrel cage rotor, one can write the following equations:

$$v_r = R_r .i_r + \frac{d\phi_r}{dt} \qquad (17)$$

With :

$$\phi_r = L_{sr}^T i_s + L_r .i_r \qquad (18)$$

$$v_r = \begin{bmatrix} v_{r1} & v_{r2} & \dots & v_{rk} & \dots & v_{rN_r} & v_e \end{bmatrix}^T = 0$$

$$i_r = \begin{bmatrix} i_{r1} & i_{r2} & \dots & i_{rk} & \dots & i_{rN_r} & i_e \end{bmatrix}^T$$

$$\phi_r = \begin{bmatrix} \phi_{r1} & \phi_{r2} & \dots & \phi_{rk} & \dots & \phi_{rN_r} & \phi_e \end{bmatrix}^T$$

The rotor resistance matrix (N_rxN_r) is:

$$R_r = \begin{bmatrix} R_0 & -R_b & 0 & \dots & 0 & -R_b \\ -R_b & R_0 & -R_b & \dots & . & 0 \\ 0 & -R_b & R_0 & \dots & . & . \\ .. & & & & & \\ . & . & . & \dots & . & . \\ . & . & . & \dots & R_0 & -R_b \\ -R_b & . & . & \dots & -R_b & R_0 \end{bmatrix}$$

With :

$$R_0 = 2(R_b + R_e)$$

$$L_0 = L_{mr} + 2(L_b + L_e)$$

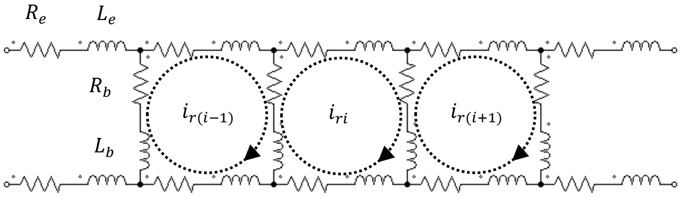


Fig. 2. Equivalent circuits of a squirrel cage rotor for Healthy induction generator

Self-inductance of the meshes is equal, and they are given by:

$$L_{mr} = \frac{\mu_0 lr}{g}\int_0^{2\pi} N_i^2(\theta)d\theta =$$

$$\frac{\mu_0 lr}{g}\left(\int_0^{\theta_i}\left(-\frac{\alpha_r}{2\pi}\right)^2 d\theta + \int_{\theta_i}^{\theta_{i+1}}\left(1-\frac{\alpha_r}{2\pi}\right)^2 d\theta + \int_{\theta_{i+1}}^{2\pi}\left(-\frac{\alpha_r}{2\pi}\right)^2 d\theta\right)$$

$$= \frac{\mu_0 lr}{g}\alpha_r\left(1-\frac{\alpha_r}{2\pi}\right) \tag{19}$$

Rotor-rotor mutual inductance between two rotor meshes is deduced from Eq.(19)

$$L_{ki} = \frac{\mu_0 lr}{g}\int_0^{2\pi} N_k(\theta)N_i(\theta)d\theta = -\frac{\mu_0 lr}{g}\left(\frac{\alpha_r^2}{2\pi}\right) \tag{20}$$

Electromagnetic torque produced by the machine can be expressed from the magnetic co-energy W_{co} as,

$$\Gamma_e = \left[\frac{\partial W_{co}}{\partial \theta}\right]_{(i_s, i_r constant)}$$

It is previously defined that inductances L_s and L_r contain only constant elements and the electromagnetic torque is a scalar quantity. The final expression of the torque is thus reduced to

$$\Gamma_e = \frac{1}{2}P i_s^T \frac{\partial L_{sr}(\theta_r)}{\partial \theta_r} i_r$$

Where P is the number of pole pairs and θ_r is the rotor displacement in electrical radians.

$$L_r = \begin{bmatrix} L_0 & L_{12}-L_b & L_{13} & \cdots & L_{1(N_r-1)} & L_{1N_r}-L_b \\ L_{21}-L_b & L_0 & L_{23}-L_b & \cdots & . & L_{2N_r} \\ L_{31} & L_{32}-L_b & . & \cdots & . & . \\ \vdots & . & . & \cdots & . & . \\ \vdots & . & . & \cdots & . & L_{(N_r-1)N_r}-L_b \\ L_{N_r 1}-L_b & . & . & \cdots & L_{N_r(N_r-1)}-L_b & L_0 \end{bmatrix}$$

$$\Gamma_e = \frac{P}{2}L_m\left\{\left(i_a - \frac{1}{2}i_b - \frac{1}{2}i_c\right)\sum_{k=1}^{N_r} i_{rk}\sin(\theta_r + (k-1)\alpha_r) + \frac{\sqrt{3}}{2}(i_c - i_b)\sum_{k=1}^{N_r} i_{rk}\cos(\theta_r + (k-1)\alpha_r)\right\} \tag{21}$$

C. Broken rotor bar faults

1) Generator with one broken bar

The rupture of a bar decreases the number of equations to the rotor of 1, because the defect imposes the following condition:

If it is the bar traversed by i_{ri} and $i_{r(i+1)}$ which is broken, one has $i_{ri} = i_{r(i+1)}$ which wants to say that current i_{ri} traverses a mesh twice broader and the mesh $i+1$ is eliminated, Fig.(3).

In the inductance matrix the line and the column $i+1$ are eliminated and the terms relating to the column i are thus recomputed by using the expression (5) by taking account of the new function of winding for the rotor mesh i,

$$N_i = \begin{cases} -\alpha_r/2\pi & 0 < \theta \le \theta_i \\ 1-\alpha_r/2\pi & \theta_i < \theta \le \theta_{i+2} \\ -\alpha_r/2\pi & \theta_{i+2} < \theta \le 2\pi \end{cases}$$

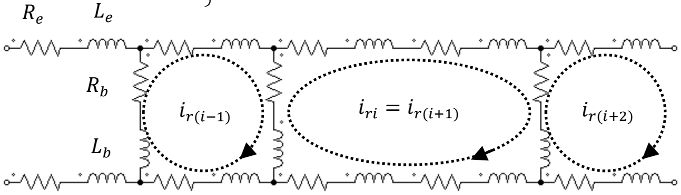


Fig. 3. Equivalent circuits of a squirrel cage rotor with one broken bar

Calculation of new mutual inductance L_{ai} between the mesh i and stator winding a:

$$L_{ai} = \frac{\mu_0 lr}{g}\int_0^{2\pi} N_a(\theta)N_i(\theta)d\theta = L_m \sin(2\delta)\cos(\theta_i + 2\delta) \tag{22}$$

The mutual inductance matrix stator rotor (3x(N_r-1)) becomes:

$$[L_{sr}] = \begin{bmatrix} L_{a1} & L_{a2} & \cdots & L_{ai} & L_{a(i+2)} & \cdots & L_{aN_r} \\ L_{b1} & L_{b2} & \cdots & L_{bi} & L_{b(i+2)} & \cdots & L_{bN_r} \\ L_{c1} & L_{c2} & \cdots & L_{ci} & L_{c(i+2)} & \cdots & L_{cN_r} \end{bmatrix}$$

The same calculation is applied for the i^{th} line, the new matrix inductance of rotor is $(N_r\text{-}1)$ x $(N_r\text{-}1)$ becomes:

$$L_r = \begin{bmatrix} L_0 & L_{12}-L_b & L_{13} & \dots & L_{1(N_r-1)} & L_{1N_r}-L_b \\ L_{21}-L_b & L_0 & L_{23}-L_b & \dots & . & L_{2N_r} \\ L_{31} & L_{32}-L_b & . & \dots & . & . \\ \vdots & \vdots & \vdots & L_{0i} & \vdots & \vdots \\ \vdots & . & . & \dots & L_0 & L_{(N_r-1)N_r}-L_b \\ L_{N_r1}-L_b & . & . & \dots & L_{N_r(N_r-1)}-L_b & L_0 \end{bmatrix}$$

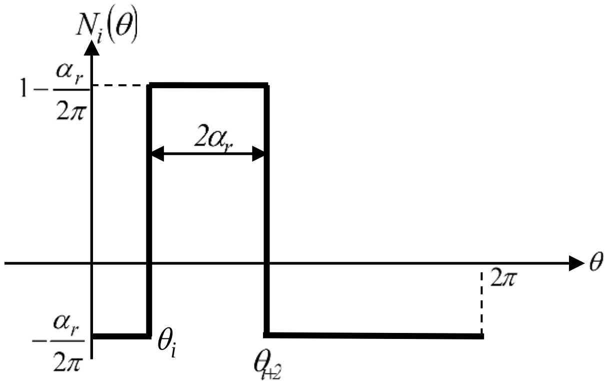


Fig. 4. Rotor loop winding function of faulty generator (1bb)

The rotor resistance matrix is affected by the break of the bar, and is obtained from the new representation $(N_r\text{-}1)$ x$(N_r\text{-}1)$ of the rotor:

$$R_r = \begin{bmatrix} R_0 & -R_b & \cdots & 0 & \cdots & -R_b \\ -R_b & R_0 & \cdots & \dots & . & 0 \\ \vdots & \vdots & \cdots & \vdots & . & . \\ . & . & . & R_{0i} & . & . \\ . & . & . & \vdots & \vdots & \vdots \\ -R_b & . & . & \dots & \cdots & R_0 \end{bmatrix}$$

With:

$$R_{0i} = 2(R_b + 2R_e)$$

Determination of the new self-inductance L_{0i} of the rotor:

$$L_{0i} = L_{ii} + 2(L_b + 2L_e)$$

The self-inductance of the i^{th} rotor mesh becomes:

$$L_{ii} = \frac{\mu_0 lr}{g}\int_0^{2\pi} N_i^2(\theta)d\theta = \frac{\mu_0 lr}{g}\alpha_r\left(2-3\frac{\alpha_r}{2\pi}\right) \qquad (23)$$

The mutual inductance between the k^{th} mesh and the i^{th} mesh of the rotor is recalculated as follows:

$$L_{ki} = \frac{\mu_0 lr}{g}\int_0^{2\pi} N_k(\theta)N_i(\theta)d\theta = -2\frac{\mu_0 lr}{g}\left(\frac{\alpha_r^2}{2\pi}\right) \qquad (24)$$

2) Generator with two adjacent broken bars

In this case, the number of rotor meshes falls by 2, if the first bar traversed by the currents i_{ri}, $i_{r(i+1)}$ and second bar traversed by the currents $i_{r(i+1)}$ and $i_{r(i+2)}$ are broken, one have $i_{ri}=i_{r(i+1)}=i_{r(i+2)}$ means that i_{ri} traverses a mesh three times broader as shows it in Fig.(5). At the time of the break of two bars the function of winding becomes:

$$N_i = \begin{cases} -\alpha_r/2\pi & 0 < \theta \le \theta_i \\ 1-\alpha_r/2\pi & \theta_i < \theta \le \theta_{i+3} \\ -\alpha_r/2\pi & \theta_{i+3} < \theta \le 2\pi \end{cases}$$

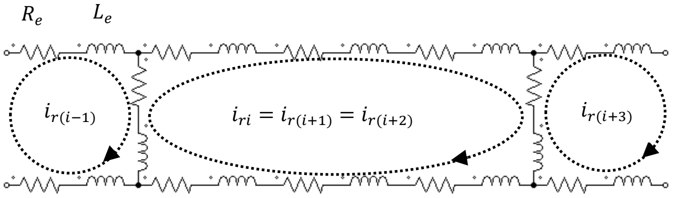


Fig. 5. Equivalent circuits of a squirrel cage rotor with two broken bars

The same type of relation applied in the case of a broken bar is used for the calculation of the new inductance and resistance matrices.

The mutual inductance matrix stator rotor (3x$(N_r\text{-}2)$) becomes:

$$[L_{sr}] = \begin{bmatrix} L_{a1} & L_{a2} & \dots & L_{ai} & L_{a(i+3)} & \dots & L_{aN_r} \\ L_{b1} & L_{b2} & \dots & L_{bi} & L_{b(i+3)} & \dots & L_{bN_r} \\ L_{c1} & L_{c2} & \dots & L_{ci} & L_{c(i+3)} & \dots & L_{cN_r} \end{bmatrix}$$

The same calculation is applied for the i^{th} line, the new matrix inductance of rotor is $(N_r\text{-}2)$ x$(N_r\text{-}2)$ dimensions, with:

$$L_{0i} = L_{ii} + 2(L_b + 3L_e)$$

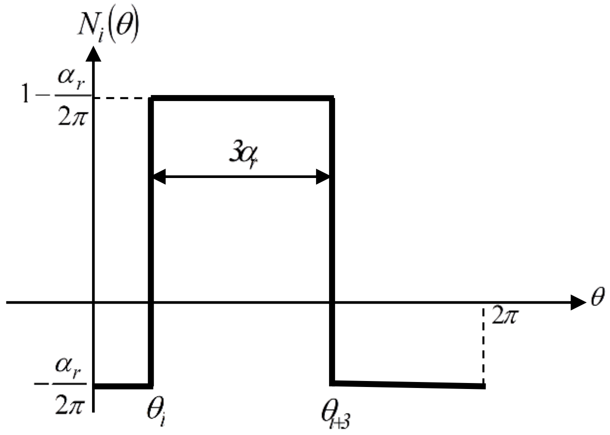


Fig. 6. Rotor loop winding function of faulty generator (2 adjacent bb)

And the self-inductance of winding *i* is

$$L_{ii} = \frac{\mu_0 lr}{g} \int_0^{2\pi} N_i^2(\theta) d\theta = \frac{\mu_0 lr}{g} \alpha_r \left(3 - 5\frac{\alpha_r}{2\pi} \right) \qquad (25)$$

The mutual inductance between the k^{th} mesh and the i^{th} mesh of the rotor is recalculated as follows:

$$L_{ki} = \frac{\mu_0 lr}{g} \int_0^{2\pi} N_k(\theta) N_i(\theta) d\theta = -3\frac{\mu_0 lr}{g} \left(\frac{\alpha_r^2}{2\pi} \right) \qquad (26)$$

Determination of mutual inductances

$$L_{ai} = \frac{\mu_0 lr}{g} \int_0^{2\pi} N_a(\theta) N_i(\theta) d\theta = L_m \sin(3\delta) \cos(\theta_i + 3\delta) \quad (27)$$

The rotor resistance matrix is affected by the break of the bar, and is obtained from the new representation (N_r-2) x(N_r-2) of the rotor, with :

$$R_{0i} = 2(R_b + 3R_e)$$

3) Generalized model by mathematical recurrences

The same reasoning, that in the case of a bar and two broken bars, is valid in the case where there are several broken successive adjacent bars. By mathematical recurrences, we determine the generalized model. The number of rotor equations will be fallen according to the number of broken bars, and the meshes concerned with the break are eliminated; the mesh *i* will be *n* time broader.

The mutual inductance matrix stator rotor (3x(N_r-*n*)) becomes:

$$[L_{sr}] = \begin{bmatrix} L_{a1} & L_{a2} & \cdots & L_{ai} & L_{a(i+(n+1))} & \cdots & L_{aN_r} \\ L_{b1} & L_{b2} & \cdots & L_{bi} & L_{b(i+(n+1))} & \cdots & L_{bN_r} \\ L_{c1} & L_{c2} & \cdots & L_{ci} & L_{c(i+(n+1))} & \cdots & L_{cN_r} \end{bmatrix}$$

The same calculation is applied for the i^{th} line, the new matrix inductance of rotor is (N_r-*n*) x(N_r-*n*) dimensions, with:

$$L_{0i} = L_{ii} + 2(L_b + (n+1)L_e)$$

And the self-inductance of winding *I* is

$$L_{ii} = \frac{\mu_0 lr}{g} \int_0^{2\pi} N_i^2(\theta) d\theta = \frac{\mu_0 lr}{g} \alpha_r \left((n+1) - (2n+1)\frac{\alpha_r}{2\pi} \right) (28)$$

The mutual inductance between the k^{th} mesh and the i^{th} mesh of the rotor is recalculated as follows:

$$L_{ki} = \frac{\mu_0 lr}{g} \int_0^{2\pi} N_k(\theta) N_i(\theta) d\theta = -(n+1)\frac{\mu_0 lr}{g} \left(\frac{\alpha_r^2}{2\pi} \right) \qquad (29)$$

Determination of mutual inductances

$$L_{ai} = \frac{\mu_0 lr}{g} \int_0^{2\pi} N_a(\theta) N_i(\theta) d\theta = L_m \sin((n+1)\delta) \cos(\theta_i + (n+1)\delta)$$

(30)

The rotor resistance matrix becomes (N_r-*n*) x(N_r-*n*) with:

$$R_{0i} = 2(R_b + (n+1)R_e)$$

4) Generator with two non-adjacent broken bars

Now we suppose that the two broken bars are not adjacent, it is obvious that the matrices resistances and inductances of the previous section are not the same. We will see that there is a new mutual inductance between the two cells and involved two broken non adjacent bars as shown in the following figure:

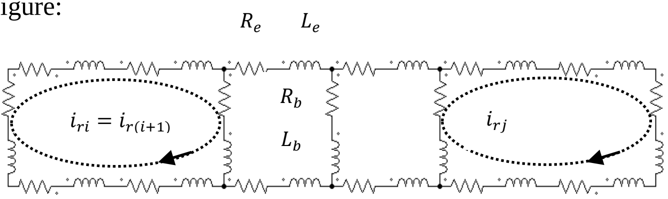


Fig. 7. Equivalent circuits of a squirrel cage rotor with two non-adjacent broken bars

This failure reduces the number of rotor equations with two equations, but one will have:

$$i_{ri} = i_{r(i+1)}$$
$$i_{rj} = i_{r(j+1)}$$

The mutual inductances (stator-rotor, rotor-rotor) and the self-inductance of the mesh ir remains the same as we saw in the first part. With the same procedure we can calculate the inductances of the mesh jr , such as:

$$N_j = \begin{cases} -\alpha_r / 2\pi & 0 < \theta \le \theta_j \\ 1 - \alpha_r / 2\pi & \theta_j < \theta \le \theta_{j+2} \\ -\alpha_r / 2\pi & \theta_{j+2} < \theta \le 2\pi \end{cases}$$

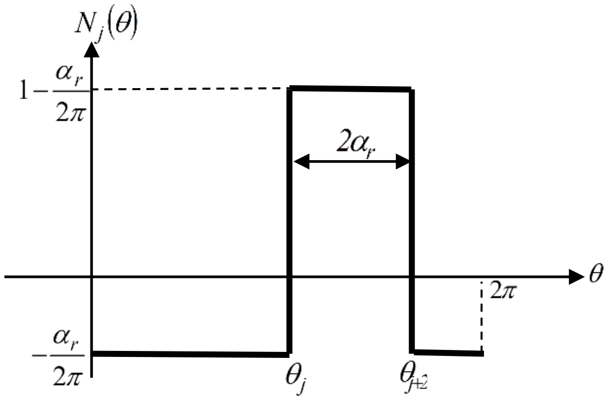


Fig. 8. Rotor loop winding function of faulty generator (2 non-adjacent bb)

Calculation of stator-rotor mutual inductances of the two mesh i_r and j_r :

$$L_{ai} = \frac{\mu_0 lr}{g} \int_0^{2\pi} N_a(\theta) N_i(\theta) d\theta = \frac{\mu_0 lr}{g} N_s \sin(2\delta) \cos(\theta_i + 2\delta)$$

$$L_{aj} = \frac{\mu_0 lr}{g} \int_0^{2\pi} N_a(\theta) N_j(\theta) d\theta = \frac{\mu_0 lr}{g} N_s \sin(2\delta) \cos(\theta_j + 2\delta)$$

The matrix of the stator-rotor mutual inductance becomes:

$$[L_{sr}]=\begin{bmatrix} L_{a1} & L_{a2} & \dots & L_{ai} & L_{a(i+2)} & \dots & L_{aj} & L_{a(j+2)} & \dots & L_{aN_r} \\ L_{b1} & L_{b2} & \dots & L_{bi} & L_{b(i+2)} & \dots & L_{bj} & L_{b(j+2)} & \dots & L_{bN_r} \\ L_{c1} & L_{c2} & \dots & L_{ci} & L_{c(i+2)} & \dots & L_{cj} & L_{c(j+2)} & \dots & L_{cN_r} \end{bmatrix}$$

Calculation of self-inductances of i^{th} et j^{th} rotor meshes :

$$L_{ii}=\frac{\mu_0 lr}{g}\int_0^{2\pi} N_i^2(\theta)d\theta=\frac{\mu_0 lr}{g}\alpha_r\left(2-3\frac{\alpha_r}{2\pi}\right)$$

$$L_{jj}=\frac{\mu_0 lr}{g}\int_0^{2\pi} N_j^2(\theta)d\theta=\frac{\mu_0 lr}{g}\alpha_r\left(2-3\frac{\alpha_r}{2\pi}\right)$$

Calculation of mutual inductances between k^{th} and i^{th} meshes firstly, on the other hand between k^{th} and j^{th} meshes of rotor :

$$L_{ki}=\frac{\mu_0 lr}{g}\int_0^{2\pi} N_k(\theta)N_i(\theta)d\theta=-2\frac{\mu_0 lr}{g}\left(\frac{\alpha_r^2}{2\pi}\right)$$

$$L_{kj}=\frac{\mu_0 lr}{g}\int_0^{2\pi} N_k(\theta)N_j(\theta)d\theta=-2\frac{\mu_0 lr}{g}\left(\frac{\alpha_r^2}{2\pi}\right)$$

Calculation of mutual inductance between i^{th} and j^{th} meshes:

$$L_{ij}=\frac{\mu_0 lr}{g}\int_0^{2\pi} N_i(\theta)N_j(\theta)d\theta=-3\frac{\mu_0 lr}{g}\left(\frac{\alpha_r^2}{2\pi}\right)$$

The matrix of the rotor inductances becomes:

$$[L_r]=\begin{bmatrix} L_0 & L_{12}-L_b & \dots & L_{1i} & L_{1(i+2)} & \dots & L_{1j} & L_{1(j+2)} & \dots & L_{1N_r}-L_b \\ L_{21}-L_b & L_0 & \dots & . & . & \dots & . & . & \dots & . \\ . & . & \dots & . & . & \dots & . & . & \dots & . \\ . & . & \dots & . & . & \dots & . & . & \dots & . \\ L_{i1} & . & \dots & L_{0i} & L_{i(i+2)}-L_b & \dots & L_{ij} & L_{i(j+2)} & \dots & L_{iN_r} \\ L_{(i+2)1} & . & \dots & L_{(i+2)i}-L_b & L_0 & \dots & L_{(i+2)j} & L_{(i+2)(j+2)} & \dots & L_{(i+2)N_r} \\ . & . & \dots & . & . & \dots & . & . & \dots & . \\ . & . & \dots & . & . & \dots & . & . & \dots & . \\ L_{j1} & . & \dots & L_{ji} & L_{j(i+2)} & \dots & L_{0j} & L_{j(j+2)}-L_b & \dots & L_{jN_r} \\ L_{(j+2)1} & . & \dots & L_{(j+2)i} & .L_{(j+2)(i+2)} & \dots & L_{(j+2)j}-L_b & L_0 & \dots & L_{(j+2)N_r} \\ . & . & \dots & . & . & \dots & . & . & \dots & . \\ . & . & \dots & . & . & \dots & . & . & \dots & . \\ L_{N_r1}-L_b & . & \dots & L_{N_ri} & L_{N_r(i+2)} & \dots & L_{N_rj} & L_{N_r(j+2)} & \dots & L_0 \end{bmatrix}$$

The matrix of the rotor resistances is written:

$$[R_r]=\begin{bmatrix} R_0 & -R_b & \dots & 0 & 0 & \dots & 0 & 0 & \dots & -R_b \\ -R_b & R_0 & \dots & . & . & \dots & . & . & \dots & . \\ . & . & \dots & . & . & \dots & . & . & \dots & . \\ . & . & \dots & R_{0i} & -R_b & \dots & . & . & \dots & . \\ . & . & \dots & -R_b & R_0 & \dots & . & . & \dots & . \\ . & . & \dots & . & . & \dots & . & . & \dots & . \\ . & . & \dots & . & . & \dots & . & . & \dots & . \\ 0 & . & \dots & . & . & \dots & R_{0j} & -R_b & \dots & 0 \\ 0 & . & \dots & . & . & \dots & -R_b & R_0 & \dots & 0 \\ . & . & \dots & . & . & \dots & . & . & \dots & -R_b \\ -R_b & 0 & \dots & . & . & \dots & 0 & 0 & \dots & R_0 \end{bmatrix}$$

Such as:

$$R_{0i}=2(R_b+2R_e)$$

$$L_{0i}=L_{ii}+2(L_b+2L_e)$$

$$R_{0j}=2(R_b+2R_e)$$

$$L_{0j}=L_{jj}+2(L_b+2L_e)$$

There is a difference between the two models of equations of a cage rotor with two broken bars, in the first model the two bars are adjacent and in the second are not adjacent. This difference appears particularly at resistance, self and mutual inductances of the relevant mesh breaking bars and the mutual inductance between them and the stator windings.

III. Parameter Determination for Rotor with Several Broken Bars

The suggested approach was tested on a 4 kW, 2 pole pairs, 28 bars squirrel cage induction generator which detailed parameters are given in Appendix.

The defects are introduced into the program by removing the blocks defining the bars concerned with the break, and by introducing new inductances and resistances calculated in the previous section. The variations of the equivalent resistance and the inductances of the rotor are given in table1.

TABLE I. Parameter Values for Various Broken Rotor Bars

Number of BRB				
	L_{ii}(μH)			

Number of BRB	$L_{ii}(\mu H)$	$L_{0i}(\mu H)$	$L_{ki}(\mu H)$	$R_{0i}(\mu\Omega)$
0	8.15	8.78	-0.3	203.88
1	16	16.7	-0.6	213.88
2	23.84	24.61	-0.9	223.88
3	31.68	32.52	-1.2	233.88
4	39.53	40.45	-1.5	243.88
5	47.38	48.37	-1.8	253.88

Number of BRB	$L_{ii}(\mu H)$	$L_{0i}(\mu H)$	$L_{ki}(\mu H)$	$R_{0i}(\mu\Omega)$
6	55.22	56.28	-2.1	263.88
7	63.07	64.20	-2.4	273.88
8	70.91	72.11	-2.7	283.88
9	78.76	80.04	-3	293.88
10	86.61	87.96	-3.3	303.88

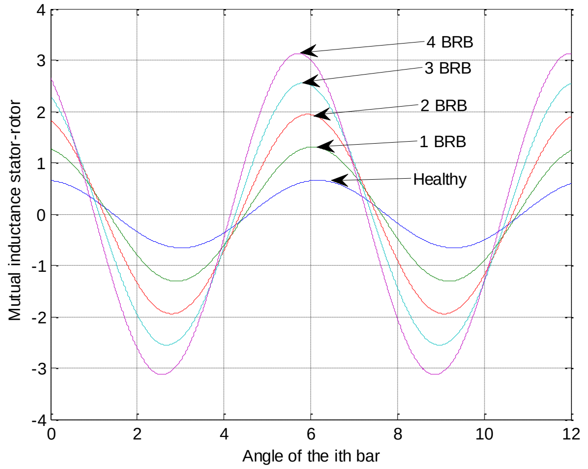


Fig. 9. Mutual inductance between *a* stator phase and i^{th} rotor loop

IV. Conclusion

A generalized model of the squirrel-cage rotor in induction generator has been developed at the base of mathematical recurrences. The determination of leakage and the mutual inductances of the rotor meshes is carried out by using the winding function theory.

This model allows for the introduction of the defects and knowledge of the real currents in the rotor bars like in the end ring; what is difficult to realize by the 3-phase model.

One can determine analytical knowledge of the machine to produce quantifiable and analytical information.

The advantage of this model is that it is applicable to cage rotors with non-integral number of rotor bars per pole pair. The proposed approach can be used also for the simulation of variable speed induction machine drives under rotor faults.

Appendix

The parameters of our machines are illustrated in the following table:

TABLE II. Induction Generator Parameters

Parameters	***Symbols***	**Values**
Rated power	P	4 kW
Rated voltage	V	220/380 V
Rated current	I	15.2/8.8 A
Rated speed	Ω	1465 rpm
Pole-pair number	p	2
Supply frequency	f_s	50 Hz
Power factor	$cos\varphi$	0.83
Rotor inertia	J	0.011 kg.m^2
Friction coefficient	f	8.73e^{-4} Nm/rad/s
Stator phase turns in series	N_s	156
Stator phase resistance	R_s	1.5 Ω
Stator phase leakage inductance	L_{ls}	7 mH
Stator magnetizing inductance	L_{ms}	0.55 H
Rotor bar resistance	R_b	96.94 μΩ
Rotor bar inductance	L_b	0.28 μH
End-ring segment resistance	R_e	5 μΩ
End-ring segment inductance	L_e	0.036 μH
Number of rotor bars	N_r	28
Mean radius air-gap	r	54 mm
Effective air-gap	g	0.28 mm
Length of air-gap	l	120 mm

References

[1] M.E.H Benbouzid. Bibliography on induction motor faults detection and diagnosis. *IEEE Trans.On Energy Conversion*, vol.14 , pp.1065-1074, Dec.1999

[2] Kumar, R.S. ; VIT Univ., Vellore, India ; Manimozhi, M. ; Enosh, M.T. A survey of fault detection and isolation in wind turbine drives. International Conference on Power, Energy and Control (ICPEC), 6-8 Feb. 2013, pp. 648 - 652.

[3] S. Williamson and M. J. Rabinson. Calculation of cage induction motor equivalent circuit parameters using finite elements," *IEE Proceedings-B*, vol. 138, No. 5, pp. 263–276, Sep. 1991.

[4] J. F. Bangura, Povinelli R.J., N.A.O Demerdash, R.H. Brown. Diagnostics of eccentricities and bars/end-ring connectors breakage in polyphase induction motors trough a combination of time series data mining and time-stepping coupled FE state space techniques. IEEE Trans. On Industry Applications, vol.39, N°4, pp.1005-1013, July/August 2003.

[5] Meshgin - Kelk, H. Milimonfared, J. and Toliyat H.A. Interbar Currents and Axial Fluxes in Healthy and Faulty Induction Motors. IEEE Trans. On Industry Applications, vol.40, N°1, pp. 128-134, Jan./Feb 2004.

[6] H.A. Toliyat and T. A. Lipo. Transient analysis of cage induction machines under stator, rotor bar and end ring faults. *IEEE Trans. On Energy Conversion*, vol.10, n°2, pp.241-247, June 1995.

[7] Vinod V. Thomas, Krishna Vasudevan, and V. Jagadeesh Kumar. Online cage rotor fault detection using air-gap torque spectra. IEEE Trans. On Energy Conversion, vol. 18, N°2, pp.265-270, June 2003.

[8] K. R. Cho, J.H. Lang, and S.D. Umans. Detection of broken rotor bars in induction motors using state and parameter estimation. IEEE Trans. On Industry Appl., pp.702-709, vol.28, N°3, May/June 1992.

[9] F. Filippetti, G. Franceshini, C. Tassoni, and P. Vas. Broken bar detection in induction machines: Comparison between current spectrum approach and parameter estimation approach. IEEE-IAS, Annual Meeting Conf., pp.95-102, vol.2, Denver USA,1994.

[10] H. R. Fudeh and Ong C. M. Modeling and analysis of induction machines containing space harmonics. Pt. I, II and III. *IEEE Trans. On Power Apparatus and Systems*, vol.PAS-102, n°8, pp.2608-2628, Aug. 1985

[11] Alfredo R. Muñoz and Thomas Lipo. Complex vector model of the squirrel cage inductance machine including instantaneous rotor bar currents. IEEE Trans. On Industry Applications, vol.35, n°6, pp.1332-1340, 1999.

[12] Omar, T. ; Lahcene, N. ; Rachid, I. ; Maurice, F. Modeling of the induction machine for the diagnosis of rotor defects. Part I. An approach of magnetically coupled multiple circuits. IECON 2005. 31st Annual Conference of IEEE Industrial Electronics Society, 2005.

[13] Issam Attoui , Amar Omeiri. Modeling, control and fault diagnosis of an isolated wind energy conversion system with a self-excited induction generator subject to electrical faults. Energy Conversion and Management 82 (2014) pp. 11–26

[14] Xiaogang Luo, Yuefeng Liao, and H.A. Toliyat. Multiple coupled circuit modeling of induction machines. IEEE Trans. On Industry Applications, vol.31, N°2, pp.311318, March/April 1995.

A Novel Ball on Beam Stabilizing Platform with Inertial Sensors

Part I: Modeling & Simulation with Detailed Geometrical Analysis

Ali Shahbaz Haider
Electrical Engineering Dept
COMSATS Institute of Information Technology
Wah, Pakistan

Basit Safir
Electrical Engineering Dept.
COMSATS Institute of Information Technology
Wah, Pakistan

Muhammad Nasir
Electrical Engineering Dept.
COMSATS Institute of Information Technology
Wah, Pakistan

Farhan Farooq
Electrical Engineering Dept.
COMSATS Institute of Information Technology
Wah, Pakistan

***Abstract*—This research paper presents dynamic modeling of inertial sensor based one degree of freedom (1-DoF) stabilizing platform. Plant is a ball on a pivoted beam. Nonlinear modeling of the plant is done. Ball position on beam is actuated by DC motor using two arms and one beam structure. Arms and beam are linked by pivoted joints. Nonlinear geometrical relations for mechanical structure are derived followed by physically realizable approximations. These relations are used in system dynamic equations followed by linearization, resulting in a linear continuous time differential equation model. State space conversion is done. Final model is simulation and system dynamics are elaborated by analysis of the simulation responses**

Keywords—stabilizing platform; ball on beam; nonlinear dynamics; inertial sensors

I. Introduction

Stabilizing platforms are among challenging control systems. Their applications are immense, especially in defense such as camera stabilizations for drones and automatic gun pointing angle control. Such plate forms have been benchmarks to practice various control techniques. Owing to such significance a lot of research work has been dedicated to these systems. One of such systems is single degree of freedom (1-DoF) ball on beam mechanism. Plant of this control problem consists of a ball capable of rolling on a beam under the action of gravity due to inclination of beam. Control objective is to stabilize the positions of the ball on the beam in the presence of external disturbances and to achieve ball position reference tracking. System is open loop unstable so feedback is inevitable [1].

This research is presented in two parts. Part-I presents modeling of the system. Dynamics of this plant has been derived and discussed in literature. Reduced order linear transfer function based model is used in [1], [2] and [4]. Similar technique has been employed in [3]. Variational techniques have been employed in [5]. However full order model without neglecting actuator dynamics and structural nonlinearities is not studied.

Our work presents geometrically accurate, nonlinear and detailed modeling which has not been presented in literature. Moreover concept of using inertial sensors i.e. rate gyro and accelerometer to measure systems states, is novel and it makes our system much closer to real stabilizing platforms in sea ships and aircraft. In this paper a nonlinear system model is developed followed by linearization and state space conversion. Various systems parameters are identified and their effects on systems dynamics are elucidated. Linear and nonlinear relations are compared and error due to linearization has been analyzed. Open loop linearized dynamics are simulated and discussed.

Organization of the paper is as follows, section-II describes the design and construction of physical hardware. Section-III comprises of derivation of system dynamics, nonlinear geometric relations governing the physical hardware, linearization and state-space conversion. Section-IV presents simulation results followed by section-V describing conclusions and future work.

II. Hardware Design

Hardware platform is shown in Figure 1. Functional description for this plant is described diagrammatically in Figure 2. Plant consists of a beam of length $2b$ hinged at its centre at the pivot point O_2. A ball is placed on this beam. Ball is capable of rolling freely. Its distance D from edge of the beam is to be controlled. Position of the ball changes under the action of gravity if the beam is inclines at some angle θ, which may be positive or negative. Inclination of beam is actuated by a DC gear motor which is connected to the beam by a servo arm and a link arm. DC gear motor is bidirectional and actuates the servo arm at an angle ϕ. Servo arm rotates about point O_1, which is also taken to be the origin of rectangular coordinate system used to model the system. Link arm is pivoted at points P_2 and P_1, these pivot points track a circular trajectory in response to angle ϕ as shown by dashed circular trajectories in the Figure 2.

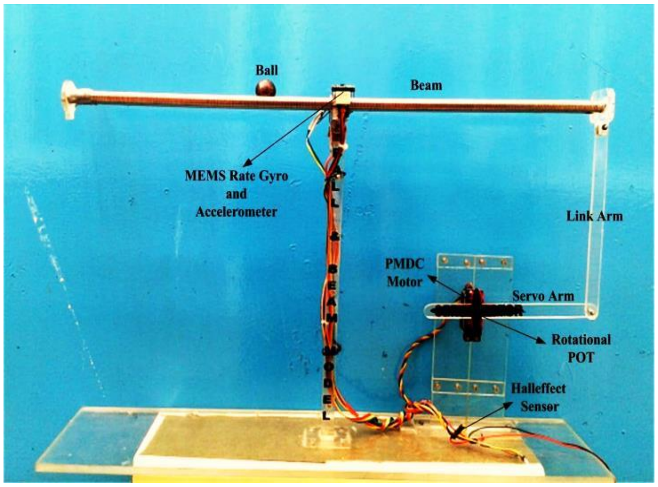


Fig. 1. Hardware platform

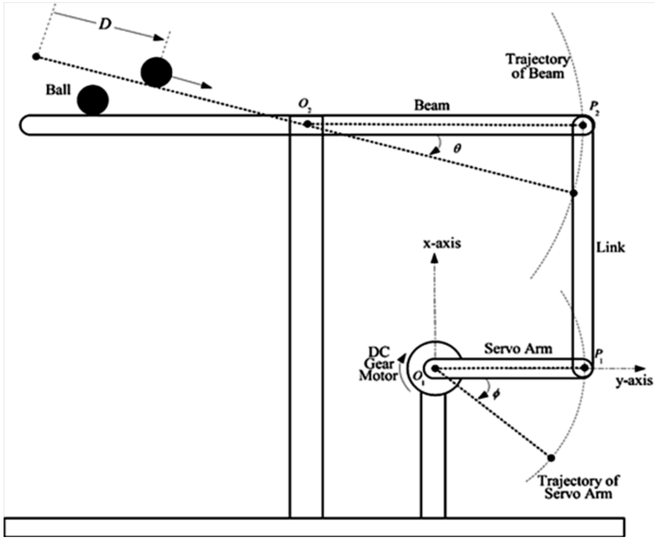


Fig. 2. Functional description of the hardware platform

Control objective is to stabilize the ball at any desired position D or to make it track a commanded position trajectory D(t) by adjusting bounded servo angle ϕ in the presence of disturbances.

III. System Dynamics

It can be seen in Figure 2 that trajectories of the servo arm and the beam are circular so mathematical relations will be nonlinear. In the following sub sections various geometrical relations for the system hardware are developed step by step to be used in the final system dynamic model.

A. DC motor dynamics

The servo arm in Figure 2 is actuated to an angle ϕ by a DC gear motor. Model of a permanent magnet (PM) DC motor is shown in Figure 3.

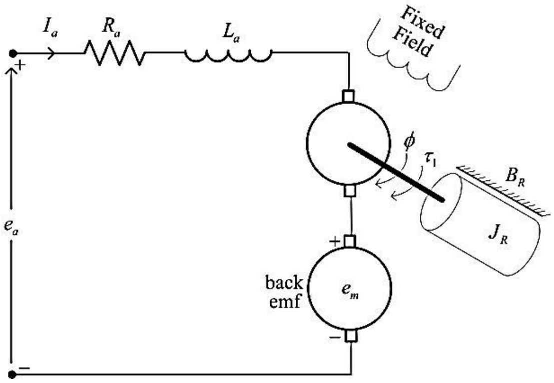


Fig. 3. Electrical and mechanical modeling diagram of PMDC motor

It has two parts namely electrical and mechanical. Armature part of PMDC motor is modelled by a series RL circuit with back emf e_m as shown in Figure 3. Input to the motor is voltage signal e_a .

$$e_a - e_m = R_a i_a + L_a \frac{di_a}{dt} \tag{1}$$

$$e_m = k_m \frac{d\phi}{dt} \tag{2}$$

Substituting equation (2) in equation (1) we get:

$$e_a - k_m \frac{d\phi}{dt} = R_a i_a + L_a \frac{di_a}{dt} \tag{3}$$

Mechanical part of the motor consists of the rotating armature with its end coupled to the servo arm and the link arm. Armature is modelled by a cylindrical moment of inertia J_R experiencing a friction B_R as shown in Figure 3. Since motor output shaft is coupled to servo arm and link arm so net moment of inertia is J_1 given by equation (4),

$$J_1 = J_R + J_S + J_L \tag{4}$$

In Equation (4), J_S and J_L are moment of inertia of the servo arm and the link arm respectively. Using Euler's equations for mechanical rotation part of motor we obtain,

$$J_1 \frac{d^2\phi}{dt^2} = \tau_1 - B_R \frac{d\phi}{dt} \tag{5}$$

Electromechanical coupling equation is given by:

$$\tau_1 = k_\tau i_a \tag{6}$$

B. Expressions for moment of inertia

Armature is modelled by a cylinder of diameter r and mass m_R . The Servo arm, link arm and the beam are modelled by rectangular parallelepiped with dimensions shown in Figure 4. Their moment of inertia are given by following equations,

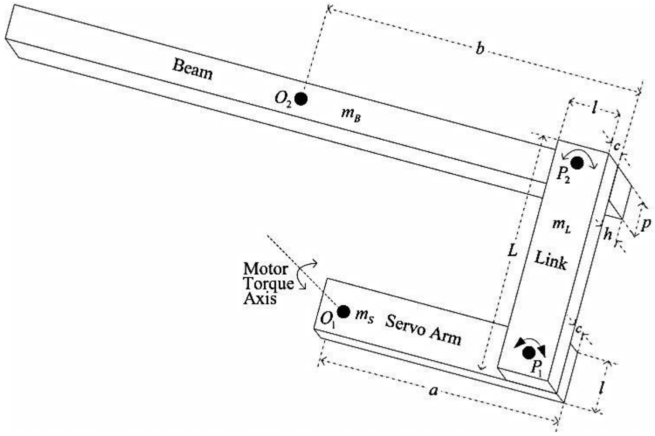


Fig. 4. Moment of inertia for the servo arm, link arm and the beam

$$J_R = \frac{1}{2} m_R r^2 \tag{7}$$

$$J_S = \frac{1}{12} m_S l^2 + \frac{1}{3} m_S a^2 \tag{8}$$

$$J_L = \frac{1}{12} m_L l^2 + \frac{1}{3} m_L L^2 + m_L a^2 \tag{9}$$

$$J_2 = \frac{1}{12}\left(m_B (2b)^2 + m_S p^2 \right) \tag{10}$$

C. Relation Between servo arm angle ϕ and beam angle θ

Position of the ball is actuated by beam angle θ as shown in Figure 2. Beam angle is itself result of servo arm angle ϕ, so a relationship is developed between these two angles. To develop this relation we select a xy-coordinate system with origin $O_1(0,0)$ as shown in Figure 5. If $\left|\overline{O_1P_1}\right| = a$ and $\left|\overline{O_2P_2}\right| = b$ then beam rotates about $O_2(a-b, L)$. Link arm end points become $P_1(\mathrm{a}\cos\phi, a\sin\phi)$ and $P_2(b\cos\theta - b + a, b\sin\theta + L)$.

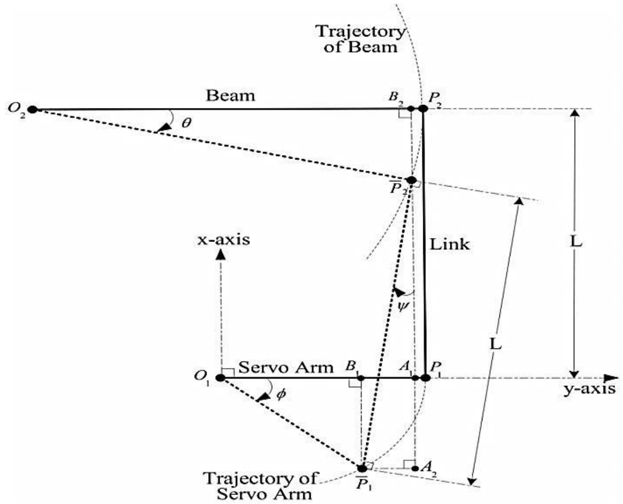


Fig. 5. Geometrical description of the hardware platform

It is clear that $\left|\overline{P_1P_2}\right| = L$ which remain constant irrespective of position of link arm. We may also write it as $\left|\overline{P_1P_2}\right|^2 = L^2$ or $\left|\overrightarrow{O_1P_2} - \overrightarrow{O_1P_1}\right|^2 = L^2$. From Figure 5 we can expand this relation using distance formula as:

$$(b\cos\theta - b + a - \mathrm{a}\cos\phi)^2 + (b\sin\theta + L - a\sin\phi)^2 = L^2,$$

$$\Rightarrow (b\cos\theta - b + a - \mathrm{a}\cos\phi)^2 + (b\sin\theta + L - a\sin\phi)^2 - L^2 = 0 \tag{11}$$

$$\Rightarrow f(a,b,L,\phi,\theta) = 0$$

Equation (11) can easily be verified by checking physically realizable trivial solution of $f(a,b,L,0,\theta)$ or $-b^2\cos\theta + bL\sin\theta$ which comes out to be θ=0 as expected. Equation (11) is tedious to be solved to find an explicit relation between θ and ϕ. However if we put constrain $\theta \in [-0.1\pi, 0.1\pi]$ we may approximate $\sin\theta = \theta$ and $\cos\theta = 1$, so equation (11) results in,

$$\begin{aligned} & f\left(a,b,L,\phi,\theta \in [-0.1\pi, 0.1\pi]\right) = 0 \\ & \Rightarrow [a^2]\cos\phi + [ab(\theta+1)]\sin\phi = a^2 + bL\theta \\ & \Rightarrow \sqrt{\left[a^2\right]^2 + \left[ab(\theta+1)\right]^2} \times \\ & \cos\left\{\phi + \tan^{-1}\left[\frac{ab(\theta+1)}{a^2}\right]\right\} = a^2 + bL\theta \end{aligned} \tag{12}$$

At this point we can get the following explicit relation between θ and ϕ:

$$\phi = \cos^{-1}\left[\frac{1 + \frac{b}{a^2}L\theta}{\sqrt{1 + \left[\frac{b}{a}(\theta+1)\right]^2}}\right] + \tan^{-1}\left[\frac{b}{a}(\theta+1)\right] \tag{13}$$

Putting $b = 2a$ in equation (13) we get,

$$\phi = \cos^{-1}\left[\frac{1 + (2L\theta / a)}{\sqrt{1 + 4(\theta+1)^2}}\right] + \tan^{-1}\left[2(\theta+1)\right] \tag{14}$$

Furthermore if $L = a$ then Equation (14) becomes:

$$\phi = \cos^{-1}\left[\frac{2(\theta+1) - 1}{\sqrt{1 + \left[2(\theta+1)\right]^2}}\right] + \tan^{-1}\left[2(\theta+1)\right] \tag{15}$$

An approximate linear relation can be found between θ and ϕ if we restrain both θ and ϕ in range $[-0.1\pi, 0.1\pi]$. In this case Equation (11) becomes:

$$\begin{aligned} & f\left(a,b,L,\phi \in [-0.1\pi, 0.1\pi], \theta \in [-0.1\pi, 0.1\pi]\right) = \\ & a^2 + [ab(\theta+1)]\phi - a^2 - bL\theta = 0, \\ & \Rightarrow \theta = \frac{a}{L}\phi. \end{aligned} \tag{16}$$

For θ $= [-0.1\pi, 0.1\pi]$, %error between ϕ calculated from nonlinear relation of equation (13) and linear relation of equation (16) have been plotted for various values of L in figure 6. It is evident that linear relation of equation (16) is valid for $\phi \in [-0.1\pi, 0.1\pi]$ and $\theta \in [-0.035\pi, 0.05\pi]$ if $L = a$, with a maximum of 5% error.

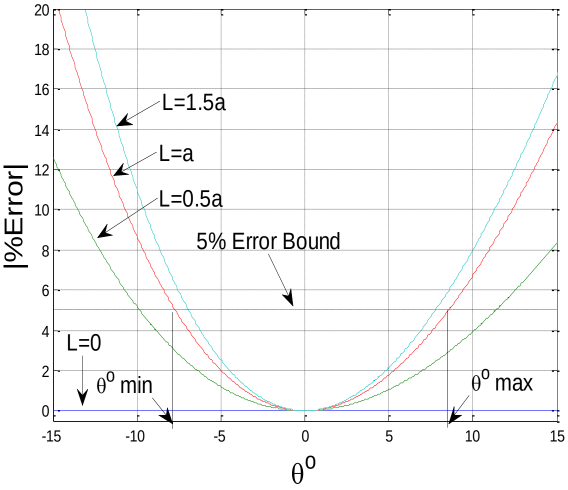


Fig. 6. % error between linear and nonlinear relation for ϕ

D. Torque and force decomposition for the Servo-arm and the link-arm

PMDC motor generates a torque τ_1 and angle ϕ which is applied to servo arm at point O_1 as shown in Figure 7.Since length of the servo arm is 'a' so force F_s normal to arm at point $\overline{P_1}$ is produces. Amplitude of this force is given by $F_s = \tau_1 / a$. However we are interested in finding the component of force F_L which is parallel to the link arm so we decompose F_s along $L_{\parallel}$-axis which comes out to be,

$$F_L = F_S \cos\left(\frac{\pi}{2} - \eta\right) \tag{17}$$

In equation (17) the unknown angle η needs to be expressed in terms of known angle ϕ. Using law of sine in right $\Delta \overline{P_1 A_2 P_2}$ in Figure 5 we get,

$$\psi = \sin^{-1}\frac{a(1-\cos\phi)}{L} \tag{18}$$

In right $\Delta O_1 B_1 \overline{P_1}$ in Figure 5 we have,

$$\eta - \psi = \frac{\pi}{2} - \phi \tag{19}$$

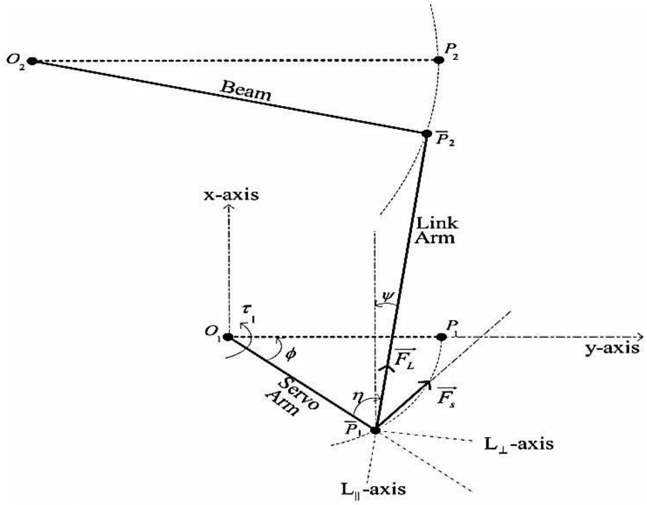


Fig. 7. Electrical and mechanical modeling diagram of PMDC motor

Using Equation (18) in Equation (19) we get:

$$\eta = \frac{\pi}{2} - \phi + \sin^{-1}\frac{a(1-\cos\phi)}{L} \tag{20}$$

Using Equation (20) in Equation (17) we get the desired expression for F_L as:

$$F_L = \frac{\tau_1}{a}\cos\left(\phi - \sin^{-1}\frac{a(1-\cos\phi)}{L}\right) \tag{21}$$

E. Torque and force decomposition for the Link-arm and the beam

The force F_L is transmitted from link arm to the beam. Figure 8 shows that this force is incident on the beam at an angle of λ from beam axis. Component of this force $F_{L\perp}$ perpendicular to beam axis is responsible for producing torque about point O_2. This torque is given by,

$$\tau_2 = bF_{L\perp} = bF_L \sin(\lambda) \tag{22}$$

The unknown angle λ needs to be expressed in terms of input angle ϕ. From Figure 8 we have,

$$\lambda = \frac{\pi}{2} - \psi - \theta \tag{23}$$

Using equations (16)-(20) in equation (23) we get,

$$\lambda = \frac{\pi}{2} - \sin^{-1}\frac{a(1-\cos\phi)}{L} - \frac{a}{L}\phi \tag{24}$$

Substituting equation (24) in (22) we get,

$$\tau_2 = bF_L \sin\left(\frac{\pi}{2} - \sin^{-1}\frac{a(1-\cos\phi)}{L} - \frac{a}{L}\phi\right) \tag{25}$$

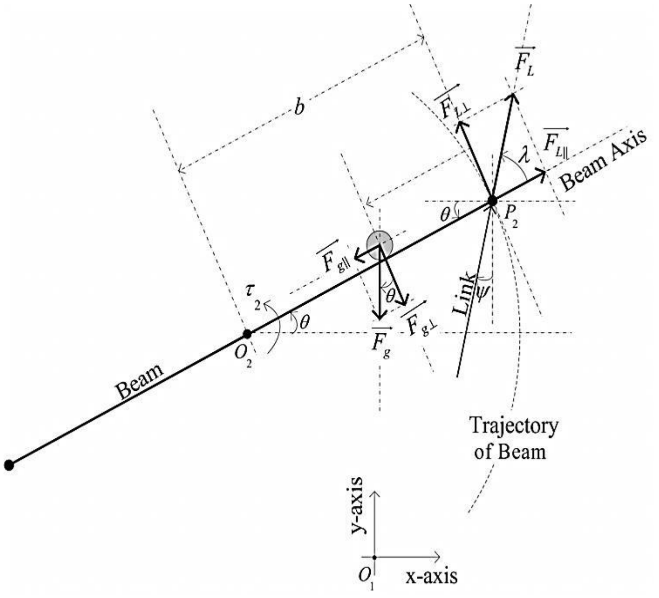

b
$\overrightarrow{F_L}$
$\overrightarrow{F_{L\perp}}$
$\overrightarrow{F_{L\parallel}}$
λ
Beam Axis
θ
P_2
$\overrightarrow{F_{g\parallel}}$
τ_2
θ
θ
Link
ψ
O_2
$\overrightarrow{F_g}$
$\overrightarrow{F_{g\perp}}$
Beam
Trajectory of Beam
y-axis
x-axis
O_1

Fig. 8. Torque and force decomposition for the Link-arm and the beam

Using equation (21) in equation (25) we get,

$$\tau_2 = b\frac{\tau_1}{a}\cos\left(\phi - \sin^{-1}\frac{a(1-\cos\phi)}{L}\right)\times \sin\left(\frac{\pi}{2} - \sin^{-1}\frac{a(1-\cos\phi)}{L} - \frac{a}{L}\phi\right) \quad (26)$$

It is also worth noting that for small ϕ and a=L equation (26) becomes:

$$\tau_2\big|_{\phi\in[-0.1\pi,0.1\pi],a=L} = \frac{b}{a}\tau_1 \quad (27)$$

F. Dynamics of motion of the ball and the beam

Beam experiences a torque τ_2 given by equations (26) and (27).

Using Euler's Law we get beam dynamics from Figure 8 as,

$$J_2\frac{d^2\theta}{dt^2} = \tau_2 - B_2\frac{d\theta}{dt} \quad (28)$$

Dynamics of the ball are given by Newton's Law,

$$m\frac{d^2D}{dt^2} = F_{g\parallel} - F_f \quad (29)$$

Substituting the values of frictional force F_f and component of gravitational force parallel to beam $F_{g\parallel}$ in equation (29) we get,

$$m\frac{d^2D}{dt^2} = mg\sin\theta - B\frac{dD}{dt} \quad (30)$$

For θ and ϕ in range $[-0.1\pi, 0.1\pi]$ equation (30) becomes,

$$m\frac{d^2D}{dt^2} = mg\theta - B\frac{dD}{dt} \quad (31)$$

G. State space description of the system

Four differential equations (32) describe the overall system dynamics. Using Equation (3) and (31) we get:

$$\begin{aligned}\frac{di_a}{dt} &= -\frac{R_a}{L_a}i_a - \frac{k_m}{L_a}\frac{d\phi}{dt} + \frac{1}{L_a}e_a\\ \frac{d^2D}{dt^2} &= g\theta - \frac{B}{m}\frac{dD}{dt}\end{aligned} \quad (32\alpha)$$

Using equations (27) and equation (6) in (28) and (5) we get:

$$\begin{aligned}\frac{d^2\phi}{dt^2} &= \frac{k_\tau}{J_1}i_a - \frac{B_R}{J_1}\frac{d\phi}{dt}\\ \frac{d^2\theta}{dt^2} &= \frac{k_\tau b}{aJ_2}i_a - \frac{B_2}{J_2}\frac{d\theta}{dt}\end{aligned} \quad (32\beta)$$

Table 1 describes state variable assignment for equation (32 α , 32 β). Using these definitions we get following state space equations (33 α , 32 β).

$$\begin{aligned}\dot{x}_1 &= \dot{D} = x_2\\ \dot{x}_2 &= \ddot{D} = gx_3 - \frac{B}{m}x_2\\ \dot{x}_3 &= \dot{\theta} = x_4\end{aligned} \quad (33\alpha)$$

$$\begin{aligned}\dot{x}_4 &= \ddot{\theta} = \frac{k_\tau b}{aJ_2}x_7 - \frac{B_2}{J_2}x_4\\ \dot{x}_5 &= \dot{\phi} = x_6\\ \dot{x}_6 &= \ddot{\phi} = \frac{k_\tau}{J_1}x_7 - \frac{B_R}{J_1}x_6\\ \dot{x}_7 &= \dot{i}_a = -\frac{R_a}{L_a}x_7 - \frac{k_m}{L_a}x_6 + \frac{1}{L_a}e_a\end{aligned} \quad (3\beta)$$

Although we are primarily interested in stabilizing the position of the ball x_1 but to get a good control of stabilizing platform i.e. beam in this case, others state variables are also measures and designated as outputs. From table 1 we have defined following output variables,

$$\begin{aligned}y_1 &= k_D x_1\\ y_2 &= k_\theta x_3\\ y_3 &= k_\phi x_5\\ y_4 &= k_i x_7\end{aligned} \quad (34)$$

The state space equations (33)-(34) can be put into matrix form resulting in the state space quadruple given by,

$$A=\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -B/m & g & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\mathrm{B}_2/\mathrm{J}_2 & 0 & 0 & 2k_\tau/J_2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -B_R/J_1 & k_\tau/J_1 \\ 0 & 0 & 0 & 0 & 0 & -k_m/L_a & -R_a/L_a \end{bmatrix}, \qquad (35)$$

$$B=[0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1]^T,$$

$$C=\begin{bmatrix} k_D & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & k_\theta & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & k_{d\theta/dt} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & k_\phi & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & k_i \end{bmatrix},$$

$$D=\begin{bmatrix}0 & 0 & 0 & 0 & 0\end{bmatrix}^T$$

Using table 2 for values of various parameters of the system in equation (35) we obtain following final state space model of equation (36).

IV. Simulation Results

System described by equation (35) is a single input multiple output (SIMO) system. Response of the system to unit step input voltage is shown in figure 9-13. It is evident that ball's position *D* in figure 9 is unstable.

$$A=\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0.82 & 9.8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -341.3e3 & 0 & 0 & 54.6e3 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -13.65 & 1.36 \\ 0 & 0 & 0 & 0 & 0 & -61.2e-3 & -3.26 \end{bmatrix} \qquad (36)$$

$$C=\begin{bmatrix} 5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4.5 \end{bmatrix}$$

Instability in the position of the ball is due to the fact that ball accelerates under the effect of gravity on inclined beam so its velocity increases linearly and distance covered increases in a quadratic fashion. In figure 10 and 12, the beam angle θ and servo arm angle ϕ keep on increasing as motor starts rotating however as motor current stabilizes due to back emf so does the velocity of the motor. This is evident from figure 11 and 13. Hence responses of *D*, θ and ϕ are bounded input bounded output (BIBO) unstable while armature current i_a and beams angular velocity *dθ/dt* are BIBO stable.

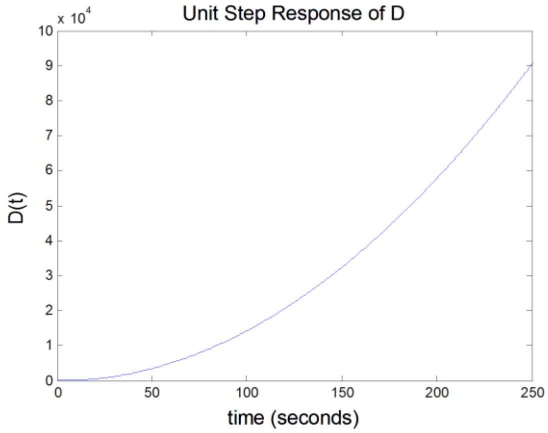


Fig. 9. Unit step response of ball position D

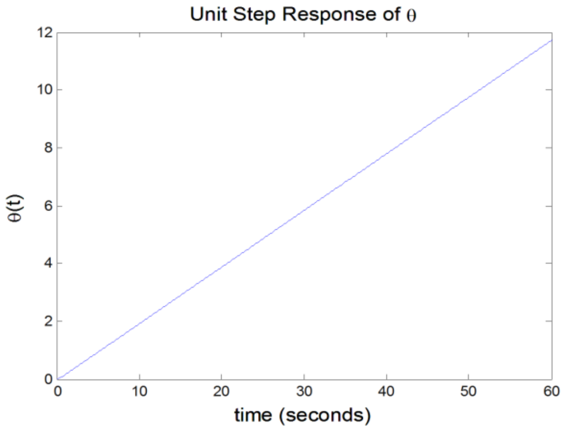


Fig. 10. Unit step response of beam angle θ

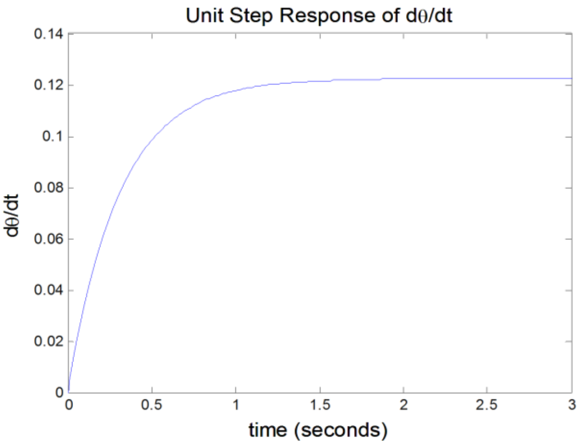


Fig. 11. Unit step response of beam angular velocity dθ/dt

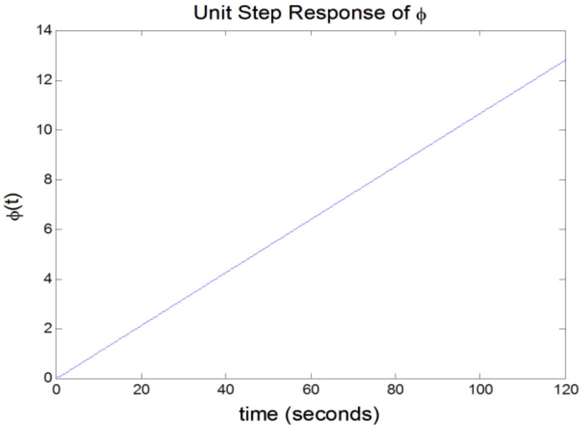


Fig. 12. Unit step response of servo arm angle ϕ

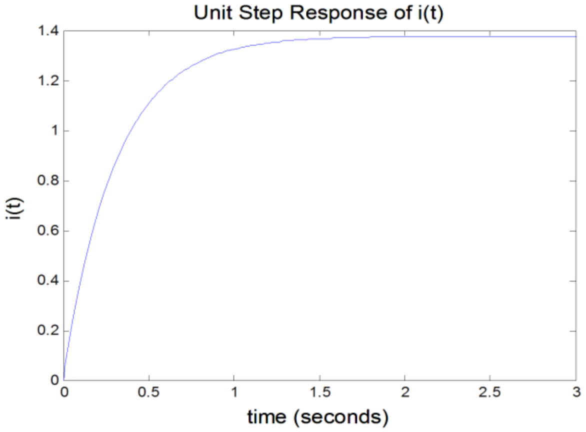


Fig. 13. Unit step response of armature current $i_a(t)$

V. Conclusions and Future Work

The detailed geometric analysis of the system shows that equations governing the system behavior are highly non-linear and relations are mostly implicit. Linearization of system dynamic is valid in very narrow operating range. Analysis of the system equations and response indicate that the system has overall unstable response and it is challenging to be controlled. Feedback is inevitable owing to the unstable nature of the system and model inaccuracies caused by linearization of dynamic. In part-II of this research work control of this system would be developed.

References

[1] K. Ogata. Modern Control Engineering. 3rd ed., New Jersey: Prentice Hall, 1997.

[2] P. R. Bélanger, Control Engineering: A Modern Approach, USA: Saunders College Pub., 1995.

[3] M. Keshmiri, A.F. Jahromi, A. Mohebbi, M. H. Amoozgar, and W. F. Xie. "Modeling and control of ball and beam system using model based and non-model based control approaches." International Journal on smart sensing and intelligent systems 5, vol. 1, pp. 14-35, 2012

[4] Y. Wen, and F. Ortiz. "Stability analysis of PD regulation for ball and beam system." In *Control Applications, 2005. CCA 2005. Proceedings of 2005 IEEE Conference on*, pp. 517-522., 2005.

[5] P. E. Wellstead., V. Chrimes, P. R. Fletcher, R. Moody, and A. J. Robins. "The ball and beam control experiment." *International Journal of Electrical Engineering Education* 15, vol. 1, pp. 21-39. 1978

TABLE I. State Variable Assignment

State variable	Physical variable	Variable description	Measured	Un-measured	Sensing mechanism
x_1	D	Distance covered by the ball on the beam	X		Wound type Linear POT
x_2	dD/dt	Velocity of the ball		X	--
x_3	θ	Beam angle	X		Analogue Accelerometer
x_4	$d\theta/dt$	Beam angular velocity	X		Analogue Rate Gyro
x_5	ϕ	Servo arm angle	X		Angular POT
x_6	$d\phi/dt$	Servo arm angular velocity		X	--
x_7	i_a	Armature current	X		Hall effect Sensor

TABLE II. VALUES OF THE SYSTEM PARAMETERS

Parameter Description	**Symbol**	**Numerical Value**
Acceleration due to Gravity	g	9.8m/sec^2
Mass of the ball	m	0.022Kg
PMDC motor rotor mass	m_R	0.050Kg
Servo arm mass	m_S	0.010kg
Link arm mass	m_L	0.010kg
Beam mass	m_B	0.025kg
Radius of PMDC motor rotor	r	0.02m
Length of link arm	L	0.375m
Width of the servo Arm	l	0.010m
Length of servo arm	a	0.1875m
Half beam length	b	0.375m
Beam width	p	0.025m
Rolling friction constant for ball	B	0.018 N.m/rad/sec
Rotational friction constant PMDC motor rotor	B_R	0.08 N.m/rad/sec
Rotational friction constant for beam about O_2	B_2	0.1 N.m/rad/sec
Torque constant for PMDC motor	k_τ	0.008 N.m/Amp
Back EMF constant for PMDC motor	k_m	0.015V/rad/sec
Ball position sensing linear potentiometer constant	k_D	5V/m
Beam angle sensing accelerometer constant	k_θ	4V/rad
Beam angular velocity gyroscopic constant	$k_{d\theta / dt}$	2.5V/rad/sec
Servo arm angular potentiometer constant	k_ϕ	3.5V/rad
Armature current hall effect constant	k_i	4.5V/A
Moment of inertia for rotor-servo-link arm assembly	J_1	5.863e-3 Kg.m^2
Moment of inertia for servo arm	J_S	4.68e-4 Kg.m^2
Moment of inertia for link arm	J_L	1.8721-3 Kg.m^2
Moment of inertia for motor rotor	J_R	3.52e-3 Kg.m^2
Moment of inertia for beam	J_2	2.93e-7 Kg.m^2
Armature resistance	R_a	0.8Ohm
Armature inductance	L_a	0.245H

Cryptic Mining in Light of Artificial Intelligence

Shaligram Prajapat
Maulana Azad National Institute of Technology
Bhopal, India

Kajol Maheshwari
International Institute of Professional Studies
Devi Ahilya University, DAVV
Indore, India

Aditi Thakur
International Institute of Professional Studies
Devi Ahilya University, DAVV
Indore, India

Ramjeevan Singh Thakur
Maulana Azad National Institute of Technology
Bhopal, India

***Abstract*—*“The analysis of cryptic text is hard problem*”, and there is no fixed algorithm for generating plain-text from cipher text. Human brains do this intelligently. The intelligent cryptic analysis process needs learning algorithms, co-operative effort of cryptanalyst and mechanism of knowledge based inference engine. This information of knowledge base will be useful for mining data(plain-text, key or cipher text plain-text relationships), classification of cipher text based on enciphering algorithms, key length or any other desirable parameters, clustering of cipher text based on similarity and extracting association rules for identifying weaknesses of cryptic algorithms. This categorization will be useful for placing given cipher text into a specific category or solving difficult level of cipher text-plain text conversion process. This paper elucidates cipher text-plain text process first than utilizes it to create a framework for AI-enabled-Cryptanalysis system. The process demonstrated in this paper attempts to analyze captured cipher from scratch. The system design elements presented in the paper gives all hints and guidelines for development of AI enabled Cryptic analysis tool.**

Keywords*—*Cipher text; Cryptic analysis; Encryption algorithm; Artificial Intelligence (AI)

I. Introduction

Originally data mining techniques are concerned with information extraction at application level or for business and commercial need of individual or organization. The term "Cryptic-Mining" is used for low level information domain. This knowledge area increases the security level of information and power of cryptic algorithms by helping cryptanalyst. In order to strengthen the cryptosystem, automated tools can be developed that intelligently exploits patterns among cipher-text, plain-text, key size, key life time and log of partially recovered plain-text-cipher text derived knowledge. Cryptic mining domain assumes that cipher texts present in the network or stored encrypted files/logs are not 100% random and exhibits some patterns. These patterns may be useful to exploit weakness using mining algorithms.

Imagine the perspective of a cryptanalyst, who is interested to know about the type of enciphering algorithm. He is also interested in obtaining the plain text from encrypted text by exploiting patters or weakness. The obvious way to deal these intractable situations is mimic different theoretical and lengthy approaches by a human mind.

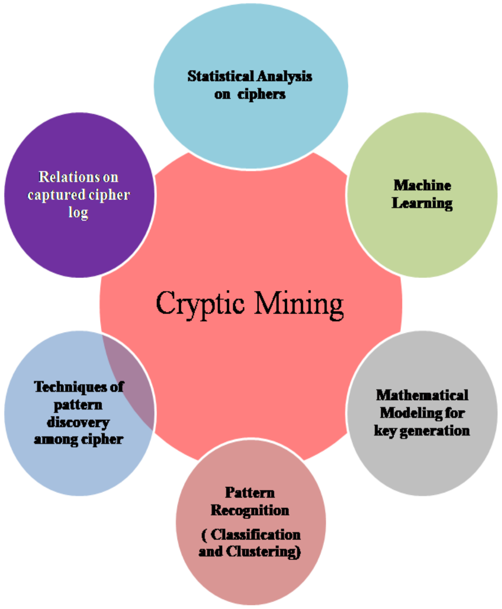

Statistical Analysis on ciphers
Relations on captured cipher log
Machine Learning
Cryptic Mining
Techniques of pattern discovery among cipher
Mathematical Modeling for key generation
Pattern Recognition (Classification and Clustering)

Fig. 1. Components of Cryptic Mining

Other alternative is to use AI and computational intelligence techniques that solves similar problems. In subsequent sections of this research work, a framework for AI enabled cryptic analysis system has been presented. This performs the cipher detection and successful conversion into plain-text in efficient way. This AI enabled system would help us to understand and analyze the various problems of cryptanalysis excluding strength and weaknesses of cryptic algorithms. This system would accept cipher texts generated from some algorithms and would try to extract meaningful information using some novel model or frameworks. Elucidation of cipher text-plain-text process has been shown on substitution cipher, such manner will resembles with the human way approach to solve the same problem. Later this concept would be generalized.

The flow diagram for schema of AI-enabled Cryptosystem has been depicted in the fig.3.It accepts a given cipher text (Substitution cipher), and attempts to transform it back to corresponding plaintext using process similar to human experts.

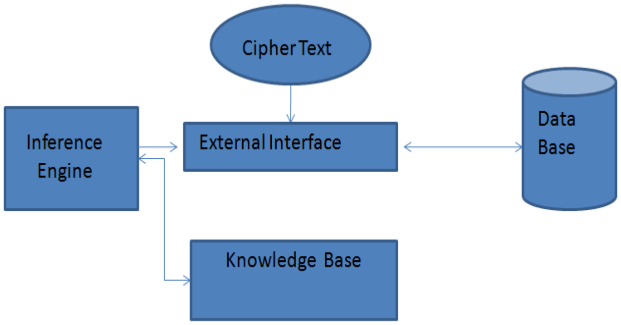


Fig. 2. Components of AI-Enabled cipher text to plain text conversion

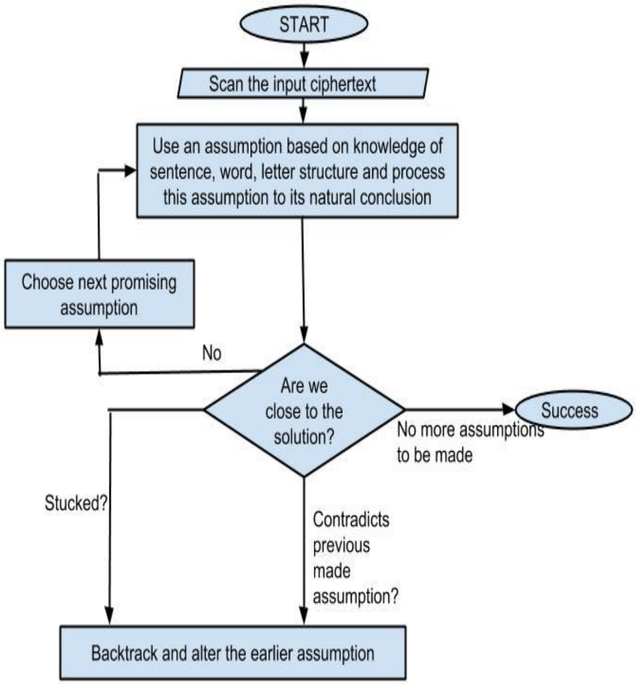

START
Scan the input ciphertext
Use an assumption based on knowledge of sentence, word, letter structure and process this assumption to its natural conclusion
Choose next promising assumption
No
Are we close to the solution?
Success
No more assumptions to be made
Stucked?
Contradicts previous made assumption?
Backtrack and alter the earlier assumption

Fig. 3. Schematic flow of AI-enabled cipher analysis system

In current typical cryptanalysis process, we limit ourselves to single substitution ciphers and we focus around "Transformation of cryptogram (cipher text) into message (plaintext) and vice-versa using single substitution cipher". In order to develop a cipher analysis system that transforms the cipher text into plaintext following steps are important: (1) Implementation of Cryptographic algorithms for producing substitution ciphertext. (2) Formulating the process of cryptanalysis. (3) Development of framework of AI-Enabled-cipher analysis system. (4) Implementation of framework for substitution ciphertext. (5) Extending the idea for categorization of cipher text generated from various symmetric key based cryptic algorithm (such as AES, DES, RC4, Blowfish and two Fish) (6) Evaluation of space and time complexity of new system.

In subsequent sections of this paper, we will describes the analysis of research topic using different examples and chalk down the system design based upon the proposed conceptual framework to be built. It includes various class diagrams and data flow diagrams describing the "dashboard". Further, system testing also has been discussed for using different examples to check functioning of each module. At the end future enhancements and new directions for further research work has been discussed in detail.

II. Basic Terminalogies

Cryptogram: A segment (word) of cipher text of length 1...n

Cryptographic Algorithms: The procedure that transforms messages (or plain-text) into cryptograms (or cipher text) and vice-versa.

Key Space: The set of possible keys K is called the key-space.

Substitution Cipher: It is the method of encoding by which units of plain-text are replaced with some other text.

Intractable Problem: Theoretically a solvable problem that takes too long time, in practice, for their providing useful solutions (e.g. deciphering cryptograms). Different alphabets are used in order to better distinguish plaintext and ciphertext, respectively. In fact these alphabets are the same.

A **cryptosystem** "S" can be defined by a 7-tuple:

S = (M, C, K_d, K_e, F, E, D) where:

M = Set of all possible **plaintext** m i.e. M= {m_1, m_2.......}. Each message m_i is the text to be encrypted (plaintext) and usually written in the lowercase alphabet: M = {a,b,c… x,y,z}.

C = Set of all possible **cipher text** c i.e. C = {c_1,c_2.......}.Each encrypted message (cipher text) c_i is usually written in uppercase alphabet: C = {A, B, C… X, Y, Z}.

K_d= Set of all possible **decryption key k** i.e. K_d = { k_1,k_2,....}

K_e=Set of all possible **encryption key k**' i.e. K_d= { k_1 ‘,k_2 ‘ ...}

F: K_d →K_e is a mapping from decryption key with corresponding encryption key. For Symmetric Cryptosystem **Kd** = **K_e** and F=I where Encryption and Decryption keys are same.

E is the relation E: K_e →(M→C) that maps encrypting keys k_e into encrypting relations e_{ke}: M→C. Each e_{ke} must be total and invertible, but need not be a deterministic function or onto.

D: K→(C→M) is the mapping that maps decrypting keys k into decrypting functions d_k: C→M. Each d_k must be a deterministic function and onto. E and D are related in that

K_e = F (k) $\subseteq$ D (k) =d_k = e_{ke}^{-1} = E $(k_e)^{-1}$ m = $D_{[k]}$ (E $_{[F (k)]}$ (M)) Often e_{ke} are one to one and onto.

III. Review of Literature

In [1], a cryptosystem has been presented that records cipher generated using information recording techniques. Then, features from this information can be extracted to distinguish one cipher from others. Also, these features can be used to transform from future information into cipher-text.

In [2] analysis of cipher text was presented by combed algorithms simultaneously to transform cipher-text into plaintext information and addressed some problems like:{Block Length detection, stream detection, entropy analysis, recurrence analysis, dictionary based analysis, decision tree based problems}.

In [3], pattern recognition based enciphering algorithms have been presented for the identification of patterns using different classification techniques like:{ SVM, Naive Bayesian , ANN, Instance based learning , Bagging , AdaBoostM1, Rotation Forest, and Decision Tree }. It can be noted that, these approaches requires improvement in accuracy with increase in number of encryption keys.

In [4], some methods have been presented with application of tools like support vector machine to identify block-ciphers for different inputs. The first one works on cipher text and second method takes partially decrypted text derived from a cipher text as input. The SVM based method performs regression using hetero-association model to derive the partially decrypted text.

Nuhn and Knight [5], worked towards automation of deciphering of ciphers. They have analyzed large number of encrypted messages found from libraries and archives, and trained by human effort only by a small and potentially interesting subset. Their work attempts to reduce human effort as well as error in decryption. Also they were interested to develop a distinguisher (first trained and then predict) to know which enciphering method has been used to generate a given cipher text.

In [6], ANN based tool has been used for decoding of a ciphertext by a pattern classification problem.

A survey of AI techniques for development of cipher analysis has been demonstrated in [7], here main objective was to investigate usage of advanced AI techniques in cryptography and they found that AI based security measures can be developed but their performance will depends on the data representation and problem formulation.

In [8], Deciphering of messages from encrypted one using genetic algorithm has been presented. It searches the key space in encrypted text. They identified limitation that it didn’t work with a two rotor problem in times comparable to those obtained using the iterative technique.

Frequency analysis in cipher-text provides a significant direction to cryptanalyst. According to Ragheb Toemeh and colleague in [9], this frequency analysis technique is used for framing objective function of cryptography. They studied the applicability of other methods like genetic algorithms for searching the key space of encryption scheme and presented cryptanalysis of polyalphabetic by applying Genetic algorithm.

Another survey based on parameters like queries, heuristics, erroneous information, group key exchange, synaptic depths has been conducted in [10], by Chakraborty and team . These parameters are suggested to improve the time complexity of algorithmic interception or decoding of the key during exchange.

In [11], A mathematical black-box model was proposed by Alallayah, AbdElwahed and Alhamami that builds the foundation for the development of Neuro-Identifier for determining the key from any given plain text-Cipher text pair. Some system identification techniques were combined with adaptive system techniques were used for the creation of the model.

All the above works and techniques follow in the direction of established long-fixed key sized algorithms. These algorithms rely on the ciphers would be secure enough if they are generated with keys of longer size. But in literature there are ciphers being generated through keys of short-fixed-length keys[12,13] varying with session to sessions. Ciphers generated through these AVK mechanism [14,15] are to be converted back into plain text.

IV. Experimental Design

For designing experimental setup it is necessary to first understand the complete mechanism of how the cipher analysis process works? How cryptanalysis applies rules of English grammar?

For this various grammar rules will be applied on the given cryptogram at different stages for each replacement which will aid in obtaining the desired plain-text.

Given following examples will be used to develop design model. Let us assume that cryptanalyst has captured following cryptogram: “*q azws dssc kas dxznn dasnn*”. Now cryptanalyst may process according to following steps:

1) To develop a model we take a hypothesis of solving a plain-text [Table 1]with one initial seed point .[Hint : wv]

2) Secondly the sentence is searched for smallest word (word with least number of letters), which in this case is the one-letter word ‘q’. This word is replaced by plain letter ‘A’ as it has the highest priority for one-letter word according to the English grammar.

3) Next the first occurrence of double letter is searched in the sentence which is ‘ss’. As it is in the middle of consonants, therefore it has to be a vowel according to English grammar and ‘s’ is replaced by plain letter ‘E’ which has highest priority in this case.

4) Further the next smallest word is searched which is ‘kea’. With this pattern the word with highest priority is ‘THE’. Hence ‘k’ and ‘an’ are replaced by ‘T’ and ‘H’ respectively.

5) Now the word having the maximum number of letters replaced is ‘HzVE’ which can possibly be ‘HIVE’(‘have’ cannot be taken as ‘A’ is already used). Therefore ‘z’ is replaced with plain letter ‘I’.

6) Next word ‘dEEc’ can be ‘SEEN’,’BEEN’,’FEEL’ etc. This word will be a verb, so we replace this word with ‘SEEN’.

7) Now our sentence includes ‘A HIVE SEEN’, which is not possible as a hive cannot see. This states that we have possibly made some mistake with our assumptions before. Backtracking to the first assumption which was qa and *changing qi to correct the sentence. Also the assumption zi has to be changed to za.*

8) Further in the next word 'SxAnn', the double letter 'nn' will be a consonant according to the English language. Therefore 'n' is replaced by plain letter 'L' which has the highest possibility in this case.

9) Now 'SxALL' can possibly be 'SMALL' or 'SHALL'. But observing the sentence structure it can be a noun or an adjective so 'SMALL' is used. Hence 'x' is replaced by plain letter 'M'.

10)Finally we obtain the plaintext from the cryptogram given.

The above process can be summarized in Table1:

TABLE I. CRYPTANALYSIS STEPS WITH KNOWLEDGE SOURCE USED INTERFERENCE

Sno	Cryptogram	Inference	Knowledge Source	Reference/ Remark
1.	q azws dssc kas dxznn dasnn	wv	using hint /KS=direct substitution	
2	q azVs dssc kas dxznn dasnn	qa	KS=small word (n-gram :n=1)	
3	A azVs dssc kas dxznn dasnn	se,	KS=double letter	
4	A azVE dEEc kaE dxznn daEnn	kt, ah	KS=small word (n-gram: n=3)	
5	A HzVE dEEc THE dxznn dHEnn	zi	pattern matching (valid small word dictionary)	Dictionary
6	A HIVE dEEc THE dxInn dHEnn	ds, cn	pattern matching ,valid smallworld dictionary, sentence structure (position of word)	KS=Patterns
7	A HIVE SEEN THE SxInn SHEnn	qi, za	Sentence structure , KS=IsSolved	Backtracking
8	I HAVE SEEN THE SxAnn SHEnn	nl	KS=Double letter, KS=word structure	
9	I HAVE SEEN THE SxALL SHELL	xm	KS=word structure, pattern matching,KS=Sentence structures	
10	I HAVE SEEN THE SMALL SHELL		KS=IsSolved	

V. EXPERIMENTAL FINDING

It can be observed that a central place (like Dashboard) is needed to apply sources of knowledge. It would be useful to align with the assumptions made and to reason the consequences. Knowledgebase (a Data structure) KS will maintain log of many different sources of knowledge such as: Knowledge about grammar, spelling and vowels. At some point of time, specialization process (moving down) is followed (General to specific) during the replace of cryptogram with n=3 and ending with "e". (for THE) and at some other points, Generalization process i.e. moving Up process is followed (from Specific to General) during the processing of cryptogram with n=4 and having pattern "?ee?"Which may be from {deer, beer, seen} but at the third position the word must be a verb instead of a noun, so "seen" should be final choice.

VI. FLOW DIAGRAM

In order to build a system flow of information from one component of system to other is depicted by fig.4, fig.5 and fig.6.

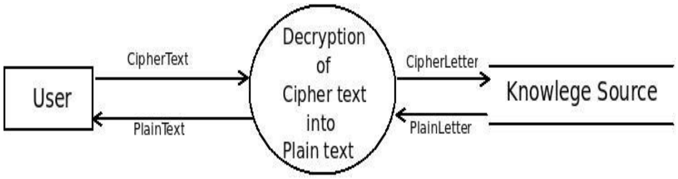


Fig. 4. Context flow diagram

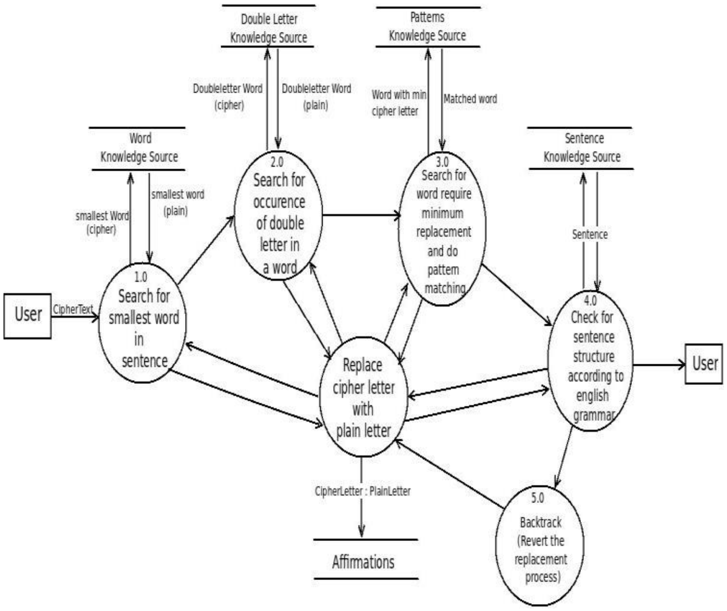

Double Letter Knowledge Source
Patterns Knowledge Source
Doubleletter Word (cipher)
Doubleletter Word (plain)
Word with min cipher letter
Matched word
Word Knowledge Source
Sentence Knowledge Source
smallest Word (cipher)
smallest word (plain)
2.0 Search for occurence of double letter in a word
3.0 Search for word require minimum replacement and do pattern matching
Sentence
User
CipherText
1.0 Search for smallest word in sentence
Replace cipher letter with plain letter
4.0 Check for sentence structure according to english grammar
User
CipherLetter : PlainLetter
Affimations
5.0 Backtrack (Revert the replacement process)

Fig. 5. First level data flow diagram

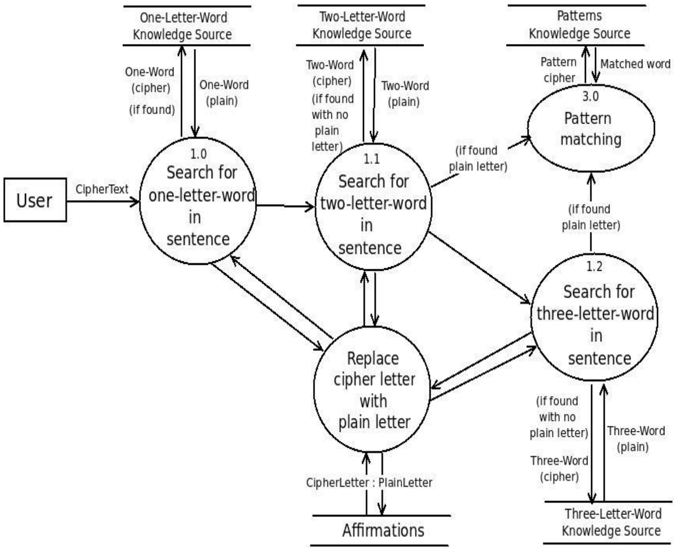

One-Letter-Word Knowledge Source
Two-Letter-Word Knowledge Source
Patterns Knowledge Source
One-Word (cipher) (if found)
One-Word (plain)
Two-Word (cipher) (if found with no plain letter)
Two-Word (plain)
Pattern cipher
Matched word
3.0 Pattern matching
User
CipherText
1.0 Search for one-letter-word in sentence
1.1 Search for two-letter-word in sentence
(if found plain letter)
(if found plain letter)
1.2 Search for three-letter-word in sentence
Replace cipher letter with plain letter
(if found with no plain letter)
Three-Word (cipher)
Three-Word (plain)
CipherLetter : PlainLetter
Affimations
Three-Letter-Word Knowledge Source

Fig. 6. Second level data flow diagram

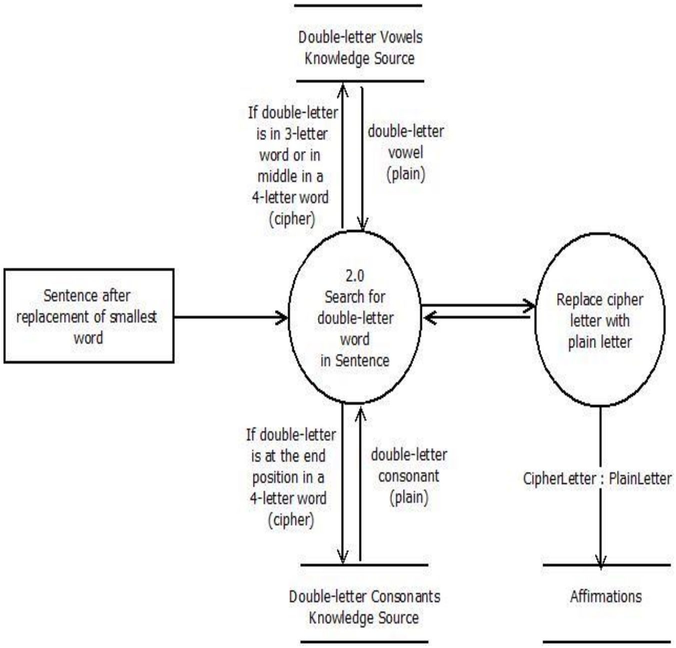

Double-letter Vowels Knowledge Source
If double-letter is in 3-letter word or in middle in a 4-letter word (cipher)
double-letter vowel (plain)
Sentence after replacement of smallest word
2.0 Search for double-letter word in Sentence
Replace cipher letter with plain letter
If double-letter is at the end position in a 4-letter word (cipher)
double-letter consonant (plain)
CipherLetter : PlainLetter
Double-letter Consonants Knowledge Source
Affimations

Fig. 7. 2-Level Data flow diagram for process 2.0

VII. Modular Structure

For implementation of cryptosystem and cryptanalysis of substitution different cipher function structures are described below:

function1-def spell_check(word)

This module checks the spelling of the word and returns true if the spelling is correct.

function 2-def replacefunc(word, file_word)

This module replaces the word with a word from file and adds the entry in assumption(dictionary containing cipher Letter-plain, Letter pair)

function 3.-def transposition( )

This function displaces the cipher letter with plain letter according to the displacement in the plain letter with its corresponding cipher letter (key) in the assumption (dictionary). If the words replaced don't have correct spelling then the transposition is reverted back and the plain letters are again replaced with corresponding cipher letters which were added to assumption dictionary.

function 4.-def backtrack(word)

If no pattern match is found for a word then that word is passed as the argument to backtrack, it will replace the plain letter with their corresponding original cipher letter as the #assumptions made before was not correct

function 5.-def trans_status( )

After doing transposition it checks whether the transposition made was correct or not.

function 6.-def revert_trans( )

If the transposition made was correct then it displays the final sentence otherwise revert all the #changes made during transposition process

function 7.-def pat_rep(lst, fil, cnt)

pat_rep function replaces the words from list with suitable word from file according to condition. It has three arguments:

lst**:** list of specific words(i.e 2-letter, 3-letter etc) if the sentence containing cipher.

fil: text file of containing 2-letter-letter etc plain-letter words corresponding to list.

cnt: counter to mention the position in the file

function 8.-def pattern(word, fil, cnt)

If the word contains one or more plain letter ***pattern function*** matches the word with every word in file and replaces if a pattern is matched. It has 3 arguments:

word: word from sentence containing a capital letter

fil: corresponding file(for ex: 4_word file for 4-letter word)

cnt: counter that mentions position in the file

function 9.-def double_letter(word)

This function checks if a word (input) contains any double letter, if yes it replaces the double letter cipher with appropriate plain letter according to its position (i.e. if in middle it will be a vowel and if end it will be a consonant according to English grammar rules)

function 10.-def one_letter( )

If the sentence contains ***one-letter-word*** in cipher then this function will replace that cipher with the possible plain one-letter-word and will make entry according to the assumption.

function11-def find_key(value)

This function finds the corresponding ***cipher(key)*** letter of the plain ***letter(value)*** given as argument from the dictionary "assumption"

VIII. Test Case Development

Test cases are developed to validate and verify the working of system in two situations. Case-1 and Case-2.

Case-1: For testing transposition

Let sentence given by user: sent_1 = “k co c iktn”

Case-2: For testing english grammar

Input Sentence supplied by user:

sent = “dwer er ed”

TABLE II. Steps Followed for Case 1

S.no	Module name	Test Cases	Result	Response
1.	Enter valid cipher sentence	Check chars of sent	Returns true if the sentence contains only alphabets otherwise false	OK
2.	**one_letter()**		Replace one-letter cipher word with the plain word chosen from the file containing one-letter words	OK
3.	one_letter() sent=”k co c iktn”	action performed on sent	sent=”I Ao A iItn”	OK
4.	**transposition()**		Finds the difference between the replaced cipher letter and its corresponding plain letter and replace remaining cipher letter with plain letter with same difference	OK

5.	transposition()
	sent = “I Ao A iItn”
		action performed on sent	sent = “I AM A GIRL”	OK

6.	**spell_check (word)**		Returns true if correct spelling else false	OK
6.1	spell_check (“I”)	correct spelled word	Returns true	OK
6.2.	spell_check (“AM”)	correct spelled word	Returns true	OK
6.3.	spell_check (“A”)	correct spelled word	Returns true	OK
6.4.	spell_check (“GIRL”)	correct spelled word	Returns true	OK
7.	**trans_status()**		Returns false if spelling of any of the word in sent is wrong else true	OK

8.	trans_status()
	sent=”I AM A GIRL”
		all words are correct spelled	returns true	OK

TABLE III. Steps Followed for Case 2

S.no	Module name	Test Cases	Result	Response
1.	main()		Calls all functions according to condition	OK
2.	pat_rep(lst,fil,cnt)		Search the word from list from file according to conditions met	OK
3.	replacefunc(word,file_word)		Replace each chars of word with the corresponding chars of file_word and made the entry of pair(cipherletter:plainletter) in the dictionary ‘assumption’	OK
4.	pattern(word,fil,cnt)		Search the word from list according to the pattern formed	OK
5.	backtrack(word)		Reverts back the previous assumptions made if pattern is not found for *word* i.e replace the plain text with their original ciphertext in sent	OK
6.	check_sent(sent)		Checks the grammar of sentence and returns true if correct else false	OK
7.	transposition()		Finds the difference between the replaced cipher letter and its corresponding plain letter and replace remaining cipher letter with plain letter with same difference	OK
8.	Enter valid cipher sentence	Check each characters of sent	Returns true if the sentence contains only alphabets otherwise false	OK
9.	main() started			
10.	sent = “dwer er ed”	check for smallest word	Two-letter word found	OK

11.	pat_rep(two_w,tw,cnt2)
	two_w = [er,ed]
		Actions performed on the words of list two_w,hence on sent	Replaces ‘er’ with ‘OF’(first word in fil tw)
			sent = “dwOF OF Od”
				OK

12.	replacefunc(er,OF)	Replacement done on sent	Replaces ‘er’ with ‘OF’(first word in fil tw) sent = “dwOF OF Od”	OK

			assumption={‘e’:’O’ , ’r’:’F’}	
13.	pattern(Od,etw,e2cnt) etw: file containing 2-letter words at ending position of sent	Search matched word	No pattern found for pattern=’O.’	OK
14.	backtrack(Od)	Action performed on sent and assumption	Replaced all the plain letter i.e. F and O with corresponding cipher letter from assumption and calls main() again	OK
15.	main() called			
16.	sent = “dwer er ed”	Check for smallest word	Two-letter word found	OK

16.1	pat_rep(two_w,tw, cnt2)
	two_w = [er, ed]
			Start search for the word in “tw” according to condition after the word last searched	OK

16.1.1	replacefunc(er,TO)	Replacement done on sent	Replaces ‘er’ with ‘TO’
			sent = “dwOF TO Od”
			assumption={‘e’:’T’ , ’r’:’O’}
				OK

16.1.2	pattern(Td,etw,e2cnt) etw: file containing 2-letter words at ending position of sent	Search matched word	No pattern found for pattern=’T.’	OK
16.1.2.1	backtrack(Td)	Action performed on sent and assumption	Replaced all the plain letter i.e. T and O with corresponding cipher letter from assumption and calls main() again	OK
17.	main() called			
18.1	sent = “dwer er ed”	Check for smallest word	Two-letter word found	OK

19.	pat_rep(two_w,tw, cnt2)
	two_w = [er, ed]
			Start search for the word in “tw” according to condition after the word last searched	OK

19.1.1	replacefunc(er,IS)	Replacem-
		ent done on sent
			Replaces ‘er’ with ‘IS’
			sent = “dwIS IS Id”
			assumption={‘e’:’I’ , ’r’:’S’}
				OK

19.1.2	pattern(Id,etw,e2cnt)
	etw: file containing 2-letter words at ending position of sent
		Search matched word	Match found for pattern=’I.’
			match = “IT”
			sent = “TwIS IS IT”
			assumption={‘e’:’I’ , ’r’:’S’, ‘d’:T}
				OK

20	return to main()			
21	transposition()	Check difference between cipher letter and plain letter	Difference is not same, therefore returns False	
22	sent = “TwIS IS IT”	Check for word having length greater than two	Four-letter word found	OK

22.1	pat_rep(four_w,fw, cnt4)
	four_w = [TwIS]
			As word contains plain letter so calls pattern()	OK

22.2	pattern(TwIS,sfw,s4cnt)
	sfw: file containing 4-letter words at starting position of sent
		Search matched word	Match found for pattern=’T.IS’
			match = “THIS”
			sent = “THIS IS IT”
			assumption={‘e’:’I’ , ’r’:’S’, ‘d’:’T’, ‘w’:’H’}
				OK

sent = “THIS IS IT” and assumption={‘e’:’I’ , ’r’:’S’, ‘d’:’T’, ‘w’:’H’}				
23	check_sent(sent)	Check grammar of sent	Returns true	

IX. Conclusion

This paper is an attempt to demonstrate the demonstrate cipher text -plain text conversion process for analysis of cryptic text. AI has been used to get the feasible solution of hard problem. By generalizing the conversion process system for obtaining plain-text from input cipher text is the central objective. The developed system would analyze and learn for pruning. This paper has demonstrated cipher text-plain text process completely and created a framework for AI-enabled-Cryptanalysis system, Data Flow Diagrams and appropriate test cases. This schema and plan would be suitable for development of AI enabled Cryptic analysis tool and in turn they will evaluate strength of any cryptosystem.

X. Future Enhancement

AI-based-crypto system works correctly for the basic cipher-text to plain-text conversion process. To extend this further to fulfill various requirements following enhancements are suggested.

1) Current work can be extended to incorporate ciphers other than substitution and transposition cipher. That is, present system response is fine for transposition cipher and substitution cipher, but cipher types are more than two. This will require testing with different algorithm, method and cipher text. So that extended version is fit and deciphers it accordingly.

2) Incorporating plain-text of multiple languages in the process is also desirable. That is, current elucidation demonstrated in this work deciphers and outputs result in English. Maximum number of ciphers gives English plain-text on decryption. But over the communication channel languages local, non-English languages are also exchanged. For decryption of cipher text yielding other language plain text, the grammar rules of that particular language has to be applied.

3) Extension of character set with adding special characters and symbols will make the current system more flexible. The reason behind this is, day-by-day increasing amount of data transferred, and the need to encrypt it in a more complex way is mandatory for securing information from unauthorized users. Hence special characters and numbers are used to generate a more complex cipher patterns. Deciphering these ciphers using algorithm with condition for checking these symbols together with the English alphabets will be necessary.

4) Extension for n-gram (n>4) will increase the power of cipher analysis. That is checking cipher words with having length more than 4 and words which are not present in any knowledge source, needs to be worked out. Currently the Knowledge source, include files having upto 4-letter words. More generalized approach is needed for words having length more than 4. This may require a tool for checking the spellings of every possible word which states that the spelling is correct or not.

ACKNOWLEDGMENT

This work is supported by research project under Fast Track Scheme for Young Scientist from DST, New Delhi, India. Scheme 2011-12, No. SR/FTP/ETA-121/ 2011 (SERB), dated 18/12/2012.

REFERENCES

[1] Khadivi P, Momtazpour M. “Ciphertext classification using data mining”,International Symposium on Advanced Networks and Telecommunication Systems IEEE- ANTS, pp.64-66, 2010.

[2] Shivendra Mishra, Dr. Aniruddha Bhattacharya , “Pattern analysis of cipher text : a combined approach”, proceeding of Recent Trends in Information Technology (ICRTIT), 2013 International Conference on ,25-27 July 2013,393 - 398, DOI:10.1109/ICRTIT.2013.6844236.

[3] Sushila Omer Sharif, saad P. Mansoor , “Performance evaluation of classifiers of encryption algorithm”, ACEEE International Journal on Network Security , Vol. 02, No. 04, Oct 2011.

[4] S. Swapna, A. D. Dileep, C. Chandra Sekhar, and Sri Kant, “Block cipher identification using support vector classification and regression,” Journal of Discrete Mathematical Sciences and Cryptography, vol. 13, no. 4, pp. 305- 318, August 2010.

[5] Malte Nuhn, Kevin Knight , “Cipher type detection”, http://www-i6.informatik.rwth-aachen.de/~nuhn/2014-classification-poster.pdf, pp.1769–1773,2014

[6] Sambasiva Rao Baragad, P. Satyanarayana Redd , “Studies on the advancements of Nerual Networks and Neural Network based cryptanalytic works”, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),Volume 2, Issue 5, September – October 2013 ISSN 2278-6856

[7] E.C. Laskari, G.C. Meletio, Y.C. Stamatiou, M.N.Vrahatis ,“Cryptography and Cryptanalysis through Computational Intelligence”, proceeding of Computational Intelligence in Information Assurance and Security, Studies in Computational Intelligence Volume 57,2007, pp 1-49

[8] A.J. Bagnall, G.P. Mckeown, V.J. Rayward Smith , “Cryptanalysis of a three rotor machine using a genetic algorithm”, In proceedings of 7th International Conference on genetic algorithms,ICGA97,1997.

[9] Sandip Chakraborty, Jiban Dalal, Bikramjit Sarkar, Debaprasad Mukherjee , “Neural Synchronization based secret key exchange over public channels: A Survey”, Journal of Engineering Science and Technology Review 8(2) (2015) 152 – 156.

[10] Khalid Alallayah, Moamed Amin, Wail AbdElwahed, Alaa Ahamami ,“Applying Neural Networks for simplified data encryption standard (SDES) cipher system cryptanalysis”, Vol. 9, pp.163-169,2012.

[11] Shaligram Prajapat, A. jain, R.S.Thakur, “A Novel Approach For Information Security with Automatic Variable Key Using Fibonacci Q-Matrix”, International Journal of Computer & Communication Technology (IJCCT) ISSN (ONLINE): 2231 - 0371 ISSN (PRINT): 0975 – 7449 Vol-3, Iss-3, 2012, p.p. No. 54-57.

[12] Shaligram Prajapat, D. Rajput, R. S. Thakur, "Time variant approach towards Symmetric Key " ,Science and Information Conference 2013,October 7-9, 2013, London, UK p.p.398-405.,Technically Co-Sponsored by: IEEE Computer Society, UKRI Section, IEEE Computational Intelligence Society,UKRI Section, IEEE Consumer Electronics Society,, IEEE Sponsored and Organized by The Science and Information (SAI) Organization

[13] Shaligram Prajapat, R.S. Thakur, "Towards Optimum size of key for AVK based cryptosystem", Communicated and CJICT, Nigeria in June-Dec. 2015.ISSN (Online): 2354 - 3507; ISSN (Print): 2354 - 3566

[14] Shaligram Prajapat, R.S.Thakur, "Markov Analysis of AVK Approach of Symmetric Key Based Cryptosystem ",Computational Science and Its Applications, ICCSA 2015,Springer LNCS: Volume 9159, 2015, pp 164-176,Jun 2015,doi:10.1007/978-3-319-21413-9_12,ISBN:9783319214139 and 9783319214122.

[15] Shaligram Prajapat, R. S.Thakur, "Cryptic-Mining: Association Rules Extractions Using Session Log ", Computational Science and Its Applications, ICCSA 2015,Springer LNCS: Volume 9158, 2015, pp 699-711,Jun 2015,doi:10.1007/978-3-319-21413-9_12.

[16] Claudia Oliveira, J. A. Xex´eo, C. A. Carvalho,”Clustering and Categorization Applied to Cryptanalysis”,Taylor and Francis 2007

[17] M.F. Uddin and A.M. Youssef,Cryptanalysis of simple substitution ciphers using particle swarm optimization, Evolutionary Computation, 2006. CEC 2006.IEEE Congress on, 0-0 2006, pp. 677 -680.

[18] George Nagy, Sharad C. Seth and Kent Einspahr, “Decoding Substitution Ciphers by Means of Word Matching with Application to OCR”, 1987

[19] Amrapali Dhavare, Richard M. Low & Mark Stamp ,“Efficient Cryptanalysis of Homophonic Substitution Ciphers”, 2013

[20] Grady Booch, Robert A. Maksimchuk, Michael W. Engle, Bobbi J. Young(Ph.D.), Jim Conallen, Kelli A. Houston, “Object-oriented Analysis and Design with applications”, Addison-wesley publishing company, Rational, Santa Clara, California 3rd Edition

[21] S. William and Stalling, Cryptography And Network Security, 4/E. Pearson Education India, 2006.

[22] http://www.nltk.org

[23] http://what-when-how.com/artificial-intelligence/automated-ryptanalysis-artificialintelligence/

[24] http://cse.ucdenver.edu/~rhilton/docs/Cryptanalysis-Against-Monosub-Ciphers.pdf

[25] http://people.csail.mit.edu/hasinoff/pubs/hasinoff-quipster-2003.pdf

[26] http://scottbryce.com/cryptograms/stats.htm

[27] http://jeremykun.com/2012/02/03/cryptanalysis-with-n-grams/

Evaluation and Improvement of Procurement Process with Data Analytics

Melvin Tan H.C., Wee-Leong Lee
School of Information Systems
Singapore Management University
80 Stamford Road
Singapore 178902

***Abstract*—Analytics can be applied in procurement to benefit organizations beyond just prevention and detection of fraud. This study aims to demonstrate how advanced data mining techniques such as text mining and cluster analysis can be used to improve visibility of procurement patterns and provide decision-makers with insight to develop more efficient sourcing strategies, in terms of cost and effort. A case study of an organization's effort to improve its procurement process is presented in this paper. The findings from this study suggest that opportunities exist for organizations to aggregate common goods and services among the purchases made under and across different prescribed procurement approaches. It also suggests that these opportunities are more prevalent in purchases made by individual project teams rather than across multiple project teams.**

***Keywords*—procurement; text mining; clustering; data analytics; fraud detection**

I. Introduction

Several procurement lapses in Singapore have brought procurement risk into the spotlight [1]. Today, stakeholders are demanding to know whether their money is being spent wisely and not used on fraudulent expenditure. While steps were taken to improve internal control by strengthening procurement guidelines and policies, another approach has gathered pace – integrating data analytics into the procurement process to help prevent and detect fraud. Although this has been the key objective of the application of analytics in procurement, this study has shown that there are other benefits to be reaped from its application. Analytics improve visibility of procurement patterns and empower stakeholders with better insight for developing more efficient sourcing strategies, in terms of costs and effort.

In Organization X (the organization involved in this study chose to remain anonymous for privacy reasons), under their procurement rules and principles, the prescribed procurement approach is based on the Estimated Procurement Amount (EPA) of the intended procurement. This can be summarized in Table 1.

TABLE I. Procurement Approach

EPA	Procurement Approach	Sourcing Methods
Up to $3,000	• Small Value Purchase (SVP) • Established Term Contract (ETC)[1]	• Verbal or written quotes • Off-the-shelf purchase • Call For Quotation[2] (CFQ)
Between $3,001 to $70,000	• Open Quotation (OQ) • ETC	• Invite For Quotation (IFQ) • CFQ
Above $70,000	• Open Tender (OT) • ETC	• Invite For Tender (IFT) • CFQ

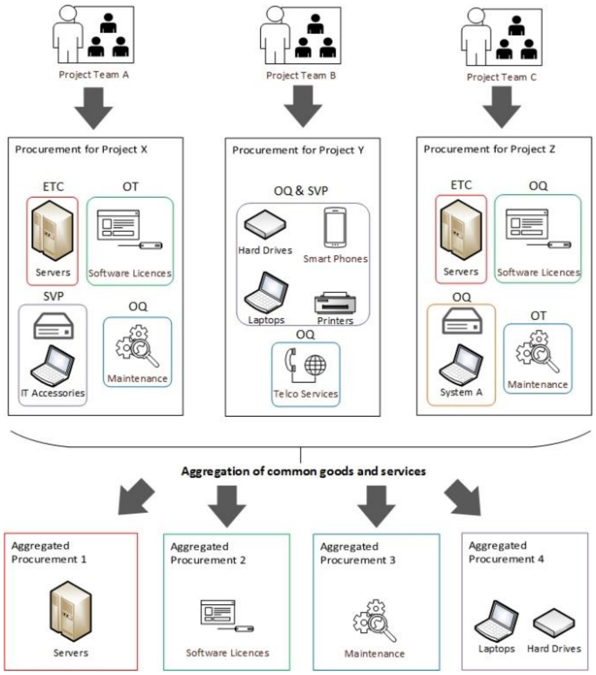


Fig. 1. Current situation in Organization X and the desired outcome

[1] Organization X has combined purchases of common goods and services by establishing a Term Contract to yield better value for money through economies of scale. With an Established Term Contract (ETC), Organization X can then procure directly from the appointed supplier(s) when the product or service is required during the contractual period.

[2] If the particular item or an equivalent functional item can be obtained from more than one supplier, all the contracted supplier(s) in the Established Term Contract (ETC) who are deemed to be capable of supplying the item should be approached for quotations.

The mission of Organization X is to provide effective and timely information and communications technology (ICT) support and solution. In Organization X, procurement of goods and services is proposed by project teams focuses on various areas and projects with different objectives and timelines. Based on the EPA of their intended procurement, their procurement needs are subsequently carried out in separate transactions using the prescribed procurement approaches stated in Table 1. There could be opportunities to aggregate common goods and services across the various areas and projects to achieve possibly higher economies of scale and lessen the administrative efforts. By consolidating frequent purchases into a contractual agreement, Organization X can exploit economies of scale to obtain favorable prices and reduce the transactional overhead of subsequent acquisitions of the same item by performing it upfront. An illustration of the current situation in Organization X and the desired outcome is shown in Fig. 1.

The main purpose of this study is to propose a model where analytics can be applied to detect such opportunities and derive meaningful insights that would lead to improvement in the current procurement process. This paper is organized as follows: Section II provides a literature review of efforts made in application of analytics in procurement and the proposed data mining techniques. Section III presents the data and input variables used in the analysis. Section IV shares the design considerations of the analysis, proposed model and methodology for the analysis. In Section V, the results and discussion of the findings are presented. Section VI concludes the paper and proposed recommendations to the findings and suggest areas for future improvements.

II. Literature Review

Kemp [2] commented that analytics have been used at an advanced level for years to combat fraud in the private sector, especially in financial services. He advocated the following approaches which are proven within the private sector:

- Rules-based detection - identifies potential instances of fraud based on behaviors already proven to be fraudulent.
- Anomaly detection - spots unknown or unexpected patterns by comparing like-for-like data within groups.
- Advanced analytics – applies the latest data, text and web mining technologies to identify fraudulent and errant behaviors that have not already been spotted by rules-based and anomaly detection approaches.

Byrne [3] suggested that much effort has gone into ensuring minimization of fraud, misconduct and other unethical behaviors in procurement and it is time for procurement to start adding value to Organizational strategies and to move from what many perceive as a policing role to a value added role. He added that procurement should be managed strategically and this requires analysis of past procurement spending to determine if you can combine individual purchases to cut costs.

National Fraud Authority [5] highlighted a number of inefficiencies in public procurement, some of which were due to departmental autonomy over procurement. It was recommended that government should leverage its purchasing power by seizing opportunities to procure as a single entity. The Organization for Economic Co-operation and Development (OECD) [6] has also identified that savings are being sought, through a variety of measures including centralization of the procurement function, the aggregation of purchases in order to achieve economies of scale.

Chae and Olson [4] discussed the role of analytical capability for sourcing in Supply Chain Management. There is a strong application of analytical IT to support supplier selection within supply chains. Prescriptive analytics has been a key enabler of manufacturer's sourcing-related decision making. Predictive analytics techniques are increasingly available these days for intelligent material planning, inventory management, and supplier relationship management. For instance, advanced machine learning techniques such as artificial neural networks and support vector machines are promising tools to enable effective sourcing. Pattern recognition, when used with large sets of historical purchase orders and supplier delivery data, can reveal hidden facts and potential problems with processes and performances.

Kantardzic [7] proposed that market search, business-intelligence gathering, e-mail management, claim analysis, e-procurement and automated help desk are only a few of the possible applications where text mining can be successfully deployed. Miner et al [9] discussed efforts on clustering in Natural Language Processing (NLP) and how necessary information extraction from the meaning of the text can be performed.

To analyze demand before it can be aggregated, Chia and Chen [8] discussed how Business Intelligence (BI) methodologies using the online analytical processing (OLAP) concept such as drilling, pivoting, dicing and aggregating can be applied to the unstructured content found in procurement databases by performing any ad-hoc query. This would allow users to derive transaction trends at any hierarchy and resolution.

The literature review suggests the following:

a) There is no doubt on the benefits that analytics can bring to procurement, beyond its application in the prevention and detection of fraud.

b) The principles behind the analytical approaches applied successfully for fraud detection and prevention could be used to identify opportunities for aggregation of purchases.

c) Advanced data mining techniques such as cluster analysis, text mining etc. could be applied in procurement.

d) BI tools such as OLAP are commonly used to analyze procurement data. There is limited literature to suggest widespread application of advanced data mining techniques to analyze procurement data.

III. Data Set and Input Variables

The data set consists of procurement transactions from Year 2011 to 2014 (inclusive), related to the three main types of procurement approaches, namely SVP, OQ and ETC highlighted in Table 1. This period was selected based on completeness of records (for more accurate insights drawn from the results of the analysis) and recency (for more meaningful follow-up action on the insights derived).

The records for SVP are provided by the Finance Section of Organization X. The Finance Section maintains this information in Excel Spreadsheets. There are 20,861 records for SVP. The records for OQ and ETC are extracted from the procurement databases. There are 267 and 118 records for OQ and ETC respectively.

As there are many data fields available in these records, only the data fields containing information relevant to this study need to be identified. Data fields of interest would include information on description of the purchased items, amount spent, period of transaction, supplier and buyer details.

The records underwent extensive data exploration, manipulation and cleaning to prepare them for analysis:

a) SVP

i. Filter transactions amounting between $0 and $3,000.

ii. Remove transactions relating to:

- Claims (transport, dental, medical, travel)
- Training-related payment
- Public utilities (Power, water etc.) and telecommunication-related payment
- Contract-related payment.

iii. Remove transactions with missing fields.

iv. The number of SVP transactions was reduced to 785 records.

b) OQ and ETC

i. Categorize the transactions by year.

IV. Methodology

Given the nature of the procurement process in Organization X, it would be interesting to see how the three analytic approaches prescribed by Kemp [2] for combating fraud could be adapted and applied to Organization X's procurement process to meet the objective of this study:

a) Rules-based detection – Logically, if frequent purchases can be consolidated, they can also be split. In order to avoid the rules put in place for a higher EPA procurement approach, which one might perceive as being more stringent in terms of approval and more time-consuming in terms of administrative efforts, a high-value purchase might be split up into multiple low-value purchases to be carried out separately. For instance, an EPA of $10,000 which rules prescribed a procurement approach via an OQ might be split into multiple lower value purchases via the SVP procurement approach. Hence, the scope of analysis should cover the transactions related to the different procurement approaches determined by the EPA of the intended procurement, namely:

i. SVP - Goods and services of similar nature can be aggregated into OQs.

ii. OQ - Goods and services of similar nature can be aggregated into OTs.

iii. ETC – While goods and services available under the ETCs are in general already aggregated, a better understanding of how they were purchased will improve procurement planning and possibly reduce the administrative efforts involved in issuing CFQs.

b) Anomaly detection – In the context of consolidating frequent purchases, observations which are beyond the norm i.e. abnormal trends and patterns could be:

i. Recurring transactions of similar goods and services.

ii. Dominant suppliers, in terms of number of transactions and amount spent.

c) Advanced analytics – Text Mining and Clustering techniques could be appropriate given the nature of the data.

In the organization currently, the account codes i.e. expense codes are used to categorize purchases into categories that describe the nature of the goods or services purchased. However, these could be too generic to derive any further useful information on the goods or services purchased. For instance, an IT-related equipment purchase is only categorized into hardware, software or communication equipment and network. In addition, the account codes are allocated by the purchasers on a best-effort basis and these are subjected to judgment errors. Hence, the description fields which contained information on the actual goods or services purchased would reveal more details about the transactions than the account codes.

A key problem with the description fields is that they are captured in free text format. Therefore, they are likely to include a significant amount of irrelevant and noisy information such as dates, names, teams, etc. To the untrained eye, the goods or services purchased would not be intuitively noticeable.

Taking the above into account, a text mining with cluster modelling approach (a conventional data mining technique) using RapidMiner (a software platform that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics) is used in this study. The text processing algorithms in the RapidMiner's text mining extension will transform the text data i.e. the description fields into a format that can be easily analyzed using cluster modelling technique. Each record is characterized in terms of the occurrence frequency of words in it and the cluster modelling will place together the records that have a similar distribution of word frequencies. The higher the occurrence of the words would imply a higher number of transactions i.e. purchased more times. However, it is important to bear in mind that a higher number of transactions might not necessarily mean that a larger amount is spent correspondingly. In other words, a drawback of the text mining with cluster modelling is that it would not be able to tell whether a specific good or service with many transactions is of relatively small value. Similarly, it would also miss out a specific good or service with few transactions but of relatively higher value.

The text mining with cluster modelling process in Rapid Miner is illustrated in Fig 2.

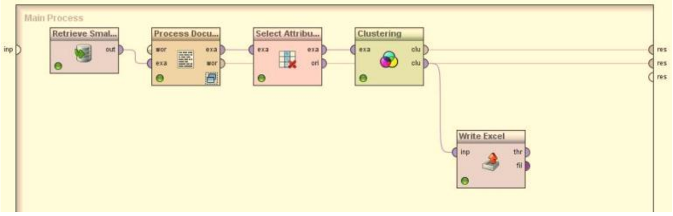


Fig. 2. Text mining with cluster modelling process in Rapid Miner

The functions of the 5 main operators are:

a) Retrieve data - Reads data from the file

b) Process Documents from Files – Processes the text using eight operators nested within:

i. Tokenize Non-letters (Tokenize).

ii. Tokenize Linguistic (Tokenize).

iii. Filter Stopwords (English).

iv. Filter Stopwords (Dictionary).

v. Filter Tokens (by Length).

vi. Stem (Porter).

vii. Transform Cases.

viii. Generate n-Grams (Terms).

c) Select Attributes – Only columns with numeric values are selected for clustering (due to the choice of the clustering technique, K-Means, in the following step).

d) Clustering – K-Means clustering algorithm is used to group the records in terms of the occurrence frequency of words in it. K-Means is selected for its simplicity and speed among the clustering techniques.

e) Write Excel – The output i.e. cluster groups is saved into an Excel file so that they can be combined with the other data fields i.e. amount spent, period of transaction i.e. Year, supplier and buyer details to derive further insights.

For each group of data (SVP, OQ and ETC) the methodology for the analysis is as follows:

a) Approach 1

i. Input the data into the designed process in Rapid Miner.

ii. Run the process.

iii. Examine the results – Word List, Cluster Model and Centroid Table.

iv. Re-calibrate the parameters accordingly and repeat from Step ii. Compare the results of multiple runs with different k classes and choose the best one.

v. Combine the results of the selected cluster with other data fields i.e. amount spent, period of transaction, supplier and buyer details.

vi. Explore and analyze the results for insights.

b) Approach 2

i. Tabulate the data by the supplier details against the period of transaction, amount spent, number of transactions made and the buyer details.

ii. Identify suppliers from whom purchases were made in every year.

iii. Extract the records for the identified suppliers.

iv. For those with many transactions, follow the steps in Approach 1 using the data obtained in Step 3.

v. Explore and analyze the results for insights.

Approach 1 begins with the identification of the common goods and services purchased. The output is analyzed together with related information on the amount spent, period of transaction, supplier and buyer details. Approach 2 begins with the identification of suppliers based on value and volume of transactions. Both approaches incorporate further data points to address the drawback of the model which would not be able to tell us whether a specific good or service with many transactions is actually of significance. Both approaches aim to complement each other and their results can be compared against each other for verification and completeness when drawing the conclusion.

V. Results and Discussion

A. *SVP*

Approach 1

An examination of the Wordlist generated from all the records, sorted by total number of occurrences and number of documents that contain these words, revealed high occurrences of irrelevant words such as purchase, service, supply, etc. Including these words in the clustering process will affect the results. As such, these words will be identified and added into the "Stopwords" list under "Filter Stopwords (Dictionary)". The cleaned-up wordlist now presented words such as PhoneX, printer, TabletX, cable, rubber stamp, book, fruit, camera, biscuit, BTH, bowl, certificate, screen, etc. with highest occurrences. This provided a clearer indication of the purchases made and the keywords that should be watch out for in subsequent steps of the analysis.

In determining the 'k' value (i.e. the number of clusters for k-mean clustering), the "rule of thumb" [10] ($k= \sqrt{(n/2)}$ where n is the number of data points) was used as it is a quick and simple method. For 785 records, k=20 was derived. For verification and comparison purposes, additional runs were also made for k=25 and k=30.

The Centroid Cluster Model in Table 2 shows the results, interpreted based on the term frequency of the keywords generated from each cluster (for k= 20, 25 and 30). The results at k=25 and k=30 seem to indicate that the keywords in the majority of the clusters were recurring e.g. PhoneX, screen, fruit, juice, biscuit, printer, book etc., indicating a strong presence, albeit further breakdown of each into smaller clusters e.g. fruits appeared in Cluster 3, 19 and 21 at k=30. New keywords which appeared in the cluster at k=30 consist of fewer items which were not significant.

TABLE II. SVP: CENTROID CLUSTER MODEL FOR DIFFERENT "K" VALUES

k=20	***k=25***	***k=30***
0 – H Tags, Access (Rooftop)	0 – HT, certificate (ET, Net ID)	0 – HT, certificate (ET, Net ID)
1 – PhoneX	1 – Biscuit, Tin, Voucher, Bowling, BTH, Camera, Tape, Trophies	1 – PhoneX
2 – PhoneX	2 – Camera, Photocopier	2 –Nil
3 – Nil	3 – PhoneX	3 – Mandarin Oranges (Fruits)
4 – Mandarin Oranges (Fruits)	4 – PhoneX	4 – Camera (CCTV)
5 – Rubber stamp	5 – Camera (CCTV)	5 – Power (Supply)
6 – Camera (CCTV)		6 – Security Holograph sticker
7 – Card, pouch	6 – Mandarin Oranges (Fruits)	7 – Office (Phones, Chairs), Certificates (Appreciation, SSL)
8 – Tea capsule, voucher	7 – Juice	8 – Printer
9 – Power (Supply)	8 – Ops, Note	9 – Case (Peli, TabletX)
10 – Book, Biscuit, Voucher, BTH, Tape, Battery, Juice, Christmas decor, Grocery	9 – SSL	10 – Access (Rooftop)
11 – Case (Peli, TabletX)	10 – Office (Phones, Chairs), Certificates (Appreciation, SSL)	11 – Capsules (Coffee)
12 – Apparel (Shirts)	11 – Capsules (Tea, Coffee)	12 – Screen (Privacy, Protector)
13 – Bowling	12 – Rubber stamp	13 – PhoneX
14 – Printer	13 – Book	14 – H Tags
15 – Transparent doc box, Tin deposit (Biscuit)	14 – Screen (Privacy, Protector)	
16 – ChipsM (Biscuit)	15 – Nil	15 – Memory (Stick, Built-in)
17 – Capsules (Coffee)	16 – Access (Rooftop)	16 – Nil
18 - Nil	17 - TabletX	17 – Rubber stamp
19 - Book	18 – Christmas decor	18 – Christmas decor
20 - Screen (Privacy, Protector)	19 – Fruits	19 – Fruit
	20 – Transparent doc box	20 – Net ID
	21 – Batteries	21 – Book, Biscuit, Voucher, BTH, Tape, Battery, Juice
	22 – Apple (Fruit, drink)	22 – Apparel (Shirts)
	23 – X developer program	23 – X developer program
	24 – Grocery	24 – Bowling
		25 – Printer
		26 – Grocery
		27 – ChipsM (Biscuit)
		28 – Tin deposit (Biscuit)
		29 – Nil

Cluster for k=25 was selected for further analysis since it covered most of the keywords generated from the different 'k' values. Using the excel file generated, the results of the clustering is combined with other data fields i.e. amount spent, supplier and buyer details for further analysis. A closer examination of the clusters revealed that for most of the clusters, they were not perfect i.e. not all similar goods and services were grouped together by the clustering process. For example, in Cluster 0 shown in Table 3, the clustering is probably based on the occurrences of the words "Certificate" but these items are distinct, it consist of ET Certificates, HT certification and Electrical Certification.

TABLE III. ITEM IN CLUSTER_0

	Goods/Services	***Amount($)***	***Year***	***Cluster***
103	ET Certificate management Services	$1,073.00	2011	Cluster_0
118	To service/recalibrate HT (C/W In-House Calibration Certificate)	$570.00	2011	Cluster_0
145	Color paper for certificate of appreciation	$5.85	2011	Cluster_0
272	To provide LEW service for the certification on the DC cable	$180.00	2011	Cluster_0
399	ET Certificate Management Services	$1,614.00	2011	Cluster_0
446	ET Certificate Management Services	$600.00	2012	Cluster_0
556	Electrical Certification for installation of ICT equipment	$700.00	2012	Cluster_0
557	Electrical Certification for installation of ICT equipment	$700.00	2012	Cluster_0
558	Electrical Certification for installation of ICT equipment	$700.00	2012	Cluster_0
559	Electrical Certification for installation of ICT equipment	$700.00	2012	Cluster_0

The clusters would have to undergo further verification by a keyword search of the description field based on the keywords identified. This step is done manually to determine the final clusters.

The identified clusters, types of good or service, number of suppliers, frequency of purchases (by Year), teams which made the purchases and amount spent are summarized in Table 4. For goods and services where total amount spent is more than S$3,000, they present possible opportunities for the aggregation of purchases via OQs to achieve economies of scale or improve administrative efficiency in purchasing when the product or service is required during the contractual period (via establishing an ETC). Goods or services provided by a single supplier made every year such as X developer program, ET Certificates, Security Holographic Stickers and HT services and accessories are such examples. It is interesting to note that for majority of the transactions, purchase made for a specific good or service was by a single team, contrary to the earlier assumption that aggregation of purchases could be made across teams.

Approach 2

The nature of SVP meant that numerous purchases could be made from one supplier, either within a single year or across different years. The next step is to investigate suppliers from whom purchases were made from them every year. These suppliers, the frequency of the transactions with them, amount spent and the types of goods and services are summarized in Table 5. It is observed that most of the goods and services identified in Approach 2 were all present in Approach 1 except for two main ones, “Standby Technician Support for Video Conferencing System” and “Rental of Cherry Picker”. This is probably because each appeared as one single transaction only in most of the years; hence the term occurrence in the Text Mining analysis was low i.e. it did not feature significantly in the generated Word List or Centroid Cluster Model. However, these observations were noteworthy as they were repeated purchases of relatively significant values in the context of SVP (averaging between S$2,500 and S$2,800) in most years.

B. *OQ*

Approach 1

Similar the approach used in SVP, irrelevant words were removed from the wordlist to improve the result of clustering. The cleaned-up wordlist, sorted in terms of total and document occurrences, now presented words such as licenses, servers, network, CCTV, video, audio, anti-virus, TabletX etc. with highest occurrences.

In determining ‘k’ i.e. the number of clusters for k-mean clustering, the “rule of thumb” is used. For 267 records, k=12 was derived. Similarly for verification and comparison purposes, additional runs were also made for k=20 and k=30.

TABLE IV. SVP: RESULT OF APPROACH 1

Cluster	*Good/ Service*	*No. of suppliers*	*Freq (Year)*	*Project Team (s)*	*Total amount*
1	PhoneX	Multiple	11-14	A	$28,171
2	Printer and accessories (Toners and cartridges)	Multiple	11-14	A	$23,621
3	Fruits & juices	Multiple	11-14	B	$12,876
4	Vouchers	Multiple	11-14	B	$8,850
5	Privacy Screen filters and Screen Protectors	Multiple	11-14	A	$7,829
6	X developer program	Single	13-14	C	$7,452
7	ET Certificates	Single	11-14	C	$6,953
8	Security Holographic Stickers	Single	11-14	D	$5,406
9	Biscuits	Multiple	11-14	B	$5,316
10	Books	Multiple	11-14	G	$5,209
11	CCTV	Single	11-13	E	$4,428
12	Service/recalibrate HT & accessories	Single	11-14	D	$3,613
13	Bowling	Single	11-14	F	$2,823
14	NetID	Single	11-13	C	$1,673
15	Rubber Stamp	Single	11-14	B	$1,047
16	Christmas decor	Single	11,13-14	B	$600
17	Capsules (Coffee and Tea)	Multiple	13-14	B	$568

TABLE V. SVP: RESULT OF APPROACH 2

Supplier	*2011*	*2012*	*2013*	*2014*	*Total transactions*	*2011*	*2012*	*2013*	*2014*	*Total spent*	*Remarks*
Buy & claim	71	19	24	34	148	$14,716	$3,448	$3,935	$13,862	$35,961	Books, festive decorations, X developer program, PhoneX/TabletX accessories
1	18	36	29	14	97	$5,427	$8,030	$10,573	$5,199	$29,229	PhoneX
2	2	9	8	1	20	$20	$732	$275	$20	$1,047	Rubber stamps
3	6	4	4	3	17	$1,084	$640	$1,465	$2,460	$5,649	Fruits
4	6	3	4	1	14	$3,288	$938	$2,695	$2,216	$9,137	ET Certificate management Services
5	4	3	3	3	13	$2,025	$1,264	$1,264	$1,124	$5,678	Service/recalibrate HT & accessories
6	5	1	2	3	11	$5,093	$2,200	$1,885	$5,390	$14,567	Printers and accessories
7	2	4	1	2	9	$2,800	$6,947	$2,600	$5,138	$17,485	Rental of Cherry Picker, CCTV, rectification of cameras, FCV repair
8	2	1	3	1	7	$463	$523	$627	$1,270	$2,883	Bowling
9	3	1	2	1	7	$3,984	$307	$260	$1,400	$5,950	Hard disk, scanner, laptop, ink cartridge, network switch
10	1	3	1	1	6	$630	$5,276	$2,850	$2,850	$11,606	Standby Technician Support for Video Conferencing System
11	1	2	2	1	6	$180	$350	$330	$150	$1,010	Trophies for bowling
12	2	1	1	1	5	$1,500	$760	$1,626	$1,520	$5,406	Security Holographic stickers

The Centroid Cluster Model in Table 6 shows the results, interpreted based on the term frequency of the keywords generated from each cluster (for k= 12, 20 and 30). The results at k=20 and k=30 seem to indicate that the keywords in the majority of the clusters were recurring e.g. licenses, servers, TabletX etc., indicating a strong presence, albeit further breakdown of each into smaller clusters e.g. licenses appeared in Cluster 0,15, 24 and 29 at k=30. New keywords which appeared in the clusters consist of fewer items.

Cluster for k=30 was selected as a basis for further analysis since it covered most of the words generated. The further breakdown of the licenses into smaller clusters was also useful. After the combination of the results of the clustering with other data fields i.e. amount spent, supplier and buyer details for further analysis, a closer examination of the clusters revealed that for most of the clusters, they were not perfect i.e. not all similar goods and services were grouped together by the clustering process – similar to what was observed in SVP.

The identified clusters, types of good or service, number of suppliers, frequency of purchases (by year), teams which made the purchases and amount spent are summarized in Table 7. Notwithstanding the fact that Organization X might have already made efforts to consolidate these purchases (some of these purchases did not occur in 2014), for goods and services where total amount spent is more than S$70,000, they present possible opportunities for the aggregation of purchases via OT to achieve economies of scale. It is interesting to note that for majority of the transactions, purchase made for a specific good or service was by a single team, contrary to the earlier assumption that aggregation of purchases could be made across teams. Opportunities for aggregation of purchases across teams include anti-virus licenses, servers, rental of network equipment, maintenance of Video Conferencing System and maintenance of UPS. It was also observed that some of the items that are bought under OQ were also bought under SVP e.g. printer, Hard Disk and TabletX.

Approach 2

Unlike the SVP, goods and services purchased via an OQ could last for a contractual period beyond one year due to its higher value, for instance, putting in place an ETC where the department could procure directly from the appointed supplier(s) when the product or service is required during the contractual period. Therefore, identifying only suppliers where purchases were made from them every year would not be sufficient. Instead, the cumulative value of the contracts awarded over the period of the four years would be used as a measure to identify suppliers. As a benchmark, a cumulative value of $70,000 is used. These suppliers, sorted by total cumulative amount spent and total number of transactions is shown in Table 8. The remarks column indicated the predominant type of good or service purchased.

It was observed that a specific type of good or service is usually provided by a dominant supplier, more for services than for goods, e.g. supply and maintenance of IT systems. However, it was interesting to note that the different types of software licenses required by different teams are provided predominantly by a single supplier, Supplier 13.

TABLE VI. OQ: CENTROID CLUSTER MODEL FOR DIFFERENT "K" VALUE

k=12	*k=20*	*k=30*
0 – Network (Device E, switch and router, printer)	0 – Electronic (degausser, map, System S, System E, storage media, shredder)	0 – Licenses (iP, S, CS)
1 – Audio, accessories, TabletX	1 – AC System, New W Sys	1 – Scanners (Fingerprint, Document, Barcode)
2 – Electronic (degausser, map, System S, System E, storage media, shredder)	2 – Security (Demand Aggregate)	2 – New W Sys
3 – System (IP, VW, CT)	3 – Leased Line Circuit	3 – Licenses (I)
4 – Security (Demand Aggregate)	4 – Audio-visual, accessories	4 – Nil
5 – Leased Line Circuit	5 – Licenses (I)	5 – Network Device E
6 – Interface (fibre, cards, ports)	6 – Printers	6 – Mobile, accessories, IDN
7 – Servers	7 – System A, Thumbprint	7 – System E, man-days, security
8 – ISO certification, Data Centre cleaning	8 – System E, mandays	8 – Data Centre Cleaning
9 – CCTV	9 – TabletX	9 – AC System
10 – Licenses (anti-virus, SY, SW, O, iP)	10 – Licenses (anti-virus, SY, SW, O, iP)	10 – ISO certification
11 - Video	11 – Video Conferencing, Camera	11 – TabletX
	12 – Network (Device E, switch and router)	12 - mandays, power points
	13 – System (IP, VW, CT)	13 – Servers
	14 – Interface (fibre, cards, ports)	14 – System P
	15 – S alert	15 – Licenses (O, SY, SW)
	16 – Servers	16 – Audio accessories, Video (Conferencing, Wall)
	17 – ISO certification, Data Centre	17 – Network (switch and router, printer, cards)
	18 – CCTV	18 – Malware Analysis
	19 – Room (Partition, cable)	19 – Leased Line Circuit
		20 – NGNBN
		21 – Management System (K, V, C)
		22 – Anti-Virus
		23 – CCTV
		24 – B O licenses
		25 – RP, SSS
		26 – Electronic (map, System S, System E, storage media, shredder)
		27 – Audio-Visual
		28 – Hard Disk (Desktop, Notebook)
		29 – Licenses (O)

Another noteworthy observation was the maintenance of several IT systems by a single supplier, Supplier 14. This provided evidence that there are opportunities for different goods and services to be aggregated to achieve better pricing, contrary to the earlier assumption that aggregation of purchases could be made for similar goods and services only.

Approach 1 was more effective at identifying purchase of common goods and services across different suppliers e.g. for provision of Anti-Virus licenses, Approach 1 identified five different suppliers while Approach 2 identified one only (without going through the entire list of suppliers). Approach 2 was effective at identifying the dominant suppliers and provided insights such as consolidation of different goods and services which would have been missed out using Approach 1.

C. ETC

Approach 1

Similar to previous approach, irrelevant words were removed from the wordlist to improve the result of clustering. The cleaned-up wordlist sorted in terms of total and document occurrences, now presented words such as licenses, engineers, engineering and project management, UPS, License I, License O, switch etc. with highest occurrences. This gave a clearer indication of the purchases made and the keywords (items) that should watch out for in subsequent steps of the analysis.

In determining ‘k’ i.e. the number of clusters for k-mean clustering, the “rule of thumb” is used. For 118 records, k=8 is used. For verification and comparison purposes, additional runs were also made for k=12 and k=16.

By examining the Centroid Cluster Model in Table 9, the results, interpreted based on the term frequency of the keywords generated from each cluster (for k= 8, 12 and 16), are shown in. The results after runs at k=12 and 16 seem to indicate that the keywords in the majority of the clusters were recurring e.g. licenses, servers, racks, Engineering and Project Management etc., albeit further breakdown of each into smaller clusters e.g. licenses appeared in Cluster 7,8 and 15. At k=16, new keywords which appeared in the clusters consist of fewer items, with quite a few clusters with zero or one item.

Cluster for k=12 was selected as a basis for further analysis since it covered most of the words generated from the different ‘k’ values. After the combination of the results of the clustering with other data fields i.e. amount spent, supplier and buyer details for further analysis, a closer examination of the clusters revealed that for most of the clusters, they were not perfect i.e. not all similar goods and services were grouped together by the clustering process – similar to what was observed in SVP and OQ.

TABLE VII. OQ: RESULT OF APPROACH 1

Cluster	***Good/ Service***	***No. of suppliers***	***Freq (FY)***	***Project Team(s)***	***Total amount***
1	Purchase of O Licenses	Multiple	11-13	C	$412,582
2	Purchase of Anti-Virus Licenses	Multiple	11-14	A,D	$290,627
3	Rental of Audio Visual Equipment	Multiple	11-14	A	$252,950
4	Purchase/Maintenance of Biometric Fingerprint Scanners	Single	11-12	C	$249,350
5	Maintenance of Security Equipment	Multiple	13-14	A	$247,282
6	Maintenance of servers	Multiple	11-14	C,D,H,I	$246,112
7	Repair/Supply of Audio Accessories	Multiple	11-12,14	J	$231,859
8	Rental/Purchase of Network equipment/Device E	Multiple	11-13	D,K	$223,154
9	Maintenance of RP and SSS System	Single	11-12	C	$219,112
10	Installation of CCTV System	Single	11-13	A	$211,495
11	Maintenance and Service Request man-days for System E	Multiple	11,13	L	$209,440
12	Purchase of TabletX	Multiple	11-13	A	$196,148
13	Maintenance of IP System	Single	12-14	D	$191,680
14	Maintenance of New W System	Single	11,13	C	$182,140
15	Maintenance of Video Conferencing System	Multiple	11-13	A,D	$181,234
16	System P relocation	Multiple	11,13	E	$158,070
17	Purchase of SY Licenses	Multiple	11-14	D	$155,295
18	AC System	Multiple	11-14	C	$145,230
19	Purchase of Notebook for CCTV Clients	Multiple	12,14	A	$127,030
20	Subscription of Leased Line Circuit	Multiple	12	E	$112,738
21	Purchase of Hard Disk	Multiple	12-14	A	$101,335
22	Maintenance for System A	Single	11-12	C	$98,484
23	Installation of CCTV System	Single	11-13	A	$76,750
24	Maintenance of UPS	Multiple	12,13	D,E	$64,720
25	Data Centre Cleaning	Multiple	11,14	D	$62,540
26	Provision of ISO Consultancy and IQA Services	Multiple	11-14	I	$35,600
27	SA CS 2011 Licenses	Single	12-13	D	$27,276
28	Renewal of I Licenses	Multiple	11-13	D	$15,108
29	Purchase of printers	Multiple	12,14	A	$11,974

TABLE VIII. OQ: RESULT OF APPROACH 2

Supplier	*2011*	*2012*	*2013*	*2014*	*Total transactions*	*2011*	*2012*	*2013*	*2014*	*Total spent*	*Remarks*
13	13	14	7	4	38	$618,208	$528,615	$190,822	$132,341	$1,469,986	Licenses O, SY, anti-virus, TabletX, notebooks
14	3	3	0	1	7	$179,250	$187,200	-	$25,560	$392,010	System C, V, P/I, SSMS, HR, CRC
15	4	3	1	0	8	$172,879	$137,835	$10,790	-	$321,504	Biometric Fingerprint scanners, System ADS
16	2	1	2	0	5	$137,920	$55,480	$99,340	-	$292,740	System E, M infrastructure
17	1	1	3	0	5	$68,883	$59,584	$158,631	-	$287,098	O licenses, System R
7	2	2	3	0	7	$45,694	$86,165	$126,107	-	$257,966	System U, Repair of audio accessories
18	2	1	1	0	4	$156,415	$45,450	$27,950	-	$229,815	Electronic map, System T
19	2	2	0	0	4	$112,184	$106,928	-	-	$219,112	RP, SSS
20	1	1	1	0	3	$70,715	$68,590	$72,190	-	$211,495	CCTV
21	1	1	0	1	3	$68,000	$61,865	-	$79,800	$209,665	Audio accessories
22	0	3	0	0	3	-	$197,804	-	-	$197,804	IP Sys, Malware Analysis
23	1	1	1	0	3	$67,200	$55,660	$62,450	-	$185,310	Rental of audio-visual equipment
24	3	0	1	0	4	$112,580	-	$69,560	-	$182,140	New W Sys
25	0	2	1	0	3	-	$82,376	$79,876	-	$162,252	System CA and CK
26	2	1	0	0	3	$84,486	$74,000	-	-	$158,486	System M, System H, System AG
27	2	0	2	0	4	$73,925	-	$76,545	-	$150,470	System P
28	0	0	3	0	3	-	-	$143,732	-	$143,732	NGNBN
10	1	1	0	0	2	$71,820	$70,434	-	-	$142,254	Video Conferencing Sys
29	0	0	2	0	2	-	-	$138,425	-	$138,425	System W, System BD, System X
30,29,31 ,20,32	0	0	1	1	2	-	-	$67,621	$66,198	$133,819	System W, System BD, System X System N, System PA, System UV
33	2	0	0	0	2	$128,700	-	-	-	$128,700	eL and eT Sys
34	1	0	2	0	3	$27,200	-	$96,600	-	$123,800	CCTV
35	1	0	1	1	3	$40,750	-	$36,000	$46,800	$123,550	AC Sys
36	1	0	0	1	2	$60,000	-	-	$62,290	$122,290	IDN
37	3	2	0	2	7	$61,775	$29,292	-	$26,120	$117,187	Servers
38	1	0	1	0	2	$42,913	-	$69,496	-	$112,409	System MT, VMS
39	0	1	2	0	3	-	$24,678	$79,030	-	$103,708	Network switches, routers

TABLE IX. ETC: CENTROID CLUSTER MODEL FOR DIFFERENT "K" VALUE

k=8	*k=12*	*k=16*
0 – Servers, Racks	0 – Network Infrastructure	0 – Engineering and Project Management
1 – Engineering and Project Management	1 – Office Hour	1 – Nil
2 – Audit, Switch, Network	2 – Audit , SAN	2 – Audit, Switch, Network
3 – SAN, Network and System Engineer	3 – Licenses (I, H, SM)	3 – SOE
4 – Network	4 – O Licenses	4 – Network equipment
5 – UPS	5 – Network equipment	5 – Desk side Engineer
6 – Office Hour	6 – UPS	
	6 – Network Attach Storage	
7 – Licenses	7 – Engineering and Project Management	7 – I licenses
	8 – COR	8 – O licenses
	9 – Nil	9 – Server, Racks
	10 - Servers, Racks	10 - UPS
	11 – Licenses (SY)	11 – SAN, Network and System Engineer
		12 – SOE
		13 – COR
		14 – Office Hour
		15 – Licenses (SY)

Similar verification and manipulation of the clusters would have to be adopted similar to what had been prescribed previously. Goods and services available under the ETCs are in general already aggregated. A better understanding of how they were purchased will improve procurement planning and possibly reduce the administrative efforts involved in issuing CFQs e.g. by combining CFQs from teams or multiple year CFQs. However, it is also possible to achieve economies of scale as suppliers are known to offer better pricing than those stated in the ETC.

The identified clusters, types of good or service, number of suppliers, frequency of purchases (by year), teams which made the purchases and amount spent are summarized in Table 10. These goods and services present possible opportunities for the aggregation of purchases. It is interesting to note that items which were bought under ETC were also bought under OQ e.g. O and SY licenses. However, it is beyond the scope of this project to investigate the possible reasons for this. There are also possible evidence to suggest efforts made in the aggregation of goods and services e.g. in the OQs, there were rental and purchase of Network equipment made from 2011-13 but this stopped in 2014 and similar purchases were made under ETC in 2014. Another observation is that while there were outright purchases for certain goods in ETCs, the maintenance of similar goods was procured under OQ e.g. servers, UPS. However, there is no evidence to prove that these were the same equipment bought under the ETCs which were later maintained under the contracts established via OQs.

TABLE X. ETC: RESULT OF APPROACH 1

Cluster	*Good/ Service*	*No. of suppliers*	*Freq (Year)*	*Project team(s)*	*Total amount*
1	Purchase of Servers	Multiple	11-14	D,E,I,L	$7,383,583
2	Provision of Engineering & Project Management Services	Multiple	11-14	C,H,JK,L,M	$6,950,335
3	O Licenses	Multiple	12-14	C,D	$4,240,144
4	SAN-related purchases	Single	12-14	D	$3,514,283
5	Provision of Network and System Engineers	Multiple	11-13	D	$1,190,800
6	Provision of Office Hours/after Office Hours Support Services	Single	12-13	D	$1,085,821
7	I Licenses	Multiple	11-14	C,J,L	$891,940
8	UPS	Single	11-13	D	$822,714
9	SY Licenses	Multiple	11-13	C,D,L	$649,135
10	Purchase of racks (servers and network equipment)	Multiple	11-13	D	$412,078
11	Provision of IT Security and Audit Services	Multiple	11-13	I	$156,900
12	Purchase of Network Equipment	Single	14	D	$31,105

For majority of the transactions, purchase made for a specific good or service was by a single team, contrary to the earlier assumption that aggregation of purchases could be made across teams. Opportunities for aggregation of purchases across teams to include purchases of licenses (O, I, and SY), servers, Engineering and Project Management Services.

Approach 2

For ETC, the identification of a supplier based on its value and volume of transaction is less relevant in the analysis because it would already have been known upfront at the point where the ETC was established, the goods and services it is offering even before the CFQ was issued. Hence, Approach 2 will not be applicable in the analysis.

VI. CONCLUSION

Managerial perspective

The findings suggest that opportunities exist for Organization X to aggregate common goods and services among the purchases made under SVP, OQ and ETC (Table 4, 7 & 10). The analysis further suggests that these opportunities were more prevalent in purchases made by individual project teams rather than across multiple project teams. However, it must be acknowledged that in reality, circumstances such as different timelines/deadlines of projects, unanticipated changes and dynamic requirements from stakeholders make such procurement planning in the short term very challenging. These could be the most likely reasons for the separate transactions for similar goods and services detected. There were some indications suggesting that Organization X has undertaken efforts to consolidate frequent purchases e.g. for recurring purchases of O licenses, rental of network equipment, no such transactions appeared in the 2014 OQ list of transactions while appearing in the 2014 list of ETC transactions.

The results obtained from the analysis should increase Organization X's awareness and improve its visibility of the goods and services it has been procuring in recent years. It is recommended that these lists of identified goods and services to be shared with the different project teams to facilitate long-term procurement planning within teams and better synergy in coordinating procurement efforts across teams.

Research perspective

The use of advanced data mining techniques such as text mining and cluster analysis complements the OLAP tools commonly used in analyzing procurement data. It addresses the inadequacy of OLAP tools in generating new information from textual data. However, it does have its shortfalls – the nature of the natural language texts contains ambiguities and it is still difficult to analyze the semantics and to interpret meaning if the keywords e.g. a specific cluster might have both a purchase of an Apple computer and apple (fruit) grouped together. Keywords in the description fields which are spelt incorrectly and captured are useless, as the saying goes "garbage in, garbage out". Its output is not an end in itself. The process is most rewarding when the data text mining generates can be further analyzed by a subject matter expert, who can bring additional knowledge for a more complete picture. Text mining can create new relationships and hypotheses for further exploration.

The cluster analysis, while useful in grouping most of the transactions and offered a general overview of the types of goods and services purchased, required further manual manipulation of the clusters to derive more accurate output in order to derive meaningful insights.

Data fields such as the account codes could be used to categorize the records at a broad level before applying the text mining and clustering analysis, improving the semantics of the keywords extracted from the description fields. More clustering algorithms could be applied to compare the accuracy of the clustering output. In this study, only K-means clustering was used.

Further improvements

To improve the output from the text mining analysis, more efforts could be made to experiment with the various text processing algorithms in the RapidMiner's text mining extension. The better the quality of the output i.e. keywords identified, the more accurate the clustering of the records.

REFERENCES

[1] Lee, "Procurement lapses in govt agencies raise concern," Today newspaper article, April 2, 2014, link http://www.todayonline.com/singapore/procurement-lapses-govt-agencies-raise-concern

[2] G. Kemp, "Fighting public sector fraud in the 21st century. Computer Fraud & Security," vol. 2010, issue 11, pp 16-18, Nov 2010.

[3] A. Byrne, "Government procurement in Western Australia – beyond compliance," Keeping Good Companies, vol. 65, issue 7, pp 394-399, Aug 2013.

[4] B. Chae and D.L. Olson, "Business Analytics for Supply Chain: A Dynamic-Capabilities Framework," International Journal of Information Technology & Decision Making, vol. 12, no. 1, pp 9-26, 2013.

[5] National Fraud Authority, "Procurement Fraud in the Public Sector," Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/118460/procurement-fraud-public-sector.pdf, Oct 2011.

[6] OECD library, "OECD e-Government Studies: Egypt 2013," Paris. OECD Publishing DOI: http://dx.doi.org/10.1787/9789264178786-en, 2013.

[7] M. Kantardzic, "Data Mining: Concepts, Models, Methods, and Algorithms, 2nd Ed.," Wiley-IEEE Press, 2011.

[8] L.S. Chia and C.H. Leo, "Business Intelligence in Government Procurement," Retrieved from http://www.dsta.gov.sg/docs/publications-documents/business-intelligence-in-government-procurement.pdf?sfvrsn=0, 2009.

[9] G. Miner, J. Elder, A. Fast, T. Hill, R. Nisbet and D. Delen, "Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications," Academic Press, Jan 2012

[10] K.V. Mardia, J.T. Kent and J.M. Bibby, "Multivariate Analysis (Probability and Mathematical Statistics) 1st Edition," Academic Press, 1979.

Increase Efficiency of SURF using RGB Color Space

M. Wafy
Faculty of Computers and Information, Helwan
University, Cairo, Egypt

A. M. M. Madbouly
Mathematics Department,
Faculty of Sciences, Helwan University,
Cairo, Egypt

***Abstract*—SURF is one of the most robust local invariant feature descriptors. SURF is implemented mainly for gray images. However, color presents important information in the object description and matching tasks as it clearly in the human vision system. Many objects can be unmatched if their color contents are ignored. To overcome this drawback this paper proposed a method CSURF (Color SURF) that combines features of Red, Green and blue layers to detect color objects. It edits matched process of SURF to be more efficient with color space. Experimental results show that CSURF is more precious than traditional SURF and CSURF invariant to RGB color space**

***Index Terms*—SURF; Features; Local Features; Color Space**

I. Introduction

Finding similarity between different images is representing a challenge problem in computer vision application. For instance, it uses in pattern recognition, object tracking, image retrieval, etc. Searching between images can be affected by three problems, scale, rotation and general transformations in gray image. SURF (Speeded Up Robust Features) algorithm was adopted by Bay et al. [1]. Firstly detects corner points under different conditions. Secondly, construct descriptor based on detecting corner points and its neighbors. The final match between descriptor using a distance function. SURF [1] is the most robust descriptor among the other local invariant feature descriptors. It is invariant to scale and rotation transformation with respect to different to geometric changes. SURF has a main drawback it does not use color information. In fact, SURF has been constructed essentially for gray images, ignoring totally the information found in the color space as most of feature extractors. Neglecting color information leads to lose an important source of distinction between images as showed in following figure.

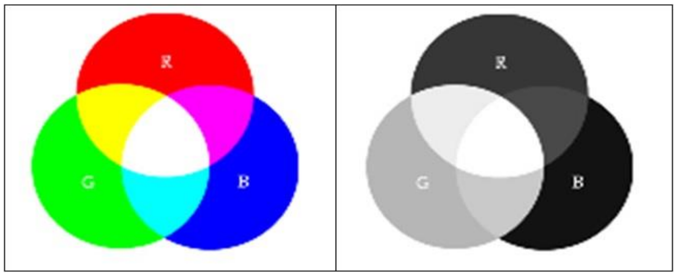


Fig. 1. Shows importance of color component

Figure 1 represents color and gray image of RGB and CMY the red and blue colors were converted to a same level of gray. The Figure 1 illustrates how studying color image cannot be done via converted it to the gray level.

This paper introduces a method that helps SURF to match color images with respect to its color in RGB space.

The rest of this paper is organized as follows: section two

Introduce survey of other color descriptors, section three represent methodology of the proposed color SURF (CSUR), section four represents experimental results, section five introduces conclusion.

II. Color Descriptors

Color information plays an important role in the computer vision application. As clear human vision system can distinctly many color values. There are several attempts that have been used for color information inside descriptor. For example, Abdel-Hakim and Farag [2], proposed a colored local invariant feature descriptor called CSIFT(Color SIFT). The new descriptor builds SIFT(Scale Invariant Feature Transform) on invariant color space. It involves three main steps: interest point detections, build descriptors and descriptor matching.

Firstly, selection of highly informative points as interest points. Secondly, descriptor of interest point is built to describe a local region around interest points. Thirdly, matching stage that decides if interest point belongs to object or not.

CSIFT is not efficient for photometrical variety of images. Ancuti and Bekaert[3] ,proposed SIFT-CCH(SIFT Color Co-occurrence Histograms) that merges the traditional SIFT with Color Co-occurrence Histograms computed from the Nrgb color space. Their method achieves as same as SIFT in the detection step, but establishes one dimension to the descriptor. Thus, features are illustrated by a two element vectors that merges the SIFT and the CCH descriptor vectors. The main limitation of such an approach is time of computation during feature matching due to the extra 128 elements added to the descriptor vector.

Cui et al.[4], proposed color descriptor called PC-SIFT (perception-based color SIFT) that invariant to illumination variation., this method builds SIFT based on new color space that formulated at [5], [6] to signify image as a replacement for using its gray scale values.

Ai et al.[7]proposed an adaptive Color Independent Component based on SIFT descriptor called (CIC-SIFT) for image classification problems. Their proposed algorithm can be summarized into three steps, firstly learning step, a transformation matrix is learned for each category by using Independent Component Analysis. Secondly color transformation step: original image components are transformed into three independent components by using the adaptive transformation matrix. Finally extraction of feature: the color independent components are used to compute CIC-SIFT descriptors.

Fan et al. [8] proposed color invariant descriptor based on SURF. It merges local kernel color histograms and Haar

Wavelet responses to build the feature vector. That means the descriptor is a two element vector. Matching process composed of two steps, firstly, SURF descriptor is matched, then unmatched points are calculated between their local kernel histograms using the distance function called Bhattacharyya. In this paper, we will introduce a novel a Color SURF descriptor called CSURF. It's based on RGB color space

III. Color SURF Descriptor

This section will illustrate our new Color SURF descriptor (CSURF). The method will divide into two main steps, firstly detect and extract features of entering the color image. Secondly, matching between two descriptor color images.

A. Detect and extract feature

CSURF firstly read color image and construct three images from it using RGB color space, where I1, I2 and I3 represent constructed images from Red, Green and Blue layers respectively. Secondly CSURF will find corner points of images I1, I2, I3 using traditional SURF to produce corner points P1, P2 and P3 of images I1, I2 and I3. Thirdly CSURF will extract features of constructing image using SURF descriptors that produce vpts1, vpts2 and vpts3 of images I1, I2 and I3 respectively.

Finally CSURF will merge the detected corner points in one vector P= [p1; p2; p3] and merge extracted features in one vector called vpts= [vpts1; vpts2; vpts3]. The previous procedure for extracting features using CSURF is summarized as figure 2.

B. Matching features between images

CSURF after extraction corner points and features of entering images will try to match between images with a simple strategy. It will match between extracted features using the traditional SURF descriptor and will divide the features into two classes matched features mf1, mf2 of image1 and image2 and unmatched feature uf1, uf2 of image1 and image2 respectively. CSURF will work on uf1 and uf2 as follows, firstly, it finds unmatched features that have nearly the same color and we will weight it with 0.3.

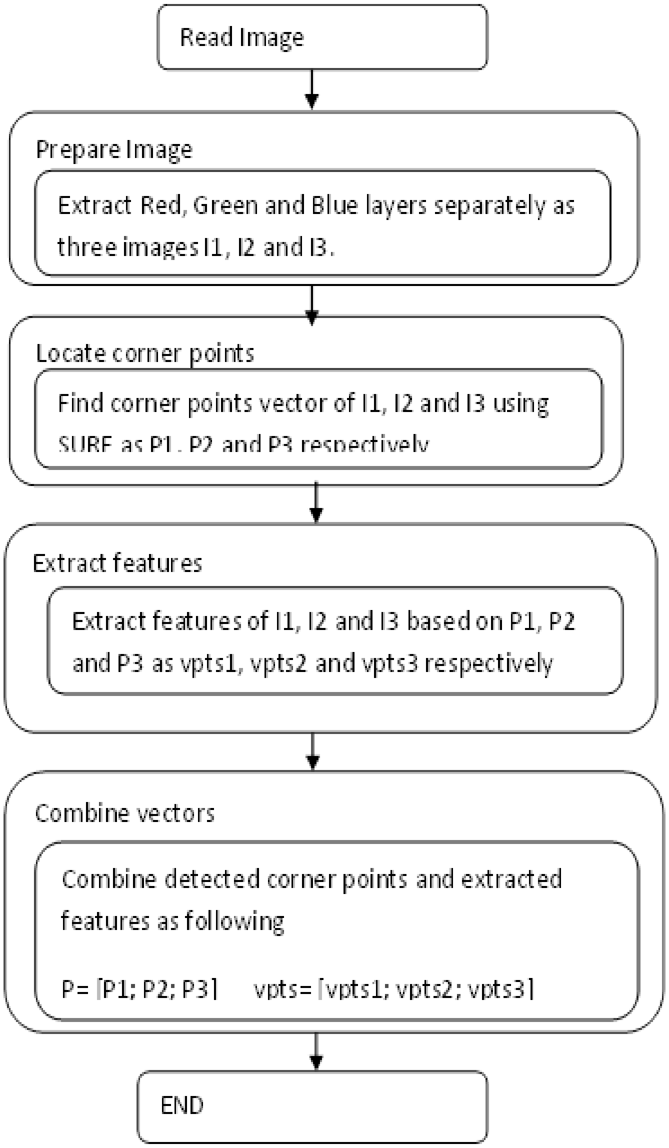

Read Image
Prepare Image
Extract Red, Green and Blue layers separately as three images I1, I2 and I3.
Locate corner points
Find corner points vector of I1, I2 and I3 using SURF as P1 , P2 and P3 respectively
Extract features
Extract features of I1, I2 and I3 based on P1, P2 and P3 as vpts1, vpts2 and vpts3 respectively
Combine vectors
Combine detected corner points and extracted features as following
P= [P1; P2; P3]
vpts= [vpts1; vpts2; vpts3]
END

Fig. 2. Shows how CSUFR extract features

Secondly, our method will try to match unmatched features with lower threshold 0.5 (traditional SURF in this experiment use threshold =0.6) and we will weigh that with 0.7. Finally method will select unmatched features that pass both of the previous two steps and store them at umf1 and umf2. The method will combine features (matched and unmatched features) in one vector called cmatch1 = [mf1; umf1] and cmatch2 = [mf2; umf2].

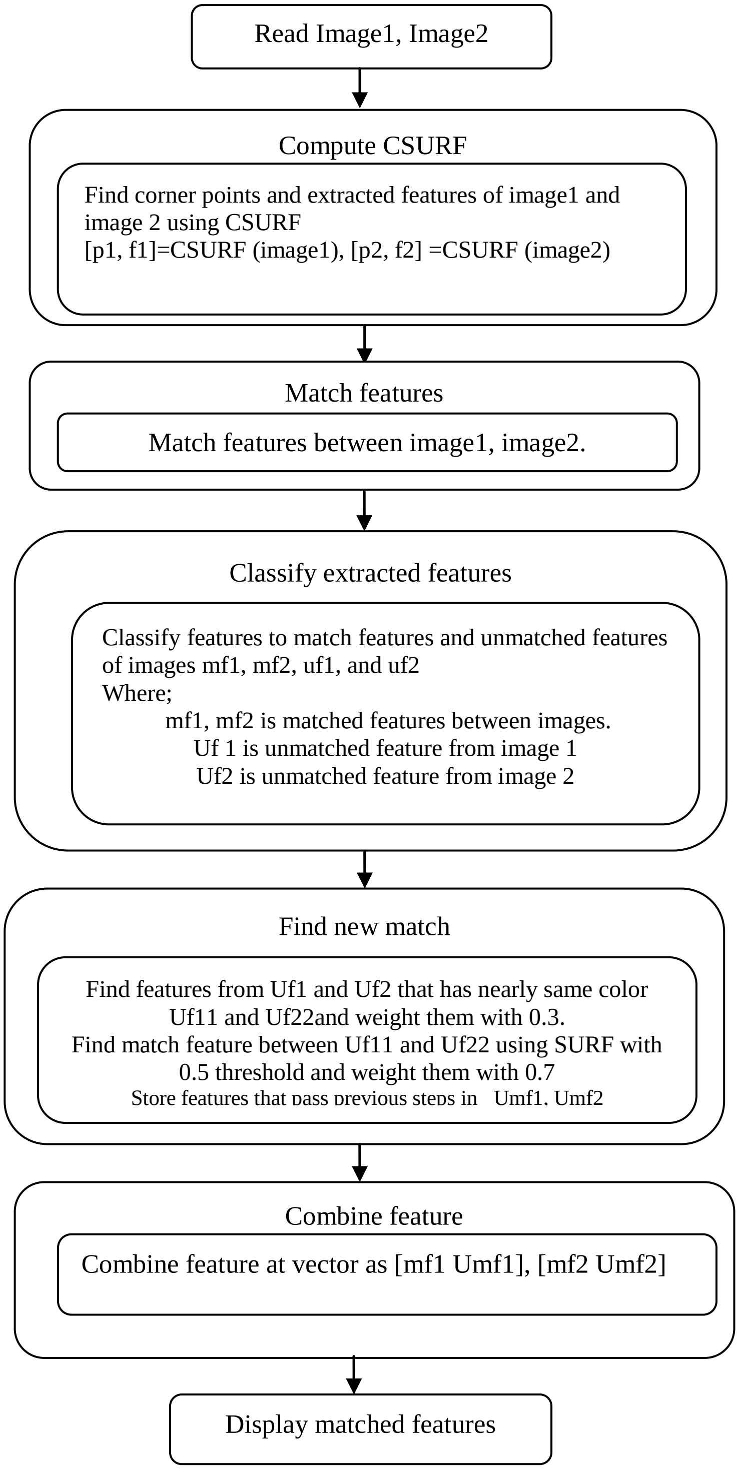

Read Image1, Image2
Compute CSURF
Find corner points and extracted features of image1 and image 2 using CSURF
[p1, f1]=CSURF (image1), [p2, f2] =CSURF (image2)
Match features
Match features between image1, image2.
Classify extracted features
Classify features to match features and unmatched features of images mf1, mf2, uf1, and uf2
Where;
mf1, mf2 is matched features between images.
Uf 1 is unmatched feature from image 1
Uf2 is unmatched feature from image 2
Find new match
Find features from Uf1 and Uf2 that has nearly same color Uf11 and Uf22and weight them with 0.3.
Find match feature between Uf11 and Uf22 using SURF with 0.5 threshold and weight them with 0.7
Store features that pass previous steps in Umf1, Umf2
Combine feature
Combine feature at vector as [mf1 Umf1], [mf2 Umf2]
Display matched features

Fig. 3. The match between extracted features using CSURF

I. Experimental Results

In this experiment weuse the"Amsterdam Library of Object Images" (ALO))[9] that is a database of colored images. It contains a large number of images under different illumination directions, illumination colors, and viewing direction. The experiment will work on ten images at different levels of illumination directions, view direction and illumination colors where every image on illumination color have 12 different copies, images that at the illumination direction have 24 copies and every type of view direction have 72 different copies.

Figure 4 shows a sample of images under different illumination color, illumination direction and view direction.

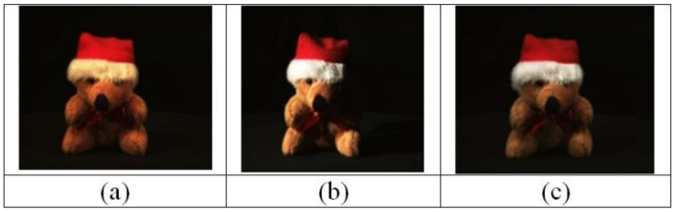


Fig. 4. Sample of ALOI color images at different illumination color, illumination direction and view direction

The method was performed on a laptop Lenovo 3000 C100 with the following configuration: Intel Pentium (R) M processor, (R) 1.73-GHz, 1.73-GHz and 2-GB RAM.

The experiment will work on three types of distortion illumination color, illumination direction and view direction. Due to importance of color component CSURF work on every layer of RGB color space separately. Figure 5 shows the difference of performance of traditional SURF and CSURF on illumination intensity distortion. Figure 5 image(a) and image(b) represent performance of traditional SURF and CSURF respectively.

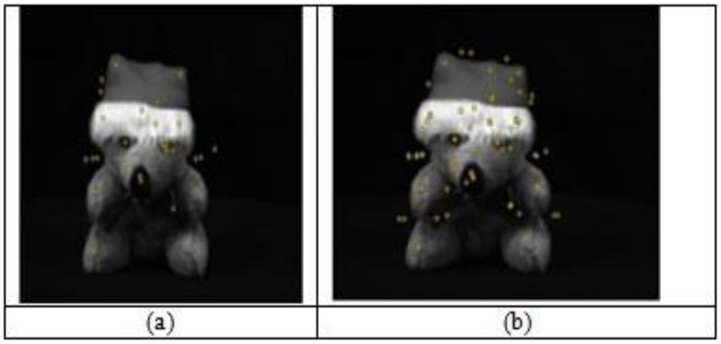


Fig. 5. Illustrate the difference in performance of match between traditional SURF and CSURF

Following figure shows how CSURF surpass traditional SURF in matching between images under illumination direction.

Figure 6 image (b) shows how CSURF matched important points that traditional SURF can't capture as shows in image(a) Important points noted that at points that helped to match a new object.

It is clear from following figure how CSURF results in image(b) have better performance than SURF result in image(a) that under different view of direction.

It is clear from figures 5, 6 and 7 that the performance of CSURF surpasses traditional SURF in number of matched points.

CUSRF detect the number of corner points more than SURF, as showed in figures 8.

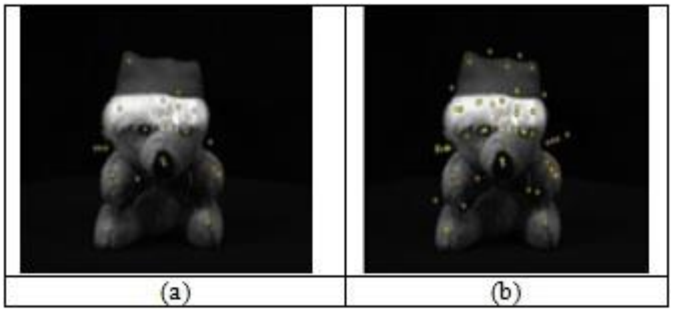


Fig. 6. Explain the difference in performance of match between traditional SURF and CSURF of images under different illumination direction

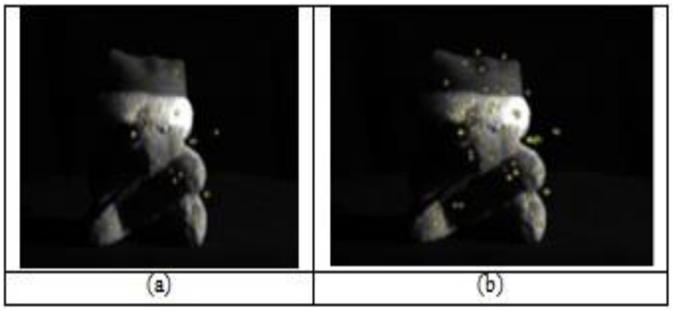


Fig. 7. Shows the difference in performance of match between traditional SURF and CSURF of images under different view direction

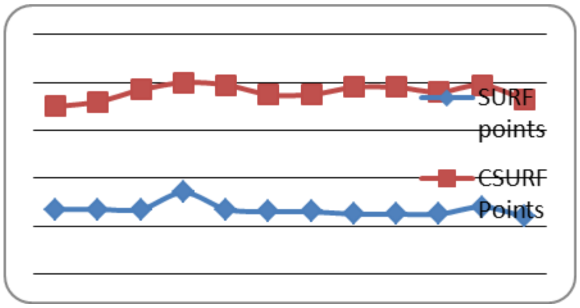


Fig. 8. Presents number of corner points detected by SURF and CSURF

Figure 8 represents the number of detected point using SURF and CSURF, where x-axis represents the level of color illumination of the image and y axis represents the number of detected points.

Most previous methods in color descriptors measure its progress by the number of detected or matched features. They show how their methods surpass traditional methods by numbers (no mentions if these numbers represent vital corner points or not. So experiment, test if CSURF detect interesting corner point tradition cannot detect it. So we will take three images on each distortion type (illumination directions, view direction and illumination colors) each image will have ten levels of its distortion. We will test the importance of the point (if point critical or not that mean it help to catch new object) and if detected by SURF or not using eye observer. The result will represent in following figure.

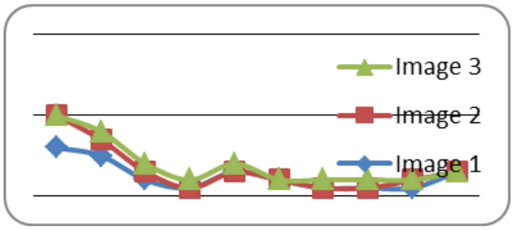


Fig. 9. Number of observed important corner points detected by CSURF at view direction distortion at different levels

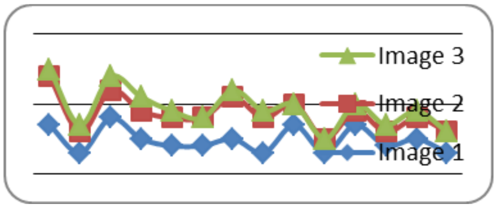


Fig. 10. Shows eye observed important corner point detected by CSURF at illumination direction distortion at different levels

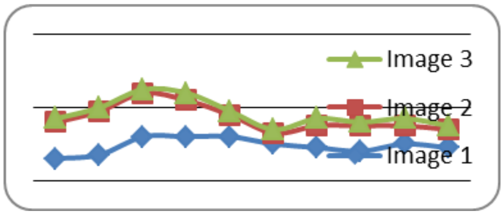
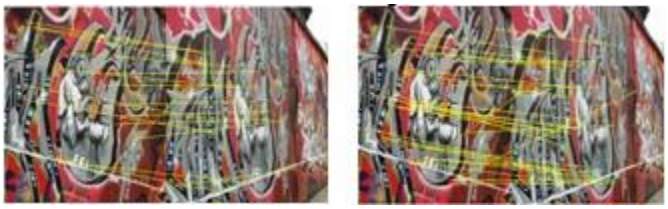


Fig. 11. Shows how CSURF detect d important corner point detected by CSURF at illumination color distortion at different levels

It is clear from the above figures (9,10 and 11), where x-axis represents the level of distortion (view direction, illumination direction and illumination color) and y axis represent the number of important detected points. CSURF can detect the important point that cannot be detected by traditional SURF because CSURF do not neglect color component as SURF did.

Proposed method will compare with [8] by matching two graffiti scene that used in [8]. Fan didn't mention the size of the image so we will compare his progress with SURF and proposed method. He mentions that traditional SURF match 43 features while his color SURF captured 51 features that mean progress 18.6%.

We will use graffiti sense similar in the sense that Fan[8] used. Traditional SURF matched 17 features while the proposed method captured 34 features that mean there is progress in number of matched as showed in figure 12.



(a) Matched SURF features (a) Matched CSURF features

Fig. 12. Comparison of SURF and Color-SURF

II. CONCLUSION

In this paper, we introduced CSURF (Color SURF) as color descriptor that works on RGB color space.

Contrary to many existing methods presented CSURF do not neglect color component due to its importance as it clearly in the human vision system. CSURF accomplished by two main steps, firstly it combined detected corner points and extracted features from Red, Green and Blue layers into two vectors, one for combined features and the other one for combined detected points. Secondly, it edits matched process of SURF to be more efficient with color space. Experimental results demonstrated the high performance of CSURF descriptor against traditional SURF.

REFERENCES

[1] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded up robust features,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2006, vol. 3951 LNCS, pp. 404–417.

[2] A. E. Abdel-Hakim and A. A. Farag, “CSIFT: A SIFT descriptor with color invariant characteristics,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, vol. 2, pp. 1978–1983.

[3] C. Ancuti and P. Bekaert, “SIFT-CCH: Increasing the SIFT distinctness by color co-occurrence histograms,” in ISPA 2007 - Proceedings of the 5th International Symposium on Image and Signal Processing and Analysis, 2007, pp. 130–135.

[4] Y. Cui, A. Pagani, and D. Stricker, “SIFT in perception-based color space,” in Proceedings - International Conference on Image Processing, ICIP, 2010, pp. 3909–3912.

[5] H. Y. Chong, S. J. Gortler, and T. Zickler, “A perception-based color space for illumination-invariant image processing,” ACM Transactions on Graphics, vol. 27, no. 3. p. 1, 2008.

[6] H. S. Fairman, M. H. Brill, and H. Hemmendinger, “How the CIE 1931 color-matching functions were derived from Wright-Guild data,” Color Res. Appl., vol. 22, no. 1, pp. 11–23, 1997.

[7] D. Ai, X. Han, X. Ruan, and Y. W. Chen, “Adaptive color independent components based SIFT descriptors for image classification,” in Proceedings - International Conference on Pattern Recognition, 2010, pp. 2436–2439.

[8] P. Fan, A. Men, M. Chen, and B. Yang, “Color-SURF: A surf descriptor with local kernel color histograms,” in Proceedings of 2009 IEEE International Conference on Network Infrastructure and Digital Content, IEEE IC-NIDC2009, 2009, pp. 726–730.

[9] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, “The Amsterdam Library of Object Images,” Int. J. Comput. Vis., vol. 61, no. 1, pp. 103–112, 2005.

Smart Grid Testbed using SCADA Software and Xbee Wireless Communication

Aryuanto Soetedjo
Dept. of Electrical Engineering
National Institute of Technology
Malang, Indonesia

Abraham Lomi
Dept. of Electrical Engineering
National Institute of Technology
Malang, Indonesia

Yusuf Ismail Nakhoda
Dept. of Electrical Engineering
National Institute of Technology
Malang, Indonesia

***Abstract*—This paper presents the development of Smart Grid testbed using SCADA software and Xbee wireless communication. The proposed testbed combines both the software simulation and the hardware simulation. The Winlog SCADA software is employed to implement the algorithm in the Smart Grid system. To communicate between nodes and the Smart Grid Center, the Xbee wireless communication is employed. The testbed is useful to test and verify the developed algorithms in the Smart Grid system. By using the hardware testbed, the more realistic simulation could be performed. While by using the software testbed, the complex model and algorithm could be implemented easily. The experimental results show that the proposed testbed works properly in simulating the continuous supply algorithm implemented in the Smart Grid system.**

Keywords—SCADA; Smart Grid; wireless communication; Xbee DigiMesh; testbed

I. Introduction

The Smart Grid is a modern technology in the electrical power systems that delivering the electricity in a smart way, where the costumers are considered as the integral part of the system [1]. In the Smart Grid, the electricity and the information flow in two-way to achieve the efficient, clean, safe, reliable, and safe electrical power systems [2].

The Smart Grid integrates the sophisticated technology in electronics, communication, information, and electrical power systems. In general, the Smart Grid is divided into three major systems [2], i.e.: the smart infrastructure system, the smart management system, and the smart protection system. The smart infrastructure system consists of the smart energy subsystem, the smart information subsystem, and the smart communication subsystem. The smart management system deals with the energy efficiency, utility, cost, pricing, and optimization methods. The smart protection system provides the system reliability and failure protection, security and privacy.

The Smart Grid offers a wide research areas such as demand response, self healing, continuous supply, dynamic pricing, and power scheduling [3]. To develop and simulate those problems, the Smart Grid testbed is usually employed [3]-[14]. The Smart Grid testbeds based on the computer software were proposed in [4]-[7]. In [4], they compared several Smart Grid software simulators such as MatPower, PSAT, AMES, InterPSS, OpenPSS, GridLab-D. The simulators simulate the power system only. In [5], they proposed the SCORE (Smart Grid Common Open Research Emulator) that emulates both the power and communication networks. SCORE runs under Linux operating system. It consists of GUI (Graphical User Interface), Service Layer, Communication Module, and Power Module.

The real-time simulation of Smart Grid was proposed in [6]. They used the parallel computation combined by the advanced hardware that could handle hundreds of switches in real-time. The switching devices were modeled inside the FPGA (Field Programmable Gate Array) devices.

Our previous work [7] proposed the Smart Grid testbed using the SCADA (Supervisory Control and Data Acquisition) software configured on several computers as the client and the server. Each computer simulated the component of Smart Grid system such as the power resource and the demander. The TCP/IP protocol was adopted to communicate between the computers.

The benefits of software simulator are: a) No need high investment; b) It could simulate the complex algorithms easily; c) The simulation time is flexible. However the software simulator has the main drawback, i.e. it could not simulate the Smart Grid system in the realistic way. Some conditions in the software simulation may differ with the real implementation.

The Smart Grid testbeds using hardware simulator were developed in [3],[8]-[14]. In [3], they proposed the SmartGridLab that consits of the Intelligent Power Switch (IPS), the different types of power sources (photovoltaic, wind turbine, and utility grid), the energy demander, and the power meter. The testbed employed the 802.15.4 wireless mesh network for communicating between the nodes in the distributed mode, or between the nodes and the management center in the centralized mode. The IPS plays the important role for switching the power network, thus controlling the interconnection of power resources and demanders.

To reduce the costs of investment and installation of the renewable energy resources, the PV (photovoltaic) array and the wind turbine emulators were proposed in [8]. The PV emulator was simulated by regulating the DC output of power supply according to the I-V curve of the PV module. The wind turbine emulator consists of the wind speed emulator, the induction generator and the power inverter.

This work was supported by the Research Grant 2015 “Excellent Research on Higher Institution” from Directorate General of Higher Education, Ministry of Research and Technology and Higher Education, Republic of Indonesia (No. : SP DIPA-023.04.1.672453/2015).

The Smart Grid testbed using a wireless sensor network within a small building was developed in [9]. It consists of the application/dashboard, the server, the gateway, the electrical devices, and the information networks. The dashboard was designed to display electricity usages on a specific period. The server was used to collect and manage the electricity usages, and send the control messages for turning on/off the devices. The gateway provided the interface between the server and the wireless sensor networks. It consists of the PLC (Power Line Communication) gateway that uses the power line communication, and the Zigbee gateway that uses the Zigbee technology. The electrical devices consist of the smart meters, the wind generators, the PV generators, the electric vehicle chargers, the light controllers, and the smart outlets. The information networks consist of the PLC communication using the power line and the Zigbee technology developed based on Zigbee specification and IEEE 802.15.4.

The GSM/GPRS modems were used as the communication channel in the Smart Grid testbed developed by [10]. Data from the Smart Meter is transmitted to the Automatic Power Switches (APSs) as a SMS every three seconds. They proposed the distributed control scheme, where each APS decides its own control startegy. The centralized control scheme was proposed in [11]. In the centralized scheme, a control station was employed to collect data from the Smart Meters and to decide the control strategy for all APs.

The Smart Grid testbed was developed in [12] to simulate and monitor three-phase transmission line. The LabVIEW software was employed to display the measurement results such as the cable position, the inclination and vibration of the tower, and the frequency of the cable current. The testbed provided the unique tool for implementing the sensor networks on the transmission line.

The testbed for simulating the loads in the Smart Grid environments was proposed in [13]. They developed three types of loads, i.e. virtual, contextual and real loads. In the virtual load, the load is fully simulated by the computer. The contextual load is similar to the virtual load, but it is generated from the pre-defined value stored in the file. The real load consists of the real load connected to the Xbee module and the load connected to the PLC (Programmable Logic Controller). The testbed is focused for studying the demand side management system in the Smart Grid.

The power system components in the power system laboratories were integrated to establish the Smart Grid testbed [14]. The testbed consists of the main component usually available in the power system laboratories such as the generation station model, the transmission line model, the bus model, and the load model. The data acquisition devices were employed for monitoring and measurement. The network infrastructure was included to provide the communication network between devices.

Due to the complexity of the Smart Grid system, the testbeds are usually developed according to the particular function as described previously. In the hardware testbed, the application software is required for displaying the measurement results and for controlling the whole system. This application is usually developed using the programming language such as C++, Java or using the popular software such as LabVIEW and Matlab. The programming language needs the high programming skill, while the LabVIEW and Matlab require the specific devices for interfacing with them. On the other hand, the SCADA software is very popular software for supervising and controlling the industrial processes. The SCADA software provides the tools for building the GUI and supports many communication protocols for interfacing with the field devices.

In this research, we exploit the capability of SCADA software for implementing the Smart Grid testbed. We extend our previous work [7] by developing the Smart Grid testbed that combining the software simulation and the real hardware. The wireless communication using Xbee protocol is employed to connect the nodes with the Smart Grid Control Center. The PV and wind turbine generators are simulated using the adjustable power supply. The DC-bus is adopted to connect the power resources and the demanders.

The rest of paper is organized as follows. Section 2 presents the architecture of the proposed Smart Grid testbed. Section 3 presents the implementation of SCADA software in the Smart Grid. The experimental results are described in Section 4. The conclusions are covered in Section 5.

II. SMART GRID TESTBED ARCHITECTURE

The architecture of power network of the proposed Smart Grid testbed is illustrated in Fig. 1. It consists of three power resources (PV generator, Wind Turbine generator, and Utility Grid), two demanders (Home-1 and Home-2), and five Power Switches. They are connected by the 12 Volt DC-bus. The Power Switch plays an important role to manage the electricity flow. The Power Switch is controlled by the microcontroller that receives the command from the Smart Grid Center via the wireless communication.

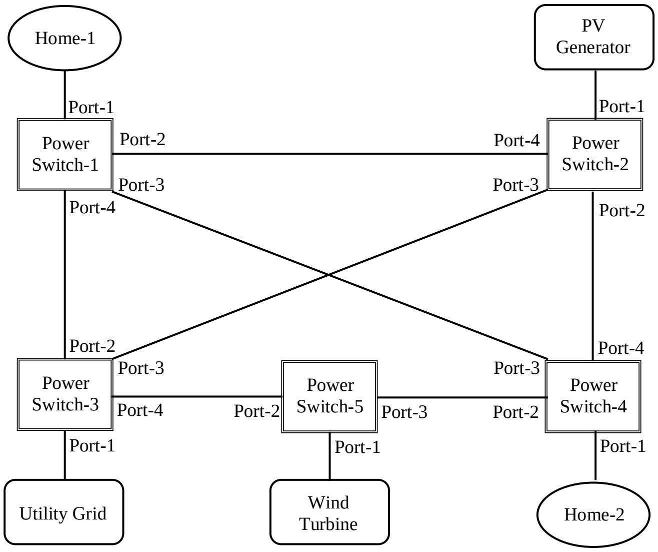


Fig. 1. Architecture of power network

Fig. 2 illustrates the architecture of information network in the Smart Grid testbed. As shown in the figure, there are five nodes and one center connected using Xbee communication. In each node, the Xbee module is attached on an embedded system using the Arduino microcontroller. The Smart Grid Center utilizes the SCADA software running on a personal computer. To connect with the nodes, the embedded system equipped with Xbee module is interfaced to the computer. The details explanation of the components are discussed in the following.

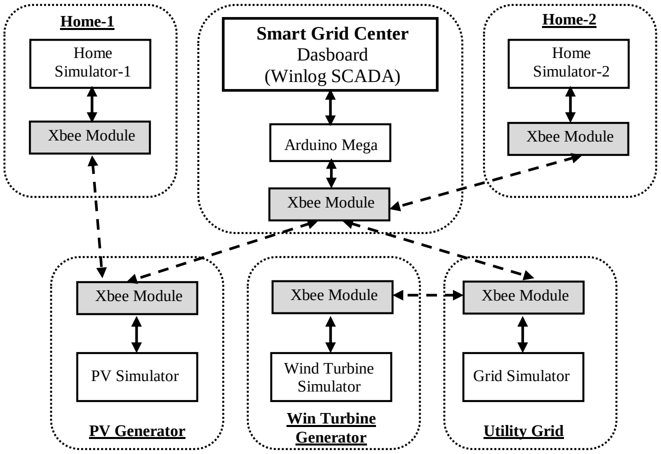


Fig. 2. Architecture of information network

A. PV and Wind Turbine Simulators

The PV simulator is designed by combining the software simulation and the adjustable power supply [15] as illustrated in Fig. 3. The simulator uses the Matlab Simulink to model the I-V characteristic of the PV module. The output of the model is interfaced to the adjustable power supply controlled by the Arduino microcontroller. To provide the simple implementation, the microcontroller controls the position of servo motor which is attached to the potentiometer of the current controlled power supply. In addition to control the adjustable power supply, the Arduino microntroller reads the current and voltage of the system from the current and voltage sensors. It sends the electrical measurements to the Smart Grid Center via the Xbee module. At the same time, the Arduino controls the Power Switch according to the command sent by the Smart Grid Center.

The Wind Turbine simulator is designed in the similar way to the PV simulator. The difference lies on the software simulation, where the Simulink models the Wind Turbine generator instead of the PV generator. The hardware interfacing is similar to the PV simulator.

B. Home Simulator

The home simulator is used to simulate the demander or the load of Smart Grid system. The main component is the load unit consists of four lamps as illustrated in Fig. 4. The load is switched-on/off by the load relay controlled by the Arduino microcontroller. The status of load is defined by the user and simulated using SCADA software in the Smart Grid Center. Therefore, it is easy for the user to configure the load conditions. Further, this scheme could be used in the direct load control algorithm, which is one of the important features in the Smart Grid application.

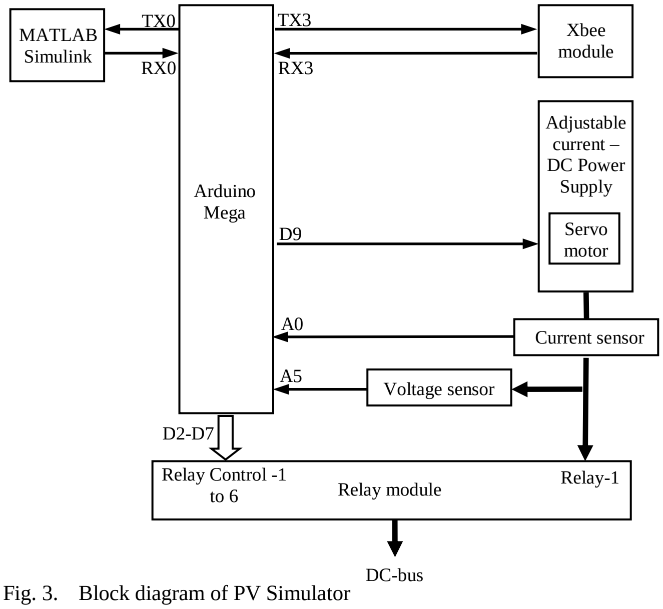


Fig. 3. Block diagram of PV Simulator

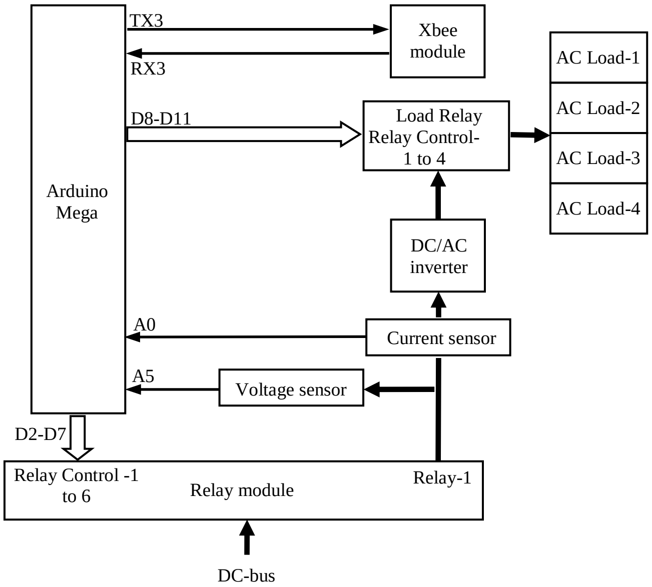


Fig. 4. Block diagram of home Simulator

C. Smart Grid Center

The proposed Smart Grid adopts the centralized scheme, thus the Smart Grid Center controls the whole procesess in the Smart Grid system. In the proposed testbed, the Winlog SCADA software[1] is employed as the software tool for developing such system. The benefits of the SCADA software are; a) It is easy to develop the application based on GUI; b) It supports many communication protocols, thus it is easy to connect to the other devices.

The Smart Grid Center is implemented on a personal computer. To provide the flexibility and reduce the cost, the Arduino microcontroller is employed as the gateway between the computer and the nodes. The Arduino communicates to the

[1] http://www.sielcosistemi.com/en/products/winlog_scada_hmi/

computer using the serial communication, while it communicates to the other nodes using the Xbee wireless communication.

D. Communication Protocols

As discussed previously, there are two communication paths from the microcontroller to the computer, and the microcontroler to the nodes. Two different communication protocols are employed. The first protocol is the SCADA protocol, which is used to communicate between SCADA software in the Smart Grid Center and the Arduino microcontroller. In this research, the Modbus protocol is employed. Modbus protocol is an open protocol and widely used in the industrial applications.

The second protocol is the Xbee protocol used to communicate between Xbee modules. In the research, the Xbee DigiMesh[2] is employed, thus the mesh network as illustrated in Fig. 5 is adopted[3]. The DigiMesh provides the simple setup and the flexibility to expand the network. In our work, all Xbee modules are configured as the DigiMesh nodes (DN). It simplifies the implementation in the Smart Grid testbed system, such as the modules are interchangeable.

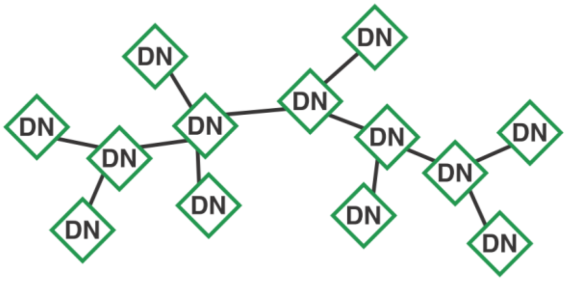


Fig. 5. DigiMesh network[3]

III. Implementation of SCADA Software in the Smart Grid

The major contribution of the paper is the implementation of SCADA software in the Smart Grid. The Winlog SCADA is the SCADA software usually used to develop the SCADA application. The objective of the research is to evaluate the capability of Winlog SCADA to be implemented in the Smart Grid application.

To implement the Winlog SCADA in the Smart Grid, four tools are utilized [7] : a) Configuration tool; b) Gate builder; c) Template builder; d) Code builder. The Configuration tool deals with the communication protocol to other devices, in this case the Modbus protocol. While the Gate builder is used to configure the variables of the process. In this system, the variables are used to define the current, voltage, power and relay status in the Smart Grid testbed.

The template builder is used to design the GUI of the Smart Grid system. Using this tool, the front panel/dashboard of the Smart Grid such as the visualization of the generators and the loads, and their corresponding electrical parameters, the lines representing the power network could be developed easily.

The Code builder is the tool for writing the scripts or program. In the research, we exploit the Code builder for implementing the Smart Grid algorithm. The simple and important algorithm is the one for solving the problem of continuous supply as desrcibed in the following.

Continuous supply is one of the simple Smart Grid feature that provides the continuous supply to the loads regardless of the power capacity of the generators or the generator's faulty. Fig. 6 illustrates the flowchart of this algorithm. The power network in Fig. 1 is used as the example.

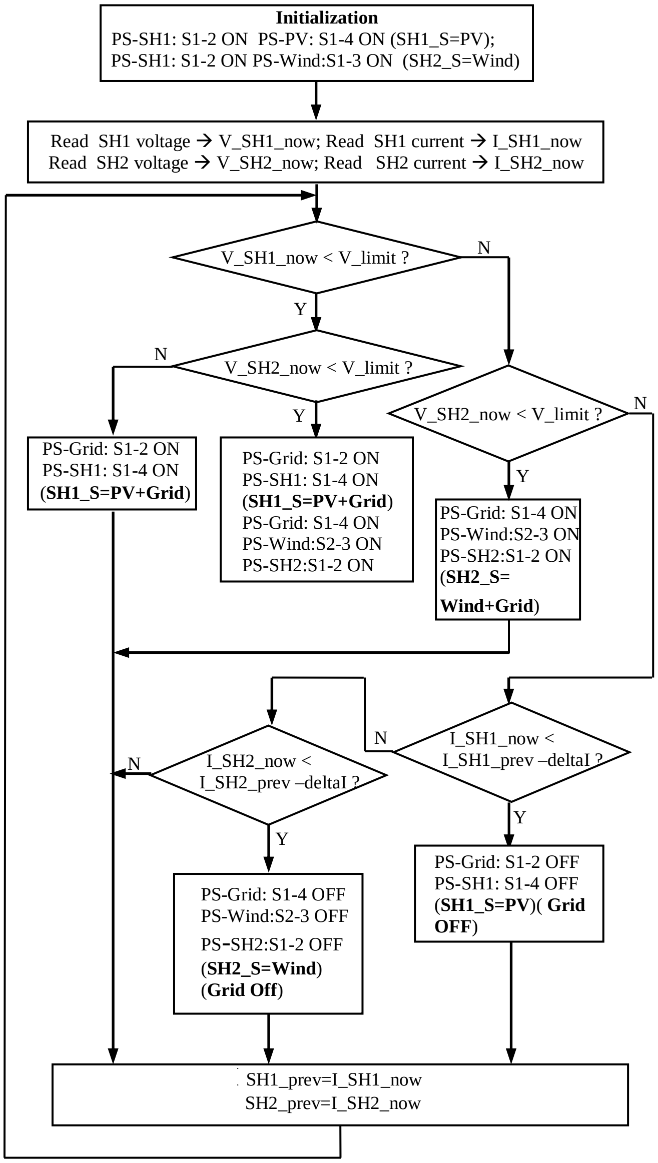


Fig. 6. Flowchart of the continuous supply algorithm

In the algorithm, it is assumed that the Home-1 gets the electricity from the PV as the primary source, and the Utility Grid as the secondary source. While the Home-2 gets the electricity from the Wind Turbine as the primary source, and the Utility Grid as the secondary source. When the power demanded by the Smarthome-1 (Smarthome-2) is lower than the power that could be delivered by the primary source, then

[2] http://www.digi.com/technology/digimesh/
[3] http://www.digi.com/pdf/wp_zigbeevsdigimesh.pdf

only the primary source (PV or Wind Turbine) works, and called as the underload condition. Otherwise, in the overload condition, both the primary and secondary sources will work together.

The algorithm starts by intializing the relay status of the Power Switches. By default, the Port-1 and Port-2 of the PS-SH1 (Power Switch of Smarthome-1) is connected by the relay S1-2 of PS-SH1; the Port-1 and Port-4 of the PS-PV (Power Switch of PV) is connected by the relay S1-4 of PS-PV; the Port-1 and Port-2 of the PS-SH2 (Power Switch of Smarthome-2) is connected by the relay S1-2 of PS-SH2; the Port-1 and Port-3 of the PS-Wind (Power Switch of Wind Turbine) is connected by the relay S1-3 of PS-Wind. Thus the PV delivers the power to the Smarthome-1, while the Wind Turbine delivers the power to the Smarthome-2.

As shown in the flowchart, to detect the overload condition, the voltage of Smarthome-1 (Smarthome-2) is compared to a certain voltage limit. If the voltage is lower than this value, the overload condition occurs and the algorithm will switch-on the Power Switch appropriately to deliver the power from the secondary source. For example, when the Smarthome-1 suffers the overload condition, then the relay S1-2 of PS-Grid and the relay S1-4 of PS-SH1 will be switched-on. Thus the Smarthome-1 is supplied by both the PV and the Utility Grid.

The algorithm needs to check whether the overload condition is already completed and changes to the underload condition. This situation is determined by comparing the decreasing current on the Smarthome-1 (Smarthome-2) with a certain threshold. When the underload load condition occurs, the algorithm will switch-off the secondary source.

IV. Experimental Results

The proposed testbed was implemented in the software and hardware. The results of Smart Home simulator developed using Winlog SCADA is illustrated in Fig. 7. In the figure, the user could control the status of load by pushing the appropriate switch. Further, the operation could be performed automatically by writing the program in the Code builder, for example by time scheduling.

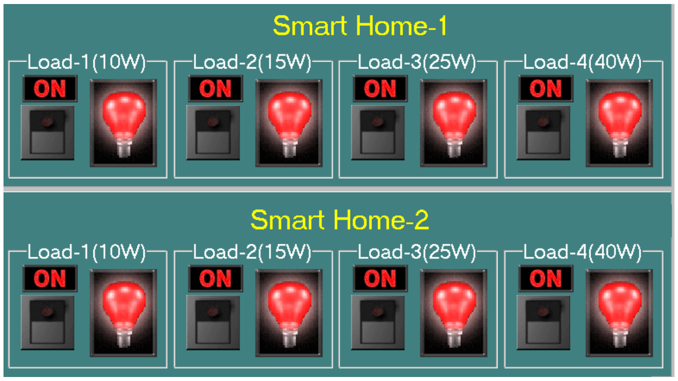


Fig. 7. Smarthome simulator developed using Winlog SCADA

To verify the proposed testbed, we conduct the experiments by varying the loads on the Smarthome-1 and the Smarthome-2 for simulating the overload and underload conditions as listed in Table 1. The power and voltage measurements on the Smart Grid testbed are illustrated in Fig. 8 and Fig. 9 respectively.

TABLE I. Load Profile

Sequence number	Load on Smarthome-1	Load on Smarthome-2
1	25 Watt	25 Watt
2	40 Watt	25 Watt
3	50 Watt	25 Watt
4	50 Watt	40 Watt
5	25 Watt	25 Watt

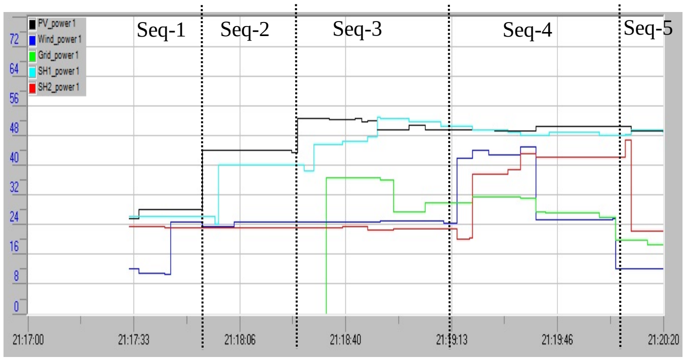


Fig. 8. Power measurement on the Smart Grid testbed

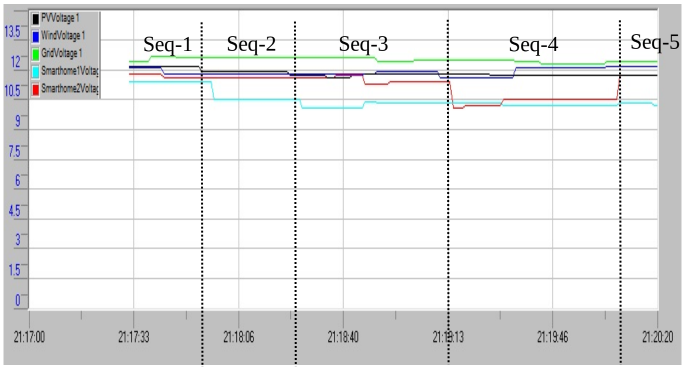


Fig. 9. Voltage measurement on the Smart Grid testbed

In the first sequence, the load of 25 Watt is applied to the Smarthome-1 and Smarthome-2. In this case, both the primary sources (PV and Wind Turbine) are sufficient to deliver the power, therefore the Utility Grid is switched-off from the grid as shown in Fig. 8. In the second sequence, the load on the Smarthome-1 is increased to 40 Watt, while the load on the Smarthome-2 remains the same. As shown in the figure, the PV is still sufficient to supply the power. The voltage on the Smarthome-1 as shown in Fig. 9 decreases, but it is still greater than the defined limit (i.e. 10.25 Volt). In the third sequence, the load on the Smarthome-1 is increased to 50 Watt. In this situation, the voltage drops below than 10.25 Volt, and the overload condition occurs. Therefore the system starts to switch-on the Utility Grid to deliver the power to the Smarthome-1. It needs the transition time about 30 second for recovery until the voltage increases above the voltage limit.

In the fourth sequence, the load on the Smarthome-1 is not changed, while the load on the Smarthome-2 is increased to 40 Watt. In this situation, the overload condition occurs and the system starts to switch-on the Utility Grid to deliver the power to the Smarthome-2. Therefore the Utility Grid delivers the power to both the Smarthome-1 and the Smarthome-2. In the fifth sequence, the load on the Smarthome-2 is decreased to 25 Watt, thus the condition changes to the underload. The system switches-off the Utility Grid for delivering the power to the Smarthome-2, but it still delivers the power to the Smarthome-1. From the experiments, it is obtained that the power measurement on the Smarthome-1 (Smarthome-2) is not exactly same with the one on the PV (Wind Turbine) or the Utility Grid. This discrepancy is caused by the error in the sensor system for reading the electrical parameters.

The algorithm is also verified by observing the operation of relays on the Power Switches and the appearance of power lines on the SCADA front panel as illustrated in Fig. 10 and Fig. 11. Fig. 10 illustrates the appearance during the first sequence when the PV supplies to the Smarthome-1 and the Wind Turbine supplies to the Smarthome-2. As shown in the figure, the respective flow line changes to red color. The relay's status on the Power Switches are also indicated.

Fig. 11 illustrates the appearance during the third sequence when the Smarthome-1 is supplied by the PV and the Utility Grid, while the Smarthome-2 is supplied by the Wind Turbine only. From the figure, it is observed that two relays on the Power Switch of the Smarthome-1 are switched-on to perform this operation.

The above results show that the proposed testbed could be used to study and research the Smart Grid system, particulary the algorithm for manage the power network. Even thought the testbed is simple and small size, it could be expanded accordingly. It is worth to note that the Smart Grid algorithms could be implemented using the SCADA software.

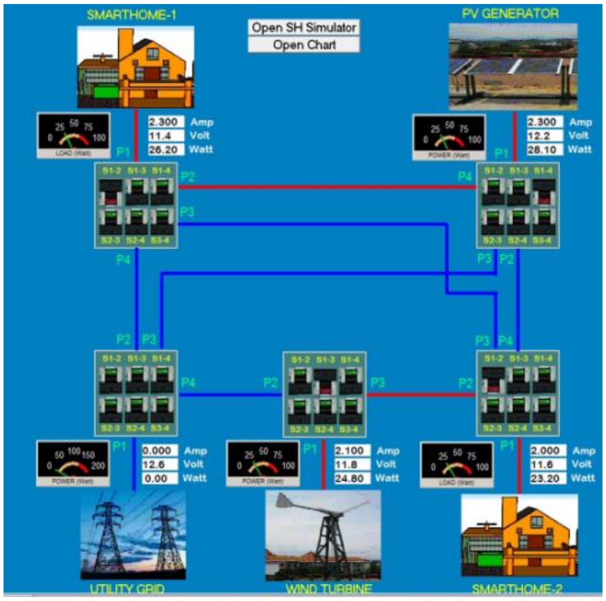


Fig. 10. Appearance of the Smart Grid front panel during the first sequence

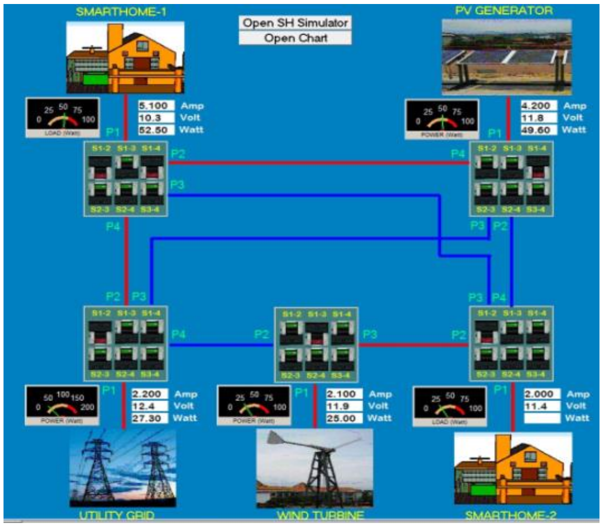


Fig. 11. Appearance of the Smart Grid front panel during the third sequence

Some of the hardware testbed simulator consists of the adjustable power supply with servo motor, the microcontroller, the Xbee module, and the Power Switch are illustrated in Fig. 12. The hardware module for each component/node is designed separately. The electrical power connections between them are conducted by the DC-bus cable. While the information connection is conducted wirelessly.

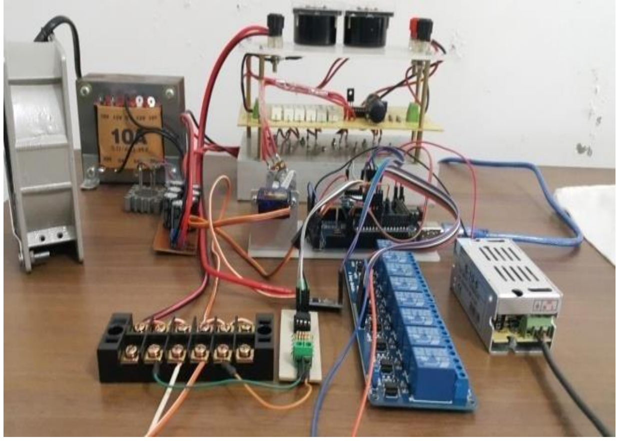


Fig. 12. PV simulator hardware.

V. CONCLUSIONS

The SCADA software is developed to implement the algorithm in the Smart Grid. The nodes on the Smart Grid testbed are communicated to the Smart Gird Center using the Xbee wireless communication. The power generators and the demanders are simulated using both the hardware and software. In the experiments, the system is tested to handle the problem of continuous supply when the power of the loads are varied. The developed Smart Grid testbed works appropriately in simulating the algorithm in the Smart Grid system. Further the front panel displayed on the Smart Grid Center could be used to help to operator/user to monitor and analyse the operation of the whole system.

In future, the complex system will be developed. Further the sophisticated algorithms in the Smart Grid will be implemented.

REFERENCES

[1] P Siano, “Demand response and smart grid – a survey,” Renewable and Sustainable Energy Reviews, Vol. 40, pp. 461-478, 2014.

[2] X. Fang, S. Misra, G. Xue, and D. Jang, “Smart Grid – The new and improved power grid: a survey,” IEEE on Communication Surveys & Tutorials, Vol. 14, No. 4, pp. 944-980, 2012.

[3] W.Z. Song, D. De, S. Tan, S.K. Das, and L. Tong, “ A wireless smart grid tesbed in lab,” IEEE Wireless Communication, Vol. 19, No. 3, pp. 58-64, 2012.

[4] M. Pochacker, A. Sobe, and W. Elmenreich, “Simulating the smart grid,” Proceedings of IEEE Grenoble PowerTech, Grenoble, France, June 16-20, 2013.

[5] S. Tan, W.Z. Song, Q. Dong, and L. Tong, “SCORE: Smart-Grid Common Open Reseacrh Emulator,” Proceedings of IEEE Third International Conference on Smart Grid Communications, Tainan, Taiwan, November 5-8, 2012.

[6] F. Guo, L. Herrera, R. Murawski, E. Inoa, C.L. Wang, P. Beauchamp, E. Ekici, and. J. Wang, “Comprehensive real-time simulation of the smart grid,” IEEE Transactions on Industry Applications, Vol. 49, No. 2, pp. 899-908, 2013.

[7] A. Soetedjo, A. Lomi, and Y.I. Nakhoda, “Simulation of smart grid using SCADA,” Proceedings of International Conference on Quality in Research (QIR) 2015, Lombok, Indonesia, August 10-13, 2015.

[8] M. Shamshiri, C.K. Gan, and C.W. Tan, “A review of recent development in smart grid and micro-grid laboratories,” Proceedings of IEEE International Power Engineering and Optimization Conference, Melaka, Malaysia, June 6-7, 2012.

[9] K.S. Kim, H. Kim, T.W. Heo, Y. Doh, and J.A. Jun, “A smart grid tesbed using wireless sensor networks in a building,” Proceedings of the Fifth International Conference on Sensor Technologies and Applications, French Riviera, France, August 21-27, 2011.

[10] S. Nithin, N. Radhika, and V. Vanitha, “Smart grid testbed based on GSM,” Proceedings of International Conference on Communication Technology and System Design, Coimbatore, India, December 7-9, 2011.

[11] S. Nithin and N. Radhika, “Centralized control station for smart grid test bed based on Windows Embedded XP 2007 and Ebox 4861S,” Proceedings of International Conference on Recent Trends in Information Processing and Computing, Kualu Lumpur, Malaysia, December 17-18, 2012.

[12] L.F. Cheung, K.S. Lui, K.K.Y. Wong, W.K. Lee, and P.W.T. Pong, “A laboratory-based three-phase smart grid sensor network testbed,” Sensors and Materials, Vol. 26, No. 5, pp. 279-290, 2014.

[13] L. Gomes, et al., “Dynamic approach and tesbed for small and medium players simulation in smart grid environments,” Proceedings of the International Federation of Automatic Control World Congress, Cape Town, South Africa, August 24-29, 2014.

[14] V. Salehi, A. Mazloomzadeh, J. Fernandez, J. Parra, and O. Mohammed, “Design and implementation of laboratory-based smart power system,” Proceedings of the 2011 American Society for Engineering Education Annual Conference, Vancouver, Canada, June 26-29, 2011.

[15] A. Soetedjo, A. Lomi, Y.I. Nakhoda, and G.E. Hendroyono, “Simulating PV Generator using Simulink Interfaced to the Adjustable Power Supply for Smart Grid Testbed,” unpublished.

A Proposed Peer Selection Algorithm for Transmission Scheduling in P2P-VOD Systems

Hatem Fetoh
Department of Information Technology
Faculty of Computers and Information
Mansoura University, Egypt

Waleed M. Bahgat
Department of Information Technology
Faculty of Computers and Information
Mansoura University, Egypt

Ahmed Atwan
Department of Information Technology
Faculty of Computers and Information
Mansoura University, Egypt

***Abstract*—Video transmission in peer-to-peer video-on-demand faces some challenges. These challenges include long transmission delay and poor quality of service. The peer selection plays an important role in enhancing transmission efficiency. For this reason, a proposed algorithm for peer selection is introduced to overcome these challenges. The proposed algorithm consists of four steps. First, the peers exchange their own buffer maps with other peers. Second, the requested segments are ordered according to their priorities. Third, neighbors of the receiver are evaluated by the efficiency estimation. Finally, the efficient sender list is applied to solve the overloading and bottleneck on the highest efficient sender. A simulation is introduced to evaluate the performance of the proposed algorithm compared to a peer selection algorithm with context-aware adaptive (CAA) data scheduling algorithm. The results show that, the proposed algorithm reduces initial buffering delay and achieves high throughput rather than CAA algorithm.**

Keywords—video-on-demand; segment transmission; chunk miss ratio; initial playback delay; peer-to-peer

I. Introduction

In P2P environment, there is no dedicated server. A peer can play as server and client at the same time. [1] The overall performance of P2P is much better than client server. There are many reasons for that. First, the peers in the network communicate and share bandwidth with each other. Second, any peer can join or leave the network dynamically. Finally, each peer serves other peers by sharing its upload bandwidth. This reduces the bandwidth costs of the server. [2]

The receiver peer must receive unavailable segments before playback deadlines. This leads to start-up buffering delay reduction. Available segments are cached in the peer's buffer. Then, peers can exchange these segments with other peers. [3]

To exchange these segments, there are two main approaches that are push-based approach and pull-based approach. In push-based approach, a sender peer pushes the available segments to a receiver. In tree-based overlays, the parent peer uses push-based approach to send available segments to children peers. In pull-based systems, each peer maintains a buffer map that is exchanged and updated periodically in P2P system. Mesh-based overlays use pull-based approaches. [4] Application examples for mesh-based overlays are PPlive [5], Coolstreaming [6].

Transmission scheduling in P2P-VOD relies on the buffer maps (BMs) exchange among peers in the network. Buffer map contains available segments, unavailable segments and playback point for each peer. In the P2P mesh-based network, the transmission scheduling algorithm is used to obtain a transmission schedule from each sender. By a transmission schedule, the sequence of requested segments is specified, their senders are selected and their transmission times for each segment are estimated. In the reason of the time constraints on the segment transmission for P2P-VOD, a transmission scheduling is a difficult task.

There are some drawbacks in the previous transmission scheduling algorithms. For random scheduling algorithm [7], it has long buffering delay to interest a video content. Overloading of fast peers is a drawback of rarest first algorithm [3]. The On-time Delivery of VBR streams (ODV) algorithm ignores the segment's playback deadline time. [3] For context-aware adaptive data scheduling algorithm (CAA), [8] the drawback is that the delivery time to pull segment from neighbors to receiver is relatively high.

There are some challenges that face Video-on-Demand in P2P. These challenges include long playback delay and the service quality. This paper proposes an algorithm for peer selection to overcome these challenges.

The proposed algorithm has two main contributions. First, the reputation factor is introduced to measure the response of each sender. Second, the efficiency of each sender is estimated to evaluate the senders and select the most efficient sender. As a result, the packet loss during transmission is reduced since the next sender is selected from the efficient sender list instead of using the first sender. This is alternative solution that is better than using the first sender since it is affected by many problems such as overloading and network bottleneck.

The remainder content of this paper is organized as follows. First, the previous transmission scheduling algorithms are discussed in section II. Second, the paper includes the efficient peer selection algorithm in section III. Third, this algorithm is evaluated by using experimental results in section IV. Finally, the conclusion is introduced in section V.

II. Related Work

P2P VOD systems' users suffer from some problems. These problems are initial buffering delays and play out lags

[9]. This reduces the video quality, increases server bandwidth costs. Hence, the previous researches suggest the heuristic transmission scheduling algorithms in peer-to-peer streaming systems such as [6], [10], and [11]. The objective of transmission scheduling algorithms is the performance improvement in P2P-VOD systems. In the following section, the most important previous segment scheduling algorithms are presented.

A. *Round-Robin Algorithm (RR)*

In this algorithm, each requested segment is received sequentially from list of senders. Each sender has the same probability to send a number of segments. This algorithm doesn't rely on the bandwidth factor or other factors to select the sender.

The main advantage of this algorithm is high balancing of transmission load to all senders. Unfortunately, the drawback of this algorithm is that it leads to non-optimal sender for transmission and causes an inefficient transmission scheduling. [8]

B. *Random Pull scheduling Algorithm (RP)*

In this algorithm, a receiver requests the segment from more peers. The requests segments are determined through a window sliding. Selection of the segments and their senders is a random selection.

This algorithm has some drawbacks. The drawbacks are long buffering delay, high packet loss and low throughput. [7]

C. *Rarest first algorithm (RF)*

Some applications use RF algorithm such as CoolStreaming/DONet. There are three main steps for this algorithm. First, the partners' count for each requested segment is estimated. Second, the segments with fewer partners are more difficult to meet deadline constraints. This leads to quickly transmission before deadline time. Hence, requested segments are ordered by partners' count factor for each segment. In the case of that there is one sender only, it must be selected. However, there are a number of partners for the requested segment; the algorithm selects the sender with highest bandwidth.

Unfortunately, this algorithm has some drawbacks. First, there are long buffering delay. Second, unavailable segments are transmitted from the highest speed senders. This causes overloading on these senders. [3]

D. *On-time Delivery of VBR streams (ODV)*

There are three main steps for ODV algorithm. First, buffer map is exchanged among peers to obtain the information needed. In addition, the expected transmission time from each sender to a receiver is computed. Finally, the selected sender is the sender can send the segment in the earliest time. [3]

Unfortunately, it has some drawbacks. First, initial buffering delay is relatively high. Second, the priority estimation of the segments ignores the playback deadline times.

E. *Context-aware Adaptive Scheduling Algorithm (CAA)*

CAA consists of four steps. First, the segment's priority is estimated for each requested segment from a receiver. Second, it estimates the bandwidth of each sender. Third, it estimates the expected delivery time that is needed to send a requested segment from sender to receiver. Finally, the sender with the earliest delivery time is selected. [8]

Unfortunately, CAA has a drawback. The initial buffering delay to stream a video content is still high.

F. *An efficient hybrid push-pull based Protocol*

In this protocol, it consists of two phases that are push-based approach and pull-based approach. The video segments are classified into urgent and non-urgent segments. First, urgent segments are transmitted by push-based approach. Second, non-urgent segments are transmitted by pull-based approach. [12]

The drawback of this protocol is that it did not discuss the sender selection approach relying on any factor. This leads to inefficient video transmission in the reason of long playback buffering delay and high chunk miss ratio.

The best one from these transmission scheduling algorithms is CAA. The reason of this is enhancing of the bandwidth estimation. This leads to delivery time reduction from the selected senders to a receiver and the throughput performance improvement.

III. Proposed Algorithm

To overcome the bottleneck on the selected sender in the previous algorithms, the efficient peer selection algorithm (EPS) is introduced. It estimates the efficiency of the senders relying on the historical reputation values in the last cycles. The proposed efficient peer selection algorithm is described in terms of four steps. These steps are the same in the proposed framework that is shown in Fig .1 [13]. First, peers exchange buffer maps over the network. Second, the requested segments are scheduled by the priority estimation of each segment. Third, neighbors are evaluated by the efficiency estimation of each neighbor peer. Fourth, the efficient senders list is applied to order the senders by the efficiency. The highest efficiency sender is selected for the transmission of requested segment to the receiver peer. TABLE .1 shows the used parameters in the efficient peer selection algorithm (EPS).

TABLE I. Parameters of the Proposed Algorithm

Parameter	Value
y	Requested segment
P_Id	Receiver's playback point
BM [x][y]	Availability of segment y in peer x
Pri[y]	Priority of the requested segment y
Neighview [x]	Neighbours set of peer x
dC	The segment with max. priority
Segset	Requested segments' set
S_{dC}	The selected sender of the segment dC
Eff (I)	The efficiency of peer I
Maxeff	The highest efficiency of all senders
S	Selected sender

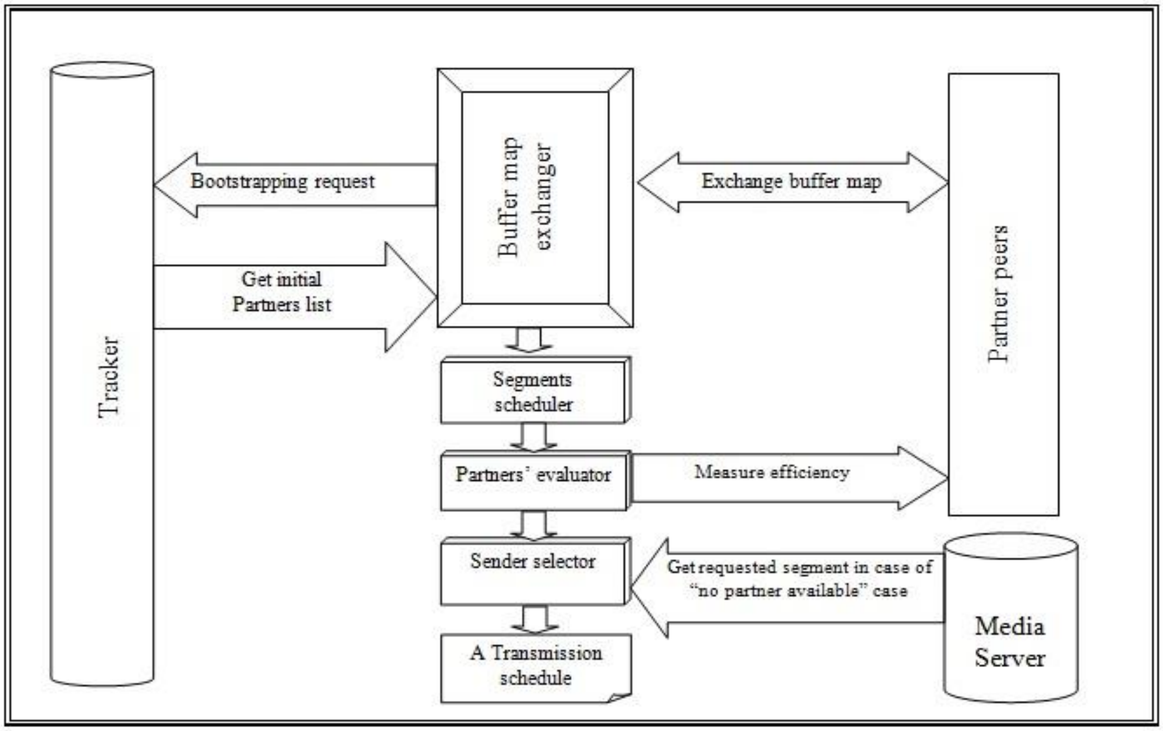


Fig. 1. The proposed framework of the transmission segment scheduling of P2P-VOD.[13]

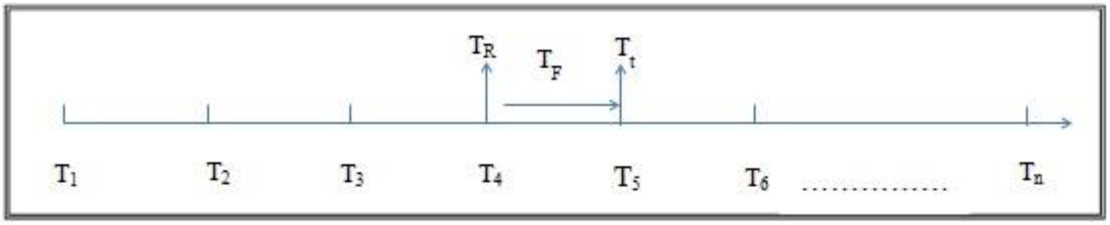


Fig. 2. Case 1 in the segment scheduler algorithm

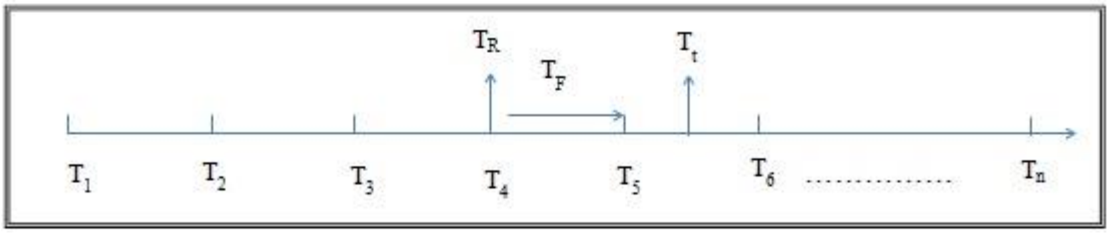


Fig. 3. Case 2 in the segment scheduler algorithm

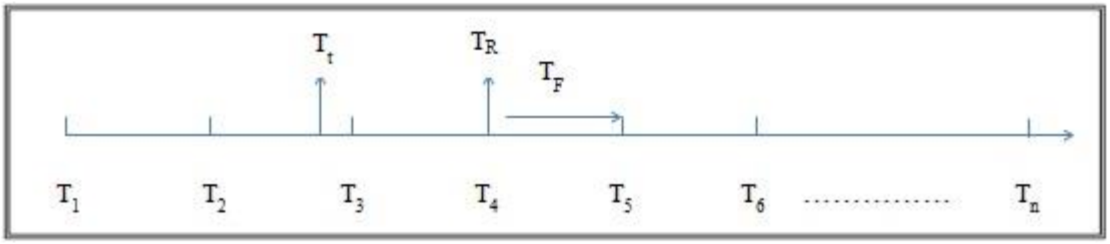


Fig. 4. Case 3 in the segment scheduler algorithm

The requested segments are determined by the buffer map of the receiver. Requested segments are ordered by the priority estimation. The requested segment's priority is estimated by a segment scheduler algorithm. The segment with the highest probability for missing is the highest priority in case 2. The parameter T_R is the playback point of the receiver. The highest priority segment must be transmitted quickly in the time Threshold T_t. The requested segment is transferred in the time T_f that is estimated as T_f =size (y) / $B_{y;}$ the parameter B_y represents the bandwidth of y. In the case 1, it's shown in Fig .2, if T_t equals ($T_R + T_F$) this means that the segment must be transmitted now. The priority of this segment is the highest. In the case 2, it's shown in Fig .3, if T_t is more than ($T_R + T_F$), the priority is estimated by the equation (1), and this means that the highest priority segment is the one that has the highest probability to be missed. In the case 3, it's shown in Fig .4, if T_t is lower than ($T_R + T_F$), this means that this segment is missed, priority equals zero.

Segment scheduler algorithm (SCA)

```
If $T_t = T_R + T_F$ Then   Pri[y] = Max.        // Case 1
Else If $T_t > T_R + T_F$ Then                   // Case 2
Pri[y] = 1/ [$T_t$ - ($T_R + T_F$)]                  (1)
Else                                             // Case 3
Pri[y] = 0
End if
```

The proposed efficient peer selection algorithm estimates the reputation of each peer that acts as a neighbor to the receiver. First, it counts the requested segments (N) which are requested from the sender I in the time cycle T_n. this algorithm relies on the historical reputation information. Second, it estimates the number of non-transmitted segments or the number of segments that are transmitted after their playback deadline times (M). Then, it estimates the reputation of the sender i in the time T_n by the equation (2).

$$R_{T_n}(I) = \frac{N - M}{N} \qquad (2)$$

Finally, it estimates the efficiency of the sender by the equation (3) that is relied on historical reputation values in the last cycles (T_1, T_2, $T_{3...}$ T_{K-1}).

$$Eff_I[j][k] = \sum_{j=1}^{m} \lambda_j . R_{T_j}(I)[j][k-1] \qquad (3)$$

The objective of EPS algorithm is selecting the most efficient sender from a large number of senders. To evaluate each sender, the efficiency parameter is introduced by the equation (3). The traditional average method is not used to estimate the efficiency of the senders according to the reputation values. The reason of this is that the weights of the time cycles are different. In the case of that the reputation value of the last cycle k-1 is low but the first cycle is very high, the average of two values gives good efficiency. The highest probability of the reputation value of the next cycle is near from the last cycle k-1. The effect of the historical reputation values of the last cycles is higher than the old cycles. Hence, the weight of the reputation in the last cycle k-1 is higher than or equals the cycle k-2 and the same in the other cycles. For the traditional average method, it leads to inefficient sender selection for transmission scheduling in P2P-VOD system. To avoid an inaccurate efficiency estimation problem, the proposed approach is introduces in equation (3). In this equation, the parameter m is a number more than zero, λ_j is a constant, $0 <= \lambda_j <= 1$, $\sum_{j=1}^{m} \lambda_j = 1$

$\lambda_j >= \lambda_{j+1} >= \lambda_{j+2} >= .... >= \lambda_m$. The algorithm estimates the efficiency of each sender I in the next cycle k. afterwards, it selects the sender with the highest Eff_I [j][k] to be an efficient sender for the requested segment. This is because that it improves throughput and can transmit segments in the earliest transmission time. This leads to selection of this peer to be the sender of the requested segment.

The proposed efficient peer selection algorithm introduces efficient senders list (ESL). This list contains the senders with high efficiency score that are ordered by the efficiency score in descend order. Then, it considers the first sender in this list as the most efficient one to be selected. In addition, if the highest sender in the list (ESL) is affected by any problem such as overloading or instability, this leads to transmission failure. The proposed algorithm solves this problem by the next sender selection in (ESL) instead of the first sender.

Algorithm(EPS) : Efficient Peer Selection Algorithm

```
For each y∈ SlideWindow
      And y > P_Id  and BM[x][y] =  0  do
     segSet = segSet υ {y} // set of requested segments
    Pri[y] = Es_Pri (y)  //compute priority by SCA algorithm
End for  y
For each i∈ Neighview [x] do
      Eff[i] = Comp_Eff (i)//estimate efficiency acc. (3)
End for i
While SegSet<> null do
dC← arg_y { Max Pri[y]  , y ∈ SegSet
$S_{dc}$ =  {ф}
For each i ∈ Neighvew [m] do
If    Eff (i) > Maxeff (S) Then {
          MaxOSS (S) = OSS(i)
        $S_{dC}$ = i
     Assign ($S_{dC}$ , dC)  }
End for P
SegSet = SegSet – {dC}
```

IV. Experimental Results

In order to evaluate the proposed efficient peer selection algorithm, Opnet Modeler 14.5 simulator is used. [14] Number of nodes in this simulation is 1500. The parameters of this simulation are shown in TABLE .2. A Dell server is used in simulation that has memory is 8GB, CPU is 8 cores Intel Pentium, and the operating system is RedHat AS5. These are the same parameters that are used in the environment of CAA algorithm.

TABLE II. Parameters Evaluation Table

Parameter	Value	Unit
Initial Bandwidth Between Peers	10	Mbps
Numbers Of Neighbours	30	
Peer Nodes	1500	
Tracker Server 1	1	
Tracker Server Bandwidth 1 Gbps	1	Gbps

Average throughput of CAA versus EPS under no. of peers

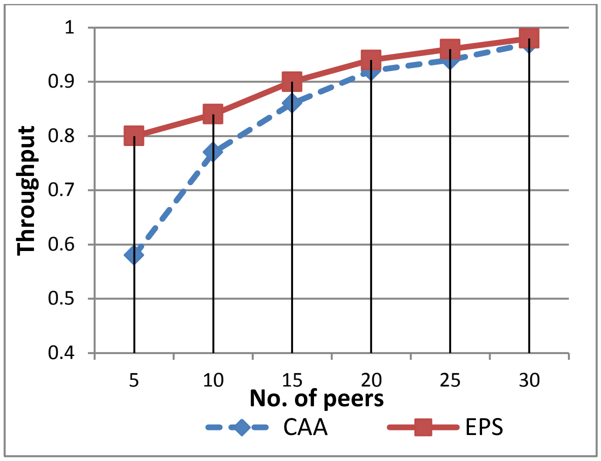


Fig. 5. Average throughput of proposed algorithm versus CAA under no. of peers

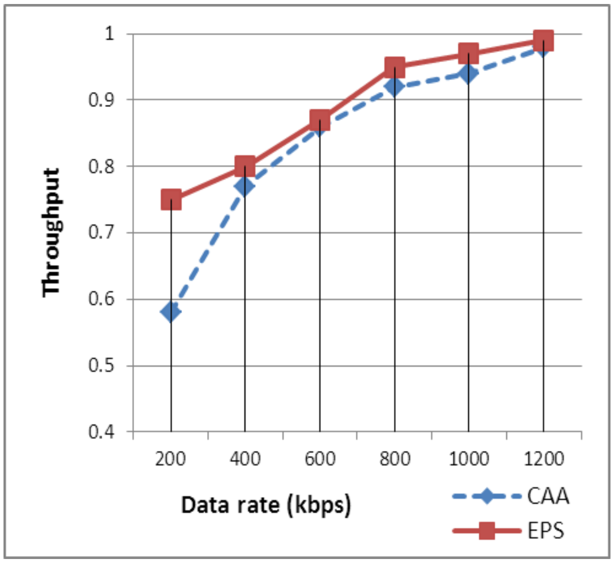


Fig. 6. Average throughput of proposed algorithm and CAA under data rate

In the first experiment, the peers exchange buffer map from each one to the others. The results are demonstrated in Fig .5. This figure shows the average throughput performance of the efficient peer selection algorithm versus CAA at number of peers (5 – 30) peer. In Fig .5, the results show the following:

- Efficient peer selection algorithm outperforms CAA under different peers' number.
- For both algorithms, in the case of that peers' number is increased, this improves the average throughput performance. This is because that the probability of high efficient senders becomes high.
- In the range (5-10), the throughput is more degraded than the other cycles. This is due to the low probability to search for an efficient peer in this region. This result indicates that the throughput of EFS is higher than CAA. The reason is that the selection of high responded peers with high reputation scores leads to improve the efficiency of the transmission from multi-senders to a receiver. This leads to throughput performance improvement.

Average throughput of CAA algorithm versus EPS under data rate

The range of data rate values are (200 – 1200) kbps.In Fig .6, these results indicates the following:

- EFS algorithm outperforms CAA under the range of data rate values (200-1200).
- For EPS algorithm, prediction of the high efficient senders leads to improvement of the throughput performance. Throughput is nearly constant in the range (800-1200). The reason is that the higher efficiency senders are affected by overload problem from large number of receivers.

Average throughput of CAA algorithm versus EPS under peer sample

Random peers are selected for this experiment. The number of these peers is twenty. The results of the experiments are shown in Fig .7. The results indicate the following:

- The EPS algorithm outperforms CAA at the range of peer samples in the most peers. The reason is that the transmission relies on selection of stable peers. These peers achieved high reputation scores. This leads to high performance for EPS algorithm.

Buffering delay of CAA versus EPS algorithm

1) Impact of Peers' Number:

The results of average buffering delay of EPS algorithm and CAA at count of peers (5 – 30) are shown in Fig .8 the results indicate the followings:

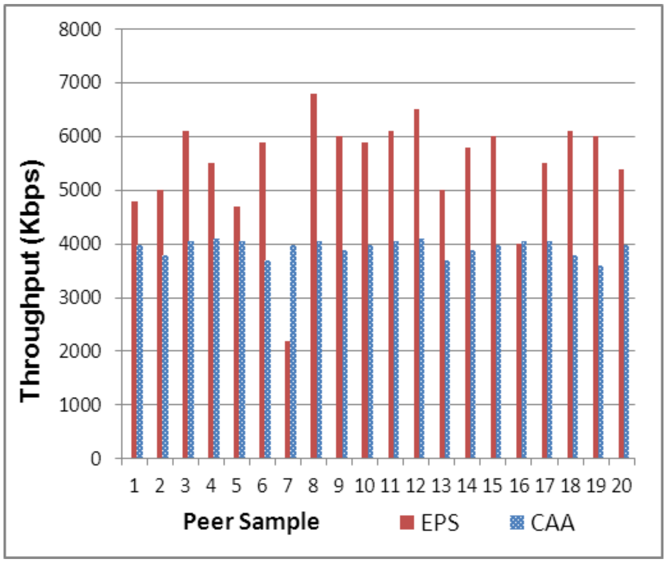


Fig. 7. Average throughput of proposed algorithm and CAA algorithm under peer sample

- The proposed EPS algorithm enhances CAA under various peers' count.

- The EPS algorithm reduces buffering delay by selecting the most efficient peer that is capable of segment transmission in the earliest time.

2) Impact of Transmission data rate:

The results of the buffering delay of EPS versus CAA at data rate range (200-1200) kbps are shown in Fig .9. The results indicate the followings:

- EPS equals CAA in data rate (200,600 and 800) kbps and it outperforms CAA in data rate (400, 1000 and 1200) kbps.
- EPS algorithm reduces buffering delay in the most from CAA. This is because that EPS is based on peer's efficiency score measurement and selection the best sender that transmits the requested segments in the lowest time. This leads to buffering delay reduction by EPS algorithm.

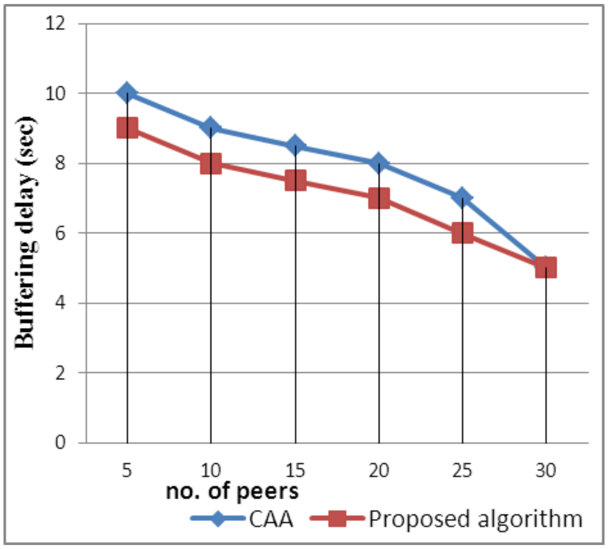


Fig. 8. buffering delay of the proposed algorithm EPS versus CAA under no. of peers

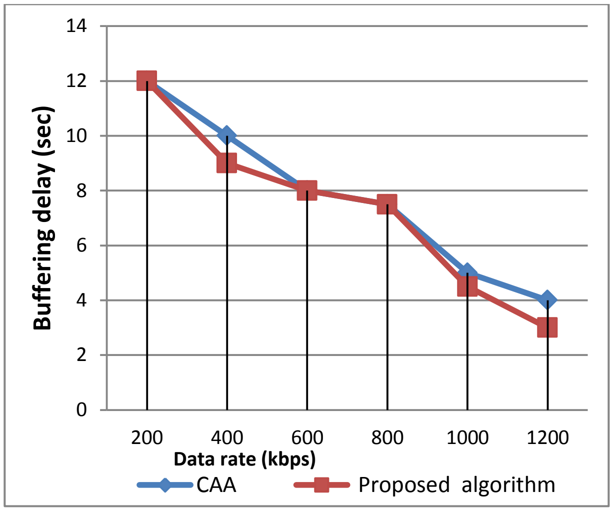


Fig. 9. buffering delay of proposed algorithm EPS versus CAA algorithm under data rate

V. Conclusion

This paper introduces a proposed algorithm for peer selection in P2P-VOD. The key contribution of the proposed algorithm is adding the reputation evaluation of peers in the historical time cycles.

This reputation is based on the number of transmitted segments before their deadlines segments compared with the number of requested segments from each sender. According to that, the proposed algorithm overcomes the bottleneck problem that occurs on the selected sender.

A simulation is introduced to evaluate the proposed algorithm compared to CAA algorithm. Our simulation results proved that there is a significant improvement in the performance of the proposed algorithm rather than CAA. The initial buffering playback delay of P2P-VOD system is decreased, continuity index is enhanced and average throughput is improved.

References

[1] García Pineda, Miguel, et al. "Controlling P2P File-Sharing Networks Traffic."Network Protocols and Algorithms 3.4 (2011): 54-92.

[2] Milojicic, Dejan S., et al. "Peer-to-peer computing." (2002).

[3] Kowalski, Greg, and Mohamed Hefeeda. "Empirical analysis of multi-sender segment transmission algorithms in peer-to-peer streaming." Multimedia, 2009. ISM'09. 11th IEEE International Symposium on. IEEE, 2009.

[4] Xu, Xiaofeng, et al. "A peer-to-peer video-on-demand system using multiple description coding and server diversity." Image Processing, 2004. ICIP'04. 2004 International Conference on. Vol. 3. IEEE, 2004.

[5] Hei, Xiaojun, et al. "A measurement study of a large-scale P2P IPTV system."Multimedia, IEEE Transactions on 9.8 (2007): 1672-1687.

[6] Zhang, Xinyan, et al. "CoolStreaming/DONet: a data-driven overlay network for peer-to-peer live media streaming." INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE. Vol. 3. IEEE, 2005.

[7] Liang, Chao, Yang Guo, and Yong Liu. "Is random scheduling sufficient in P2P video streaming?." Distributed Computing Systems, 2008. ICDCS'08. The 28th International Conference on. IEEE, 2008.

[8] Li, Wei, Quan Zheng, and Song Wang. "Context-aware adaptive data scheduling algorithm for P2P streaming systems." Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on. IEEE, 2012.

[9] Y. Liu, Y. Guo, C. Liang,"A Survey on Peer-to-Peer Video Streaming Systems," Peer-to-Peer Networking and Applications vol. 1, no. , pp. 18–28, 2008.

[10] V. Pai, K. Kumar, K. Tamilmani, V. Sambamurthy, and A. Mohr, "Chainsaw: Eliminating trees from overlay multicast," in Proc. of International Workshop on Peer-to-Peer Systems (IPTPS'05), Ithaca, NY, pp. 127–140 , February 2005.

[11] V. Agarwal and R. Rejaie, "Adaptive multi-source streaming in heterogeneous peer-to-peer networks," in Proc. of ACM/SPIE Multimedia Computing and Networking (MMCN'05), San Jose, CA, pp. 13–25, January 2005.

[12] Mohamed Ghettas , Putra Sumari," An Efficient Hybrid Push-Pull Based Protocol for VOD In Peerto-Peer network", IOSR Journal of Computer Engineering (IOSR-JCE) 16.3(2014) , 67-73

[13] Fetoh, Hatem, Waleed M. Bahgat, and Ahmed Atwan. "A proposed framework for segment transmission scheduling in P2P video streaming." Informatics and Systems (INFOS), 2014.

[14] http://opnet-modeler.software.informer.com/14.5/

Understanding Social Network Usage: Impact of Co-Presence, Intimacy, and Immediacy

Waleed Al-Ghaith
Dept. of Information Systems
Shaqra University
Riyadh, Saudi Arabia

***Abstract*—This study examines individuals' intentions and behaviour on Social Networking Sites (SNSs). The study proposed model asserts that "Co-presence", "Intimacy", "Immediacy", "Perceived Enjoyment", and "Perceived ease of use" formed individuals' "Attitude" towards "behavioral intention" to use SNSs. The results support all formulated hypotheses. The proposed model in this study explains 69% of individual's attitude or feelings towards adoption the SNSs, 59% of the variance in "Behavioral Intention" and 54% of the variance in "Usage Behaviour". The present study has shown the importance of social presence's factors, namely co-presence, intimacy, and immediacy, in explaining individuals' intentions and behavior. The study findings confirmed that social presence positively influence SNS usage indirectly through user attitude, and as aforementioned that three factors, namely co-presence, intimacy, and immediacy are framing the construct of social presence. Thus, this study is the first empirical effort to examine the impact of co-presence, intimacy, and immediacy in determining intention or behaviour.**

Keywords—Social presence theory; Social networking sites; SNS; Factors; Adoption; Usage; Saudi Arabia

I. Introduction

A Social Network Sites (SNSs) is a platform allows users to build social relations with others within and beyond their social circle [47, 19]. SNSs such as Facebook, LinkedIn, Twitter, and Google Plus have been growing rapidly in recent years [80].

This growth produced various types of SNSs. Some of SNSs have been built to be used by anyone; while others for limited purpose [45]. For example, LinkedIn and Medical Mingle, LinkedIn is the world's largest professional network and can be used by anyone, while Medical Mingle is a SNS has been built particularly for medical professionals.

With the pass of time, millions of people becoming members of one or more of these SNSs. As of March 2015, Facebook, for example, had attracted 1.44 billion active users [34]. Thus, this great success of SNSs opens a new frontier and opportunities for businesses. "More than one-third of marketers consider Facebook important for their businesses. In all, 67% of B2C and 41% of B2B companies have successfully acquired new customers through Facebook" [47]. Also, new opportunities for businesses companies can be provided by SNSs which help them to optimize their internal operations and to enhance customers' communication.

Moreover, e-commerce companies also tend to integrate SNS features into their existing web applications to enhance or retain their use in important applications [82]. Thus, due to this success, SNSs can be profitable business entities, making revenues for their stakeholders. This creates a sort of competition among the most popular SNSs, for example and "according to analytics released by ComScore in 2011, Twitter has proven to be a major competitor of Facebook, as the micro-blogging service has managed to increase its number of regular users by more than 500% since 2009" [51].

However, success of SNSs depends on their ability to attract customers using them. This is because SNS rely on number of their users and the interactions between them to increase their value. Therefore, it is important for SNSs providers to enhance their technology to attract more users and this can be done by understanding why people use SNSs. Thus, this study systematically examines what factors contribute to SNS usage. This topic matches to one of the five core research areas which forming the information systems (IS) discipline as identified by Sidorova, et al., [70]: (1) Information technology and organizations, (2) Information technology systems development, (3) Information technology and individuals, (4) Information technology and markets, and (5) Information technology and groups.

The information technology (IT) and individuals examines primarily psychological aspects of human-computer interactions, focusing on research themes such as individual technology acceptance, IT adoption, human resources issues in IS, computer self-efficacy, trust, and website design. Therefore, this study represents a mainstream area of IS research, contributing to the development of the discipline.

This study participates in this effort, theoretically and practically, by proposing an integrated theoretical model that lends itself to studying the adoption of new technologies and applies it to determine significant factors that influence adoption of SNSs. The study' proposed model brings together concepts from two distinct lines of research, the Decomposed Theory of Planned Behaviour (DTPB) from IS models and social presence theory from the social psychological theories of interpersonal communication and symbolic interactionism. Thus, the research question was:

RQ1. What are the factors that predict SNS usage?

II. Literature Review and Theoretical Framework

In the following sections, DTPB and Social presence theories are reviewed and discussed in relation to SNSs adoption in order to extract the most suitable framework for this study.

A. The Decomposed Theory of Planned Behaviour (DTPB)

Define The Decomposed Theory of Planned Behaviour (DTPB) is a decomposed version of the TPB containing several constructs from TAM and DOI. It provides the same fit as the pure TPB model but has a somewhat better predictive power relative to the TAM and TPB models [74]. According to Taylor and Todd "the decomposed Theory of Planned Behaviour provides a fuller understanding of behavioural intention by focusing on the factors that are likely to influence systems use" [74] and as "it renders more transparent and easier to grasp the relations among beliefs, attitudes and intentions, it enables application of the model to a variety of situations" [39]. In the DTPB, attitudinal beliefs, normative beliefs, control beliefs are broken down into constructs extracted from the literature and the TAM/DOI theories decomposing the attitudinal belief structure to include perceived usefulness, ease of use and compatibility. Scholars have suggested that normative belief could be decomposed into relevant reference groups such as peers, superiors, and subordinates and that each may have differing views on the use of IT. Thus, two groups (peers and superiors) have been used by Taylor and Todd [74] to represent the decomposition of normative belief structures. While the control beliefs structure can be decomposed into two groups, self-efficacy and facilitating conditions. Self-efficacy is related to the perceived ability of using a new technology, whereas the facilitating conditions construct provides two dimensions for control beliefs: one relating to resource factors such as time and money and the other focusing on technology compatibility issues that could limit usage [39, 47].

B. Social presence theory

Define Social presence is defined as the degree of salience of the other person in the interaction and the consequent salience of the interpersonal relationships [69]. Tu and McIsaac [75] redefined social presence as ''the degree of feeling, perception, and reaction to another intellectual entity in the computer-mediated communication environment''.

Three factors, namely co-presence, intimacy, and immediacy are framing the construct of social presence [15]. Co-presence, in the existing presence literature, is primarily used to refer to the sense of being together with other people, either, in a remote physical environment [58, 71], or in a technology-generated environment [33, 67]. In SNSs, user' feeling of ''being together'' is a very important perception for taking part in activities and interaction in a SNSs.

Intimacy and immediacy refer to the sense of psychological involvement which are fundamentally related to social presence theory [65]. These constructs are applied to media from the social psychological work and focused on the role of nonverbal communication in interpersonal interaction [15] (please see [9, 6, 7, 54, 8]). Intimacy can measure to which extent would people taking care, trusting, expressing themselves, and making relationships with others [69, 72, 15]. While immediacy measures interpersonal communication to assess to which degree interactivity is achieved behaviorally and perceptually (please see [17, 16, 84]).

C. Research model and hypotheses development

The research model incorporated two different theories: the Decomposed Theory of Planned Behaviour (DTPB) and social presence (see Fig. 1). A discussion of the model's constructs along with the formulated hypotheses is provided in the following sub-sections.

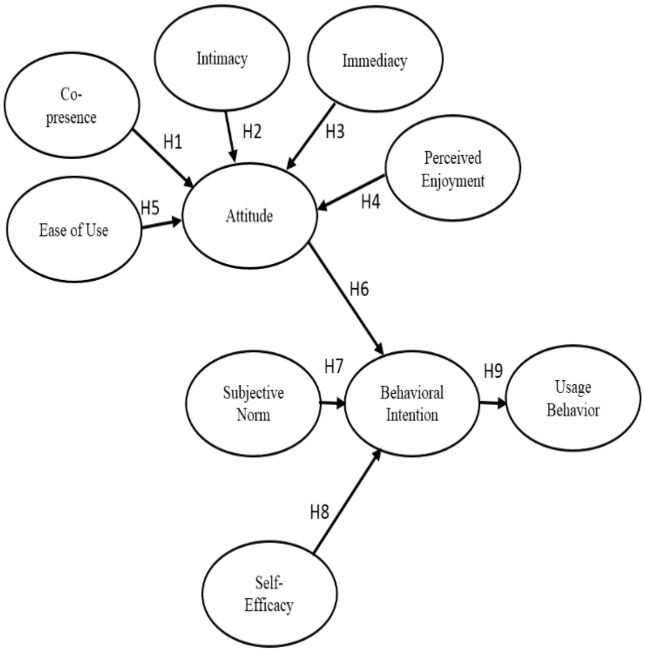


Fig. 1. The study model

1) Social presence and user attitude

The Prior studies have identified both direct and indirect impacts of social presence on intention or usage of many information technologies such as the Internet, emails, IM and e-commerce [37]. Hassanein & Head [37] found that social presence has a positive impact on attitudinal antecedents. While Xu, et al., [80] found that social presence has a positive impact on SNS usage. In this study the researcher believes that social presence will positively influence SNS usage indirectly through user attitude, and as aforementioned the three factors, namely co-presence, intimacy, and immediacy are framing the construct of social presence [15] thus, the following three hypotheses are proposed:

- Hypothesis 1. Co-presence will positively influence user attitude.
- Hypothesis 2. Intimacy will positively influence user attitude.
- Hypothesis 3. Immediacy will positively influence user attitude.

2) Perceived enjoyment and user attitude

Perceived enjoyment can be defined as the extent to which the activity of using a product or service is perceived to be enjoyable in its own right, apart from any performance consequences that may be anticipated [28, 81, 38, 43, 55]. In other words, perceived enjoyment is the individual's perception of how much fun he/she has when he/she performs an activity [11, 66].

Scholars found that there is a relationship between perceived enjoyment and perceived control, and that influences the user interaction [78, 22, 42]. In other words, within a voluntary context, if someone enjoys doing some activity he wills lose her/his self-consciousness and feeling of time and that affects his/her self-control [1, 42].

An affective reaction, such as emotion or enjoyment has been studied from a marketing perspective and it has been found that enjoyment influences cognitive perceptions or behavioral attitude [73]. Thus, the following hypothesis is proposed:

- Hypothesis 4. Perceived enjoyment will positively influence user attitude

3) Perceived ease of use and user attitude

The perceived ease of use is defined as: "the degree to which a person believes that using a particular system would be free of effort" [26]. According to the Technology acceptance model (TAM) proposed by Davis [26], two constructs (1. perceived usefulness and 2. perceived ease of use) form the behavioural beliefs that influence individuals' attitude toward information technology, which in turn predicts their acceptance of IT [26, 53, 83]. Thus, the following hypothesis is proposed:

- Hypothesis 5. Perceived ease of use will positively influence user attitude.

4) Attitude and user behavioural intention

Prior studies have shown that attitude positively influences behavioural intentions [3, 27, 2, 14, 5, 4]. Attitude is defined as an individual's feelings towards performing a specific behaviour, which is his positive or negative evaluation of performing the behaviour [27, 2].

As such, in the context of this study, attitude is defined as an individual's feelings towards adoption the SNSs and the following hypothesis is proposed:

- Hypothesis 6. Attitude will positively influence behavioural intention.

5) Subjective Norm and user behavioural intention

The subjective norm (SN) represents a person's perception of social pressure from important referents to perform or not perform a behaviour [3, 2]. In other words, individuals usually become involved in actions or an object when they have a positive attitude toward it and when they believe that important individuals think they should do so [2]. According to the theory of group influence processes, people tend to conform to others' expectations to strengthen relationships with them or in some cases to avoid a punishment [30, 50]. For example, "a student may believe that the teacher thinks that he or she should use the e-learning system. If that student is strongly motivated to comply with the expectations of the teacher, a positive impact on subjective norm may occur" [50]. This supports the effectiveness of subjective norm on behavioural intention. Therefore, the theory of reasoned action TRA [3] and the theory of planned behaviour TPB [2] measure social influence on behavioural intention through subjective norm. Accordingly, a positive relationship between subjective norms and behavioural intentions has hypothesized in many prior studies. Moreover, in the context of this study, Cheung and Lee [21] conducted a study to develop and empirically validate a research model on intentional social action in SNSs and they found that a stronger subjective norm leads to a higher level of intention to participate in an online social networking site. This finding has been also confirmed by study of Al-Debei et al. [4]. Thus, this hypothesis is proposed:

- Hypothesis 7. Subjective Norm will positively influence behavioural intention.

6) Self-efficacy and user behavioural intention

Self-efficacy and facilitating conditions. Self-efficacy is related to the perceived ability of using a new technology. Self-efficacy refers to an individual's belief in his or her own capability to perform a specific task within a particular domain [24, 50]. Bandura [13] defined self-efficacy as "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" [13].

Prior studies provided support for the relationship between computer self-efficacy and decisions involving IS adoption and usage [40, 27, 24, 44, 23, 77, 41, 32, 64].

In the context of SNS usage, online users with a higher magnitude of computer self-efficacy have a higher expectation of their ability to use a SNSs successfully with less reliance upon constant support, as a result they find using a SNS useful. Therefore, they are more likely to adopt SNS than others. Thus, the study propose that the self-efficacy construct indirectly influence usage behaviour through its direct effect on behavioural intention.

- Hypothesis 8. Self-efficacy will positively influence behavioural intention.:

7) User behavioural intention and Usage Behavior

Behavioural intention is concerned the central factor in predicting the decision to perform a certain behaviour for many information systems theories such as the Theory of reasoned action model (TRA), the Technology Acceptance Model (TAM) , the Theory of Planned Behaviour Model (TPB) and the Decomposed Theory of Planned Behaviour Model (DTPB). All these models have been widely used successfully in a range of situations and in many subject areas for predicting and understanding the performance of actual behavior [68, 20, 4] and all of them propose that an actual behaviour is significantly and directly affected by behavioural intention to perform the behavior [26, 2, 74]. Thus, the following hypothesis is proposed:

- Hypothesis 9. Behavioural intention will positively influence usage behavior.

III. Methodology

A. Measurement

Identifying the concepts or constructs that a researcher intents to measure, and then choose appropriate measuring systems to measure those constructs is essential and has a significant impact on the accuracy of findings [85]. In order to answer the research question, the researcher developed the survey instrument. The items used in the survey instrument to measure the constructs were identified and adopted from prior research; particularly from the Communication field and IS research, in order to ensure the face (content) validity of the scale used. The items were widely used in the majority of prior studies indicating potential subjective agreement among researchers that these measuring instruments logically appear to reflect accurate measure of the constructs of interest. Table 1 lists the items developed for each construct in this study as well as set of prior studies where these items have been adopted from.

B. Data Collection Procedures

Data for this study were collected in two stages (6 months apart), from samples stratified into gender groups, by means of a survey conducted in Saudi Arabia in 2014. This type of sampling technique has been chosen due to the difficulty of drawing an actual representative sample in Saudi Arabia. Most Saudi people do not have their own mail boxes and mail services are not provided for every house. Moreover, it is hard to approach women in Saudi Arabia because of cultural constraints and values. Therefore, stratified samples were drawn from several areas in the country and female relatives were engaged to distribute questionnaires to the female strata besides using electronic means to guarantee reaching females as well as males. The survey questionnaires were distributed to 1100 participants (550 male and 550 female). A total of 421 responses were received from male participants and 367 from female participants. After checking the data for validity, 657 were deemed fit for use in the analysis.

IV. Data Analysis and Results

A. Reliability and validity

A reliability and internal consistency test was performed using data obtained from the pilot study of each construct in the instrument. The alpha values from the data obtained ranged from .864 to .947 with an overall alpha value of .974. Table 2 shows the Cronbach's alpha reliability of constructs in the study. The result indicated that all constructs of the model were reliable. Therefore, the internal consistency of the instrument was acceptable.

The Kaiser–Meyer–Olkin (KMO) and principal component factor analysis were conducted to examine the adequacy of the study sample and the validity of the study instrument, respectively. As the value of KMO was 0.812 as in Table3, the study sample was considered adequate and the appropriateness of using principal component factor analysis on the collected data was assured.

TABLE I. List of Items by Construct

Construct	Items	Adapted from
Co-presence (CP)	CP1. I felt like having others with me in my SNS website. CP2. I was aware of others' presence in my SNS website. CP3. I felt others close to me in my SNS website.	[79].
Intimacy (IN)	IN1. I had a warm and comfortable relationship with others in my SNS. IN2. I received considerable emotional support from others in my SNS. IN3. I felt emotionally close to others in my SNS website.	[79].
Immediacy (IM)	IM1. I found myself respected by others in my SNS website. IM2. I found myself encouraged by others in my SNS website. IM3. I found myself assisted by others in in my SNS website.	[79].
Perceived enjoyment (NJ)	NJ1. I would/do find it fun to use my SNS website. NJ2. I would/do find it exciting to use my SNS website. NJ3. I would/do find it enjoyable to use my SNS website. NJ4. I would/do find it interesting to use my SNS website. NJ5. Interacting with my SNS website would/ does spark my imagination. NJ6. Using my SNS website would/ does make me curious. NJ7. I feel spontaneous when I use my SNS website.	[59, 60].
Perceived ease of use (ES)	ES1. Learning to use SNS website was easy for me. ES2. I find SNS website easy to use. ES3. English language is not a barrier when I use SNS website.	[26, 60, 62, 49].
Attitude (AT)	AT1. I have positive opinion in SNS website. AT2. I think usage of SNS website is good for me AT3. I think usage of SNS website is appropriate for me	[26, 2, 4].
Subjective Norm (SN)	SN1. My friends would think that I should use SNS website. SN2. My colleagues/classmates would think that I should use SNS. SN3. People who are important to me would think that I should use SNS.	[74, 31, 4].

Behavioural intention (BI)	BI1. I intend to use SNS website in next three months. BI2. I expect my use of the SNS website to continue in the future.	[74, 57, 35, 4].
Self-efficacy (SE)	SE1. I can use SNS website even if there was no one around to show me how to do it. SE2. I can use SNS website with only the online help function for assistance SE3. If I wanted to, I could easily use any of SNS website on my own. SE4. I would be able to use SNS website even if I had never used a system like it before	[74, 62].
SNS Usage (US)	US1. On average, each week I use my SNS website often US2. For each log session, I use my SNS web site long US3. On my SNS, I often post something US4. On my SNS, I often view something US5. On my SNS, I often share something US6. On my SNS, I often reply to others	[5, 80].

TABLE II. CRONBACH'S ALPHA RELIABILITY OF CONSTRUCTS IN THE STUDY

Construct	**Number of Items**	**Cronbach's Alpha**
Co-presence	3	.876
Intimacy	3	.864
Immediacy	3	.938
Perceived Enjoyment	7	.947
Perceived ease of use	3	.905
Attitude	3	.912
Subjective Norm	3	.940
Self-efficacy	4	.900
Intention	2	.889
Usage	5	.938
Overall alpha value	36	.974

TABLE III. KMO AND BARTLETT'S TEST

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.812
Bartlett's Test of Sphericity	Approx. Chi-Square	5465.154
	Df	45
	Sig.	.000

Construct validity was assessed by conducting factor analysis to calculate a principal components analysis with a Varimax rotation. This analysis helped in evaluating the convergent and discriminant validity of items. The convergent validity was evaluated by examining whether items of a variable converged together on a single construct [63], and whether the factor loading for every item was > 0.45, as suggested by Comrey and Lee [25]. Comrey and Lee [25] suggested that loadings in excess of 0.45 could be considered fair, whereas it might be considered as good if loadings were greater than 0.55, and those of 0.63 very good, and those of 0.71 as excellent. The discriminant validity was evaluated by examining the cross loading of items on different factors. As the factor pattern shows in Table 4, loadings on the target factor are in the excellent range (22 out of 36), very good (5 out of 47), good (7 out of 47), and Fair (2 out of 47). As Table 4 shows, no weak loading was found indicating the validity of constructs applied in this study.

B. *Hypotheses testing*

The study proposes a model that lends itself to studying the adoption of new technologies and applies it to determine significant factors that influence adoption of SNSs in Saudi Arabia. This model can be constituted through the test of 9 hypotheses. These hypotheses identify the relationship among factors as independent variables that impact adoption behaviour. Each accepted hypothesis represents an explanation of usage behaviour as dependent variables. Explanations are nomothetic and advance via deductive reasoning.

The simple correlation amongst all the study variables was conducted using Pearson's correlation analysis as shown in Table 5. As variables showed significant correlations ($p \leqslant 0.01$), the researcher then utilized the regression model to test multicollinearity by examining collinearity statistics; i.e. Variance Inflation Factor (VIF) and tolerance.

To determine whether any multicollinearity effects existed, the researcher checked whether there was any warning message produced by the AMOS output that signalled a problem of multicollinearity. The results showed that there was no evidence of multicollinearity. The potential problem of multicollinearity can be further examined formally in the context of regression analysis.

In Table 6, the tolerance values ranged from 1.000 to 0.556. One way to quantify collinearity is with variance inflation factors (VIF). Although a variance inflation factor (VIF) that is less than or equal to 10 (i.e. tolerance >0.1) is commonly suggested [10, 50]. Lee [48] suggested that a variance inflation factor (VIF) greater than 3 is an indicator of a serious problem of multicollinearity. In this study, a variance inflation factor (VIF) greater than 3 is considered to indicate a serious problem of multicollinearity. However, as shown in Table 3, there were no VIF values over 3 in the model; since the VIFs values ranged from 1.000 to 2.471. Thus there was no evidence of multicollinearity.

TABLE VI. MULTICOLLINEARITY TEST

Dependent variable	Path direction	Independent variables (predictors)	Collinearity Statistics	
			Tolerance	VIF
Attitude	←	Co-presence	.748	1.337
Attitude	←	Intimacy	.552	1.811
Attitude	←	Immediacy	.553	1.807

Dependent variable	Path direction	Independent variables (predictors)	Collinearity Statistics	
			Tolerance	VIF
Attitude	←	Perceived Enjoyment	.535	1.869
Attitude	←	Perceived ease of use	.405	2.471
Intention	←	Attitude	.562	1.780
Intention	←	Subjective Norm	.791	1.264
Intention	←	Self-efficacy	.667	1.499
Usage	←	Intention	1.000	1.000

To ensure the validity and reliability of the results and to use regression analysis in an appropriate manner data should be normally distributed. Jarque–Bera (skewness-kurtosis) test has been applied in this study to provide a comparison of the distributions of the research data and the normal distribution.

The symmetry of the distribution can be identified by Skewness values. If skewness value is positive, then data are clustered to the left of the distribution; otherwise data are clustered to the right of the distribution. While the height of the distribution can be measured by Kurtosis values. Positive kurtosis values indicate a peaked distribution, whilst negative kurtosis values suggest a flatter distribution (Hair, et al., 1998). Skewness–kurtosis acceptable values have been suggested by many scholars to be within the range of ±2.58 at the 0.01 significance level [12, 4]. Thus, Jarque–Bera (skewness-kurtosis) test has been applied in this study and the result is summarized in Table 7. Table 7 shows that the study data are all within the recommended range and this gives us a green light to use the regression analysis.

TABLE IV. FACTOR ANALYSIS OF ITEMS SORTED BY CONSTRUCT (ROTATED COMPONENT MATRIX (A))

	Component						
	1	2	3	4	5	6	**Its assessment**
SE1	**.758**	.401	.016	.168	.197	.105	*Excellent > 0.71*
SE2	**.790**	.103	.269	-.001	.287	.290	*Excellent > 0.71*
SE3	**.754**	.215	.224	.131	.175	.096	*Excellent > 0.71*
SE4	.555	**.586**	-.040	.290	.235	.003	*Good > 0.55*
AT1	**.616**	.442	.146	.252	.191	.308	*Good > 0.55*
AT2	.503	.489	.139	**.508**	-.092	.213	*Fair > 0.45*
AT3	**.574**	.470	.311	.400	.101	.264	*Good > 0.55*
NJ1	.347	**.695**	.294	.270	.277	.087	*Very good > 0.63*
NJ2	.137	**.631**	.297	.325	.328	.181	*Very good > 0.63*
NJ3	.271	**.723**	.319	.223	.303	.013	*Excellent > 0.71*
NJ4	.240	**.807**	.297	.002	.131	.033	*Excellent > 0.71*
NJ5	.193	**.894**	.192	.072	-.045	.094	*Excellent > 0.71*
NJ6	.132	**.739**	.090	-.080	.128	.423	*Excellent > 0.71*
NJ7	.278	**.758**	.220	.135	.074	.361	*Excellent > 0.71*
IM1	.215	.334	**.875**	.088	-.018	-.031	*Excellent > 0.71*
IM2	.197	.274	**.834**	.052	.183	.100	*Excellent > 0.71*
IM3	.246	.173	**.828**	.018	.197	.130	*Excellent > 0.71*
IN1	.310	.313	**.587**	-.094	.441	.303	*Good > 0.55*
IN2	.177	.185	.565	-.016	.069	**.612**	*Good > 0.55*
IN3	.162	.153	**.833**	.227	.090	.227	*Excellent > 0.71*
CP1	.148	-.022	.388	.452	**.634**	.007	*Very good > 0.63*
CP2	.211	.234	.194	.281	**.799**	-.049	*Excellent > 0.71*
CP3	.344	.214	.076	.275	**.746**	.158	*Excellent > 0.71*
SN1	.203	.131	.069	**.904**	.186	-.078	*Excellent > 0.71*
SN2	.118	.096	.022	**.852**	.304	.122	*Excellent > 0.71*
SN3	.214	.089	.032	**.878**	.094	.102	*Excellent > 0.71*
ES1	.354	.448	.273	.105	.054	**.678**	*Very good > 0.63*
ES2	.231	.443	.139	.357	.035	**.641**	*Very good > 0.63*
ES3	.500	**.577**	.225	.186	-.106	.376	*Excellent > 0.71*

BI1	**.552**	.198	.264	.533	.287	.188	*Good > 0.55*
BI2	**.517**	.302	.185	.460	.078	.381	*Fair > 0.45*
US1	**.576**	.479	.317	.344	.174	.312	*Good > 0.55*
US2	**.863**	.153	.114	.110	.076	.119	*Excellent > 0.71*
US3	**.743**	.278	.114	.323	.256	-.184	*Excellent > 0.71*
US4	**.801**	.059	.356	.167	.136	.165	*Excellent > 0.71*
US5	**.745**	.385	.385	.161	-.089	.075	*Excellent > 0.71*

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 7 iterations.

TABLE V. CORRELATION ANALYSIS AMONGST THE VARIABLES

	US	BI	AT	NJ	IM	CP	ES	SN	SE
BI	.747**								
AT	.785**	.749**							
NJ	.629**	.539**	.764**						
IM	.730**	.638**	.768**	.802**					
CP	.556**	.678**	.584**	.515**	.485**				
ES	.679**	.638**	.729**	.771**	.922**	.384**			
SN	.455**	.412**	.460**	.396**	.295**	.412**	.367**		
SE	.573**	.521**	.577**	.355**	.431**	.144**	.499**	.246**	
IN	.761**	.619**	.770**	.828**	.903**	.532**	.838**	.477**	.255**

US: SNS Usage, BI: Behavioural intention, AT: Attitude, NJ: Perceived Enjoyment, IM: Immediacy, IN: Intimacy, CP: Co-presence, ES: Perceived ease of use, SN: Subjective Norm, SE: Self-efficacy.
** $p \leq 0.01$

TABLE VII. NORMALITY TEST

	Skewness	Kurtosis
Subjective Norm	-.555	.153
Perceived ease of use	-.383	-.845
Co-presence	-.905	1.571
Immediacy	-.687	.090
Intimacy	-.922	.559
Perceived Enjoyment	-.394	-.525
Self-efficacy	.000	-.923
Attitude	-.639	.001
Behavioural intention	-.583	-.136
SNS Usage	-.944	.673

After assuring that necessary requirements are all adequately met, the study hypotheses were tested using multiple regression analysis.

First, “Co-presence”, “Intimacy”, “Immediacy”, “Perceived Enjoyment”, and “Perceived ease of use” were regressed on “Attitude”. As in Fig. 2, it was found that “Co-presence” (β = 0.227, Standardized path coefficient, p < 0.05), “Intimacy” (β = 0.138, Standardized path coefficient, p < 0.05), “Immediacy” (β = 0.150, Standardized path coefficient, p < 0.05), “Perceived Enjoyment” (β = 0.280, Standardized path coefficient, p < 0.05), and “Perceived ease of use” (β = 0.172, Standardized path coefficient, p < 0.05) are significantly and positively related to “Attitude” (adjusted R^2=0.69) (see Table 8, Table 9 and Fig. 2). Thus, H1, H2, H3, H4 and H5 are supported.

TABLE VIII. COEFFICIENTS FOR PROPOSED MODEL

Dependent variable	Path direction	Independent variables (predictors)	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
			B	Std. Error	Beta		
Attitude	←	Co-presence	.258	.030	.227	8.515	.000
Attitude	←	Intimacy	.170	.069	.138	2.463	.014
Attitude	←	Immediacy	.172	.082	.150	2.088	.036
Attitude	←	Perceived Enjoyment	.331	.048	.280	6.854	.000
Attitude	←	Perceived ease of use	.196	.066	.172	2.967	.003
Intention	←	Attitude	.443	.027	.488	16.149	.000
Intention	←	Subjective Norm	.137	.022	.171	6.262	.000
Intention	←	Self-efficacy	.303	.032	.286	9.514	.000
Usage	←	Intention	.772	.028	.736	27.839	.000

P values less than 0.05 were considered statistically significant

TABLE IX. STANDARDIZED REGRESSION WEIGHTS

Criterion variable	Path direction	Criterion variable predictors	Estimate	(Significance)
Attitude	←	Co-presence	.227	Significant
Attitude	←	Intimacy	.138	Significant
Attitude	←	Immediacy	.150	Significant
Attitude	←	Perceived Enjoyment	.280	Significant
Attitude	←	Perceived ease of use	.172	Significant
Intention	←	Attitude	.488	Significant
Intention	←	Subjective Norm	.171	Significant
Intention	←	Self-efficacy	.286	Significant
Usage	←	Intention	.736	Significant

Thereafter, the three independent variables (i.e. “attitude”, “subjective norms” and “Self-efficacy”) were regressed on “Behavioral Intention”. Results, as in Fig. 2, indicate that all three variables are significantly and positively related to “Behavioral Intention” (adjusted R2=0.590): “attitude” (β = 0.488, Standardized path coefficient, p < 0.05), “subjective norms” (β = 0.171, Standardized path coefficient, p < 0.05) and “Self-efficacy” (β = 0.286, Standardized path coefficient, p < 0.05) (see Table 8, Table 9 and Fig. 2). Thus, H6, H7 and H8 are supported.

Finally, the ninth Hypothesis was tested using multiple regression analysis which showed that “behavioural intention” (β = 0.736, Standardized path coefficient, p < 0.05) has a significant and positive effect on “usage behavior” (adjusted R^2=0.540) (see Table 8, Table 9 and Fig. 2). Thus, H9 is supported.

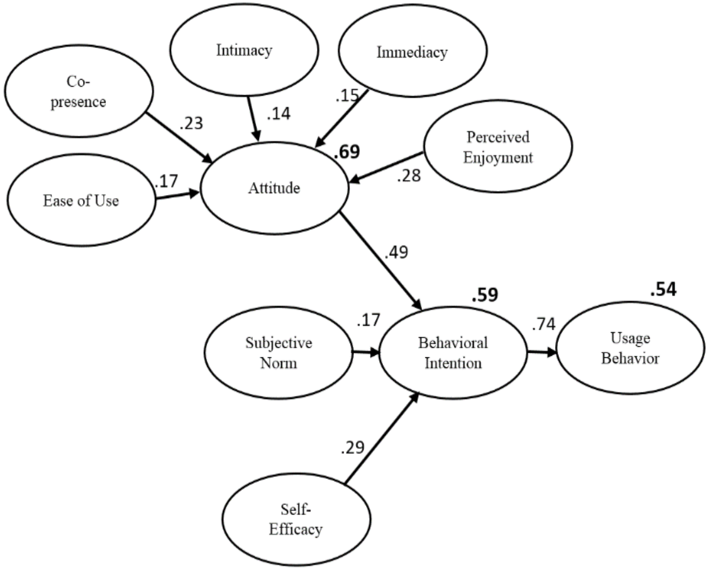


Fig. 2. The study results

V. Discussion

Overall, the results support the validity of the proposed model. The study' model shows that usage behaviour on online social networking services (SNSs) is determined by "behavioral intention" which in turn is determined by individuals' "attitude", "subjective norms", and "self-efficacy". The developed model also asserts that "Co-presence", "Intimacy", "Immediacy", "Perceived Enjoyment", and "Perceived ease of use" formed individuals' "Attitude" towards "behavioral intention" to use online social networking services (SNSs). The results support all formulated hypotheses. The proposed model in this study explains 59% of the variance in "Behavioral Intention" and 54% of the variance in "Usage Behaviour".

The results show that behavioural intention (BI) is the primary, direct determinant of usage behaviour (B) since "a person who intends to take a certain action is likely to carry out that behaviour" [46]. This result concurs with many prior studies such as Taylor and Todd study, according to their research, "behavioural intention plays an important substantive role, but is also important pragmatically in predicting behaviour" [74]. Likewise, De Guinea and Markus [29] indicate that IT use behaviour is the result of conscious, cognitive behavioural intention. The importance of behavioural intention towards usage behaviour is also reported in Venkatesh, Brown, Maruping and Bala [76]' study where they found that behavioural intention was a better predictor of duration of use and "behavioural intention will improve as a predictor of behaviour as individuals gain experience with the target behaviour" (p. 488). Thus most models position behavioural intention as an important mediating variable to be a primary predictor of behaviour, and simultaneously predicted by one or more independent variables. These independent variables or factors such as "attitude", "subjective norms" and "self-efficacy" also influence significantly and indirectly the behavior.

The results also show that the constructs "attitude", "subjective norms" and "self-efficacy" are significantly and positively related to "behavioral intention". This result confirms the role of these constructs in shaping users' behavioral intention in the SNS context. However and although all of the three constructs were found to be significant, the relation between "attitude" and "behavioral intention" is stronger (β= 0.49, pb0.001). Indeed, such a strong relation between the two constructs is evident in the literature [3, 27, 2, 14, 5, 4].

Consistent with the research hypotheses on individual's "attitude", the study' findings suggest that "Co-presence", "Intimacy", "Immediacy", "Perceived Enjoyment", and "Perceived ease of use" have a significant positive impact on individual's "attitude" towards SNSs usage.

A. Understanding Behaviour

The proposed models able to explain 54% of SNSs usage behavior. This ability relates to the diversity of the model's constructs and the diversity of relations among their constructs. In the model, behavioural intention is the primary, direct determinant of behaviour on the premise that "a person who intends to take a certain action is likely to carry out that behaviour" [46]. However, the additional explanatory power afforded by the other relative factors. An equation has been formulated and used to calculate the participation of every model's construct in the model's explanatory power. The formula was applied to the model using the total (direct and indirect) effects of each model's construct on the SNSs usage behavior (see Table10, Table11 and Table 12) as follow:

$$A_x = \frac{{\beta_x}^2}{\sum_{k=1}^{n} {\beta_x}^2} \times {R_B}^2$$

Where:

A_x = Participation of variable A_x in a model' explanatory power

${\beta_x}^2$ = Square of beta coefficients or standardized coefficients of variable

${R_B}^2$ = Model' explanatory power (*behaviour*)

$\sum_{k=1}^{n} {\beta_x}^2 =$ Total of causal effects for the model's constructs

TABLE X. Decomposition of Total Causal Effects for the Model's Constructs

	NJ	IN	IM	ES	CP	SE	SN	AT	BI
AT	.28	.138	.150	.172	.227	.000	.000	.000	.000
BI	.137	.067	.073	.084	.111	.286	.171	.488	.000
US	.101	.050	.054	.062	.082	.210	.126	.359	.736

TABLE XI. Standardized Direct Effects

	NJ	IN	IM	ES	CP	SE	SN	AT	BI
AT	.280	.138	.150	.172	.227	.000	.000	.000	.000
BI	.000	.000	.000	.000	.000	.286	.171	.488	.000
US	.000	.000	.000	.000	.000	.000	.000	.000	.736

TABLE XII. Standardized Indirect Effects

	NJ	IN	IM	ES	CP	SE	SN	AT	BI
AT	.000	.000	.000	.000	.000	.000	.000	.000	.000
BI	.137	.067	.073	.084	.111	.000	.000	.000	.000
US	.101	.050	.054	.062	.082	.210	.126	.359	.000

Table 13 shows the participating models' variables and their explanatory power. In the proposed models, behavioural intention is the primary, direct determinant of behaviour and its participation in the models' explanatory power was the highest amongst the constructs. Behavioural intention was able to explain 39% of usage behaviour. This shows that behaviour is largely driven by behavioural intention and that has a notable impact on the model's explanatory power.

TABLE XIII. Participation of Model's Variables in the Models' Explanatory Power

Constructs	**The proposed model**
Intention	39%
Attitude	9.47%

Self-Efficacy	3.24%
Subjective Norm	1.16%
Perceived Enjoyment	0.75%
Co-presence	0.49%
Perceived ease of use	0.27%
Immediacy	0.21%
Intimacy	0.18%
Total	54%

This impact can be obviously seen when behavioural intention is excluded from the model, since the prediction of behaviour decreases substantially (from $R^2(B) = 0.54$ to $R^2(B) = 0.15$. The drop in predictive power when behavioural intention was omitted concurs also with the Taylor and Todd study, according to their research, "behavioural intention plays an important substantive role, but is also important pragmatically in predicting behaviour" [74]. Likewise, De Guinea and Markus [29] indicate that IT use behaviour is the result of conscious, cognitive behavioural intention. The importance of behavioural intention towards usage behaviour is also reported in Venkatesh, Brown, Maruping and Bala [76]' study where they found that behavioural intention was a better predictor of duration of use and "behavioural intention will improve as a predictor of behaviour as individuals gain experience with the target behaviour" (p. 488). Thus most models position behavioural intention as an important mediating variable to be a primary predictor of behaviour, and simultaneously predicted by one or more independent variables. These independent variables or factors also influence significantly and indirectly the SNSs usage behavior through behavioural intention. The three antecedents of behavioural intention (“attitude”, “subjective norms” and “self-efficacy”) explain around 14% of usage behavior, while, “attitude” alone explain 9.47% of SNSs usage behavior. Participation of “attitude” in the models' explanatory power refers to five antecedents which are “Perceived Enjoyment”, “Co-presence”, “Perceived ease of use”, “Immediacy”, and “Intimacy” which all form 69% of individual’s attitude or feelings towards adoption the SNSs.

This result is consistent with the findings that the three factors, namely co-presence, intimacy, and immediacy which are framing the construct of social presence [15] have a positive impact on attitudinal antecedents [37] and in line with study of While Xu, et al., [80] which found that social presence has a positive impact on SNS usage. Moreover, the study’ findings also confirm that “perceived Enjoyment” ($\beta = 0.280$, Standardized path coefficient, $p < 0.05$), and “Perceived ease of use” ($\beta = 0.172$, Standardized path coefficient, $p < 0.05$) are significantly and positively related to “Attitude” (adjusted R2=0.69), This result is consistent with the findings of Sun and Zhang, [73]’s study which found that enjoyment influences cognitive perceptions or behavioral attitude [73]. While result of impact of “perceived ease of use” on “attitude” confirms most prior studies [26, 53, 83] which found that “perceived ease of use” form the behavioural beliefs that influence individuals’ attitude toward information technology, which in turn predicts their acceptance of IT.

The study’ findings also confirm that the self-efficacy construct indirectly influence usage behaviour through its direct effect on behavioural intention. This indicates that SNSs users, who are confident of their abilities to use Internet and SNS' sites are more likely to adopt such services. The result is consistent with the findings of most prior studies [40, 27, 24, 44, 23, 77, 41, 52, 32, 64] which provided support for the relationship between computer self-efficacy and decisions involving IS adoption.

The study findings also show that subjective norm has positive significant (.171) direct effects on intention to use SNSs. The result is consistent with the findings of Montesarchio [56]' study which found that subjective norm was positive explanatory variables of intent. Furthmore the result also is perfectly consistent with a study by Cheung and Lee which found that a stronger subjective norm leads to a higher level of intention to participate in an online social networking site [21]. This finding has been also confirmed by study of Al-Debei et al., [4].

VI. Implications for Theory and Practice

A. Implications for theory and research

SNSs in general represent a rapidly growing phenomenon that touches upon several aspects of our lives, however, there is no theory-driven empirical research in the information systems literature tackling the adoption issues in this context from a behavioural and social perspective. This study contributes to the body of knowledge by exploring the behavioural and social factors affecting users' decisions to adopt SNSs as new technology.

The present study has shown the importance of social presence’s factors, namely co-presence, intimacy, and immediacy, in explaining individuals’ intentions and behavior. Prior to the current study, only limited number of research studies examined the role of social presence in technology adoption (e.g. Xu, et al., [80]), but not in depth as in this research.

In this study the researcher found that social presence positively influence SNS usage indirectly through user attitude, and as aforementioned that three factors, namely co-presence, intimacy, and immediacy are framing the construct of social presence. Thus, this study is the first empirical effort to examine the impact of co-presence, intimacy, and immediacy in determining intention or behaviour.

The study integrated theoretical model lends itself to studying the adoption of new technologies and applies it to determine significant factors that influence adoption of SNSs. The study’ proposed model brings together concepts from two distinct lines of research, the Decomposed Theory of Planned Behaviour (DTPB) and social presence, as an attempt to build a more comprehensive model with a competitive ability to explain both technology adoption behaviour and behavioural intention.

As highlighted in the previous section, the study model explains 59% of the variance in “Behavioral Intention” and 69% of individual’s attitude or feelings towards adoption the SNSs. Moreover, the proposed models also able to explain 54% of SNSs usage behavior. This ability relates to the diversity of the model's constructs and the diversity of relations among their constructs.

B. Implications for practice

The study found that self-efficacy construct indirectly influence usage behaviour through its direct effect on behavioural intention. This indicates that SNSs users, who are confident of their abilities to use Internet and SNS' sites are more likely to adopt such services. This suggests that SNSs owners should develop effective strategies that take into account these differing levels of abilities by re-building policies and regulations for the sake of supporting users on the long run to help in increasing individuals' ability to use SNSs.

In addition to importance of self-efficacy, the study' findings also confirm that "perceived Enjoyment" ($\beta = 0.280$, Standardized path coefficient, $p < 0.05$), and "Perceived ease of use" ($\beta = 0.172$, Standardized path coefficient, $p < 0.05$) are significantly and positively related to "Attitude" (adjusted R2=0.69), thus, designing sites perceived to be easy to navigate could affect attitudes toward the site and can positively influence confidence levels. Moreover, SNSs should be designed also to be simple, user-friendly and providing users with enjoyable and pleasant experiences. SNSs used to be connected through desktop computers however it should be able to work through smart phones regardless of the operating system, thus service providers, and developers should take this design issues into their account.

VII. Conclusions

This study examines individuals' intentions and behaviour on Social Networking Sites (SNSs), from a social and behavioural perspective. The study proposed model brings together concepts from two distinct lines of research, the Decomposed Theory of Planned Behaviour (DTPB) from IS models and social presence theory from the social psychological theories of interpersonal communication and symbolic interactionism as an attempt to build a more comprehensive model with a competitive ability to explain both technology adoption behaviour and behavioural intention. The proposed model shows that usage behaviour on online social networking services (SNSs) is determined by "behavioral intention" which in turn is determined by individuals' "attitude", "subjective norms", and "self-efficacy". The developed model also asserts that "Co-presence", "Intimacy", "Immediacy", "Perceived Enjoyment", and "Perceived ease of use" formed individuals' "Attitude" towards "behavioral intention" to use online social networking services (SNSs) . The results support all formulated hypotheses. The proposed model in this study explains 69% of individual's attitude or feelings towards adoption the SNSs, 59% of the variance in "Behavioral Intention" and 54% of the variance in "Usage Behaviour".

The present study has shown the importance of social presence's factors, namely co-presence, intimacy, and immediacy, in explaining individuals' intentions and behavior. Prior to this study, only limited number of research studies examined the role of social presence in technology adoption, but not in depth as in this research. In this study we found that social presence positively influence SNS usage indirectly through user attitude, and as we aforementioned that three factors, namely co-presence, intimacy, and immediacy are framing the construct of social presence. Thus, this study is the first empirical effort to examine the impact of co-presence, intimacy, and immediacy in determining intention or behaviour.

References

[1] Agarwal, R. & Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology use. *MIS Quarterly*, 24(4), 665-694.

[2] Ajzen, I. (1991). The theory of planned behaviour. *Organizational Behaviour and Human Decision Processes*, 50(2), 179-211.

[3] Ajzen, I., & Fishbein, M. (1980). *Understanding Attitudes and Predicting Social Behaviour*. Englewood Cliffs, NJ: Prentice-Hall.

[4] Al-Debei, M., Al-Lozi, E., & Papazafeiropoulou, A. (2013). Why people keep coming back to Facebook: Explaining and predicting continuance participation from an extended theory of planned behaviour perspective. *Decision Support Systems*, 55(1), 43-54.

[5] Alghaith, W., Sanzogni, L., & Sandhu, K. (2010). Factors Influencing the Adoption and Usage of Online Services in Saudi Arabia. *Electronic Journal of Information Systems in Developing Countries (EJISDC)*, 40(1), 1-32.

[6] Argyle, M. (1969). *Social interaction*. New York: Atherton Press.

[7] Argyle, M. (1975). The syntaxes of bodily communication. In J. Benthal & T. Polhemus (Eds.), *The body as a medium of expression* (pp. 143–161). New York: E. P. Dutton & Co.

[8] Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

[9] Argyle, M., & Dean, J. (1965). Eye-contact, distance and affliation. *Sociometry*, 28, 289–304.

[10] Asher, H.B. (1983).*Causal modeling*. Newbury Park: Sage University Press.

[11] Atkinson, M.A., & Kydd, C. (1997). Individual characteristics associated with World Wide Web use: an empirical study of playfulness and motivation. *The DATA BASE for Advances in Information Systems*, 28(2), 53-62.

[12] Tabachnick, B., & Fidell, L. (2007). *Using Multivariate Statistics*. 5th ed., Boston: Allyn and Bacon.

[13] Bandura, A. (1997). Self-efficacy: toward a unifying theory of behavioural change, *Psychol Rev.*, 84(2), 191-215.

[14] Bhattacherjee, A., & Premkumar, G. (2004). Understanding changes in belief and attitude toward information technology usage: a theoreticalmodel and longitudinal test. *MIS Quarterly*, 28(2).

[15] Biocca, F., Harms, C., & Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: review and suggested criteria. *Presence: Teleoperators and Virtual Environments*, 12(5), 456–480.

[16] Burgoon, J. K., Bonito, J. A., Ramirez, A., Kam, K., Dunbar, N., & Fischer, J. (2002). Testing the interactivity principle: Effects of mediation, propinquity, and verbal and nonverbal modalities in interpersonal interaction. *Journal of Communication*, 52, 657–677.

[17] Burgoon, J., Bonito, J., Bengtsson, B., Ramirez, A., Jr., Dunbar, N., & Miczo, N. (2000). Testing the interactivity model: Communication processes, partner assessments, and the quality of collaborative work. *Journal of Management Information Systems*, *16*, 33–56.

[18] Burton-Jones, A., & Gallivan, M. (2007). Toward a Deeper Understanding of System Usage in Organizations: A Multilevel Perspective. *MIS Quarterly*, *31*(4), 657-679.

[19] Chen, A., Lu, Y., Chau, P. Y., & Gupta, S. (2014). Classifying, Measuring, and Predicting Users' Overall Active Behavior on Social Networking Sites. *Journal Of Management Information Systems*, *31*(3), 213-253.

[20] Chen, L., Gillenson, L. M., & Sherrell, L. D. (2004). Consumer acceptance of virtual stores: a theoretical model and critical success factors for virtual stores. *ACM SIGMIS Database*, 35(2), 8-31.

[21] Cheung, C., & Lee, M. (2010). A theoretical model of intentional social action in online social networks, *Decision Support Systems*, *49*(1), 24–30.

[22] Chung, J., & Tan, F. B. (2004). Antecedents of perceived playfulness: An exploratory study on user acceptance of general information-searching websites. *Information and Management*, 41, 869-881.

[23] Compeau, D.R., & Huff, S. (1999). Social cognitive theory and individual reactions to computing technology: a longitudinal study. *MIS Quarterly*, 23(2), 145-58.

[24] Compeau, D.R., & Higgins, C.A. (1995). Computer self-efficacy: development of a measure and initial test. *MIS Quarterly*, 19 (2), 189-211.

[25] Comrey, A.L., & Lee, H.B. (1992). *A first course in factor analysis*. N.J.: L. Erlbaum Associates.

[26] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319-340.

[27] Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35, 982-1003.

[28] Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1992). Extrinsic and intrinsic motivation to use computers in the workplace. *Journal of Applied Social Psychology, 22*(14), 1111–1132.

[29] De Guinea, A., & Markus, M. (2009). Why break the habit of a lifetime? rethinking the roles of intention, habit, and emotion in continuing information technology use. *MIS Quarterly*, 33(3), 433.

[30] Deutsch, M., & Gerard, H. (1995). A study of normative and informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology*, 51, 624-36.

[31] Dholakia, U., Bagozzi, R., & Pearo, L. (2004). A social influence model of consumer participation in network- and small-group-based virtual communities. *International Journal of Research in Marketing, 21*(3), 241-263.

[32] Dinev, T., & Hart, P. (2006). Internet privacy concerns and social awareness as determinants of intention to transact. *International Journal of Electronic Commerce*, 10(2), 7-29.

[33] Durlach, N., & Slater, M. (2000). Presence in shared virtual environments and virtual togetherness. *Presence: Teleoperators and Virtual Environments*, 9(2), 214–217.

[34] Facebook, Company info. (2015, July 7). Retrieved from http://newsroom.fb.com/company-info/

[35] Gardner, C., & Amoroso, D. (2004). Development of an Instrument to Measure the Acceptance of Internet Technology by Consumers. *HICSS*, 8, 80260c, Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 8.

[36] Hair, J., Anderson, R., Tatham, R., & Black, W. (1998). *Multivariate Data Analysis*. Upper Saddle River, NJ: Prentice-Hall.

[37] Hassanein, K., & Head, M. (2005). The Impact of Infusing Social Presence in the Web Interface: An Investigation Across Product Types. *International Journal Of Electronic Commerce*, 10(2), 31-55.

[38] Heijden, D. H. V. (2004). User acceptance of hedonic information systems. *MIS Quarterly*, 28(4), 695–704.

[39] Hernandez, M., & Mazzon, J. (2007). Adoption of internet banking: proposition and implementation of an integrated methodology approach. *The International Journal of Bank Marketing*, *25*(2), 72-88.

[40] Hill, T., Smith, N. D., & Mann, M. F. (1987). Role of efficacy expectations in predicting the decision to use advanced technologies: the case of computers. *Journal of Applied Psychology*, *72*(2), 307– 313.

[41] Hsu, M., & Chiu, C. (2004). Internet self-efficacy and electronic service acceptance. *Decision Support Systems*, *38*(3), 369-381.

[42] Hwang, Y. (2008). A Preliminary Examination of the Factors for Knowledge Sharing in Technology Mediated Learning. *Journal of Information Systems Education*, *19*(4), 419-429.

[43] Hwang, Y. (2005). Investigating enterprise systems adoption: Uncertainty avoidance, intrinsic motivation, and the technology acceptance model. European *Journal of Information Systems, 14*(2), 150-161

[44] Igbaria, & Iivari, M. J. (1995). The effects of self-efficacy on computer usage. *Omega,* 23 (6), 587– 605.

[45] Kane, G. P. (2014). WHAT'S DIFFERENT ABOUT SOCIAL MEDIA NETWORKS? A FRAMEWORK AND RESEARCH AGENDA. *MIS Quarterly*, 38(1), 275-304.

[46] Kim, S., & Malhotra, N. K. (2005). A Longitudinal Model of Continued IS Use: An Integrative View of Four Mechanisms Underlying Postadoption Phenomena. *Management Science*, 51(5), 741-755.

[47] Ku, Y.C., Chen, R., Zhang, H. (2013). Why do users continue using social networking sites? An exploratory study of members in the United States and Taiwan. *Information & Management, 50*(7), 571-581.

[48] Lee, J. (2003). Factors affecting intention to use online financial services. Ph.D. dissertation, The Ohio State University, United States -- Ohio.

[49] Lee, M. (2009). Factors influencing the adoption of internet banking: An integration of TAM and TPB with perceived risk and perceived benefit. *Electronic Commerce Research and Applications*, *8*(3), 130-141.

[50] Lee, Y. C. (2006). An empirical investigation into factors influencing the adoption of an e-learning system. *Online Information Review*, *30*(5), 517-541.

[51] Lin, H., Fan, W., & Chau, P. Y. K. (2014). Determinants of users' continuance of social networking sites: A self-regulation perspective. *Information & Management, 51*(5), 595.

[52] Luarn, P., & Lin, H. (2004). Toward an understanding of behavioural intention to use mobile banking. *Computers in Human behaviour*, 1-19.

[53] Ma, Q., & Liu, L. (2004). The technology acceptance model: A meta-analysis of empirical findings. *Journal of Organizational and End User Computing*, 16(1), 59-72.

[54] Mehrabian, A. (1972). *Nonverbal communication*. Chicago: Aldine Atherton.

[55] Meng, Z., Zuo, M., & Chen, Y. (2009). Which Instant Messaging System Should I Choose: a Conceptual Model. *International Journal of Hybrid Information Technology*, 2(1).

[56] Montesarchio, C. (2009). Factors influencing the unethical behavioral intention of college business students: Theory of planned behavior. Ph.D. dissertation, Lynn University, United States -- Florida.

[57] Moore, C., & Benbasat, I. (2001). Development of an instrument to measure the perception of adopting an information technology innovation. *Information Systems Research*, 2, 192-222.

[58] Muhlbach, L. M., & Prussog, A. (1995). Telepresence in videocommunications: A study on stereoscopy and individual eye contact. *Human Factors*, 37(2), 290–305.

[59] Novak, T., Hoffman, D., & Yung, Y. (2000). Measuring the customer experience in online environments: A structural modelling approach. *Marketing Science*, 19(1), 22-42.

[60] Park, J. (2003). Understanding consumer intention to shop online: A model comparison. Ph.D. dissertation, University of Missouri - Columbia, United States -- Missouri.

[61] Pikkarainen, T., Pikkarainen, K., Karjaluoto, H., & Pahnila, S. (2004). Consumer acceptance of online banking: an extension of the technology acceptance model. *Internet Research*, 14(3), 224-235.

[62] Podder, B. (2005). Factors Influencing the Adoption and Usage of Internet Banking: A New Zealand Perspective. Master thesis, Auckland University of Technology, New Zealand.

[63] Premkumar, G., & Ramamurthy, K. (1995). The role of Interorganizational and organizational factors of the decision mode for adoption of interorganizational systems. *Decision Science*, 26(3), 303-336.

[64] Ranganathan, C. & Jha S. (2007). Examining Online Purchase Intentions in B2C E-Commerce: Testing an Integrated Model. *Information Resources Management Journal*, 20(4), 48-64.

[65] Rice, R. E. (1993). Media appropriateness: Using social presence theory to compare traditional and new organizational media. *Human Communication Research*, 19, 451–484.

[66] Rouibah, K. (2008). Social usage of instant messaging by individuals outside the workplace in Kuwait: A structural equation model. *Information Technology & People*, 21(1), 34-68.

[67] Schroeder, R. (2002). Copresence and interaction in virtual environments: An overview of the range of issues. *Conference Proceedings of the 5th Annual International Workshop: Presence* 2002, 274 –295.

[68] Sheppard, B., Hartwick, J., & Warshaw, P. (1988). The theory of reasoned action: a meta-analysis of past research with recommendations for modifications and future research. *Journal of Consumer Research,* 15(3), 325–343.

[69] Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. London: John Wiley & Sons.

[70] Sidorova, A., Evangelopoulos, N., Valacich, J. S., & Ramakrishnan, T. (2008). Uncovering the intellectual core of the information systems discipline. *MIS Quarterly, 32*(3), 467-A20.

[71] Slater, M., Sadagic, A., & Schroeder, R. (2000). Small-group behavior in a virtual and real environment: A comparative study. *Presence: Teleoperators and Virtual Environments*, 9(1), 37–51.

[72] Sternberg, R. J. (1997). Construct validation of a triangular love scale. *European Journal of Social Psychology*, 27(3), 313–335.

[73] Sun, H., & Zhang, P. (2006). The role of affect in IS research: A critical survey and a research model, in HCI in MIS (I): Foundations, Zhang, P. & Galletta, D. (eds.) *Series of Advances in Management Information Systems, M.E. Sharpe Publisher*, 121-142.

[74] Taylor, S., & Todd, P.A. (1995). Understanding information technology usage: A test of competing models. *Information Systems Research*, 6(2), 144-176.

[75] Tu, C. H., & McIsaac, M. S. (2002). The relationship of social presence and interaction in online classes. *The American Journal of Distance Education*, 16(3), 131–150.

[76] Venkatesh, V., Brown, S., Maruping, L., & Bala, H. (2008). Predicting Different Conceptualizations of System Use: The Competing Roles of Behavioral Intention, Facilitating Conditions, and Behavioral Expectation. *MIS Quarterly,* 32(3), 483.

[77] Wang, Y., Shun, Wang, Y., Min, Lin, H., Hsin, & Tang, T., I. (2003). Determinats of user acceptance of Internet banking: an empirical study. *International journal of service Industry management*, 14(5), 501-519.

[78] Webster, J. & Ho, H. (1997). Audience engagement in multi-media presentation. *Data Base for the Advances in Information Systems*, 28(2), 63-77.

[79] Wei, C., Chen, N., & Kinshuk. (2012). A model for social presence in online classrooms. *Education Tech Research Dev*, 60, 529-545.

[80] Xu, C., Ryan, S., Prybutok, V., & Wen, C. (2012). It is not for fun: An examination of social network site usage. *Information and Management*, 49(5), 210–217.

[81] Yi, M. Y., & Hwang, Y. (2003). Predicting the use of web-based information systems: Self-efficacy, enjoyment, learning goal orientation, and the technology acceptance model. International *Journal of Human-Computer Studies*, 59(4) 431-449.

[82] Zhang, H., Lu, Y., Gupta, S., & Zhao, L. (2014). What motivates customers to participate in social commerce? The impact of technological environments and virtual customer experiences, *Information and Management*, *51*(8), 1017–103.

[83] Zhang, J. (2009). Exploring Drivers in the Adoption of Mobile Commerce in China. *Journal of American Academy of Business*, Cambridge, 15(1), 64-69.

[84] Zhang, Q., & Oetzel, J. G. (2006). Constructing and validating a teacher immediacy scale: a Chinese perspective. *Communication Education*, 55(2), 218–241.

[85] Zikmund, W. G. (2003). *Business research methods* (7th ed.). Cincinnati, OH: Thomson.

[86] Zmud, R., Shaft, T., Zheng, W., & Croes, H. (2010). Systematic Differences in Firm's Information Technology Signaling: Implications for Research Design. *Journal of the Association for Information Systems*, 11(3), 149-181.

Trust: A Requirement for Cloud Technology Adoption

Akinwale O. Akinwunmi
Bowen University / Computer Science and Information Technology Department, Iwo, Nigeria

Emmanuel A. Olajubu
Obafemi Awolowo University, Computer Science & Engineering Department, Ile-Ife, Nigeria

G. Adesola Aderounmu
Obafemi Awolowo University, Computer Science & Engineering Department, Ile-Ife, Nigeria

***Abstract*—Cloud computing is a recent model for enabling convenient, on-demand network access to a shared pool of configurable computing resources such as networks, servers, storage, applications, and services; that can be rapidly provisioned and released with minimal management effort or service provider interaction. Studies have shown that cloud computing has the potential to benefit establishments, industries, national and international economies. Despite the enormous benefits cloud computing technology has the potentials of offering, several issues are making intended users to pause in adopting the usage of the technology. Users need to be assured of the safety and reliability of the technology while using it. This is needed to build confidence around the technology and reduce the level of anxiety. This research attempts to investigate the effect of trust in the adoption of the technology by formulating a trust model based on Expectancy Disconfirmation Theory model and Bayesian network. A simulation experiment was carried out to determine the significance of trust in the adoption of cloud technology.**

Keywords—Cloud; User; Adoption; Trust; Bayesian Network

I. INTRODUCTION

A cloud is a type of parallel and distributed system consisting of a collection of inter-connected and virtualised computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers [1]. Studies have shown that cloud computing has the potential to benefit establishments, industries, national and international economies [2, 3]. It can greatly accelerate the way companies create new products and services, making it possible for product development professionals around the world to interconnect and collaborate more effectively and gain access to more powerful and economical computer resources[2]; Increasing the ability of organizations to mine their data for important trend information, such as customers' changing needs and competitors moves in the market[2]; Levelling the playing field between large and small companies by giving companies of all sizes access to information technology that previously was affordable for only the largest of companies[2]; and Helping emerging economies leapfrog to higher levels of technological development by providing more immediate and affordable access to next- generation applications, tools, and infrastructure [2,4]. Cloud computing delivers infrastructure, platform, and software that are made available as subscription-based services in a pay-as-you-go model to consumers [5].

These services are referred to as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) in industries [6]. The importance of these services available from the long-held dream of computing as a utility has the potential to transform a large part of the IT industry, making software even more attractive as a service [7]. Cloud computing has been the new paradigm in distributed computing in the recent times. Cloud computing has been hailed as the 5th utility after electricity, gas, water and telephony as it promises to make the computing services available anywhere, any time and pay only for what is used or consumed [8].

Despite the enormous benefits cloud computing technology has the potentials of offering, several issues are making intended users to pause in adopting the usage of the technology. Users need to be assured of the safety and reliability of the technology while using it, this is required to build confidence around the technology. A new technology must gradually build its reputation for good performance and security, earning users' trust over time [9]. Most potential users are skeptical about the trustworthiness of cloud technology and this has been affecting the rate of its adoption. There is no way for the customer to be sure whether the management of the service is trustworthy, and whether there is any hidden risk [10]. This research attempts to investigate the effect of trust on the adoption of the cloud technology.

II. RELATED WORKS

Trust can be explained in diverse fields such as psychology, sociology, and economics. Also trust can be classified based on different meaning by many writers. Trust has been the focus of researchers for a long time [11], from the social sciences, where trust between humans has been studied to the effects of trust in economic transactions [12,13,11]. Although intuitively easy to comprehend, the notion of trust has not been formally defined [14]. Control is another important issue in trust. A system is trusted less when it is difficult to have the control over the assets in its custody. Trustworthiness plays an important role in service selection and usage of cloud services. Trust in cloud computing is related more to preventing a trust violation than to guaranteeing compensation should a violation occur. For most enterprises, a security breach of data is irreparable; no amount of money can guarantee to restore the lost data or the enterprise's reputation. The cloud computing trust model thus must focus more on preventing failure than on post-failure compensation [15]. A new technology must gradually build its reputation for good performance and security, earning users' trust over time [15]. Trust is a derivation of the reputation of an entity [14]. Based on the reputation, a level of trust is bestowed upon an entity [14].

Several leading research groups both in academia and the industry are working in the area of trust management in cloud computing. Dingguo *et al*. [16] proposed a trust cloud-based subjective trust assessment and management model. The model provided the design of trust cloud, the policy of the obtainment to compute the trust information and supplied trust decision based on trust cloud model [17]. Hada *et al*. [18] proposed a trust model for cloud architecture which uses mobile agent as security agents to acquire useful information from the virtual machine which the user and service provider can utilize to keep track of privacy of their data and virtual machines [18]. Combining quality of service (QoS) with trust model, Li *et al*. [19] constructed a QoS-aware and quantitative trust-model that consists of initial trust value [19], direct trust value, and recommendatory trust value of service, making the provision, discovery [19], and aggregation of cloud services trustworthy [19]. Khan and Malluhi [15] have looked at the trust in the cloud system from users' perspective [20]. They analyze the issues of trust from what a cloud user would expect with respect to their data in terms of security and privacy [20]. They further discuss that what kind of strategy the service providers may undertake to enhance the trust of the user in cloud services and provide r[20]. Sato *et al*. [21] have proposed a trust model of cloud security in terms of social security. The authors have identified and named the specific security issue as social insecurity problem and try to handle it using a three pronged approach [21]. The family gene based cloud trust model that is fundamentally different from the Public key Infrastructure based trust models have been proposed by several researchers [22,23]. These researchers have studied the basic operations such as user authentication, authorization management and access control, and proposed a Family-gene Based model for Cloud Trust (FBCT) integrating these operations [20].

Manuel *et al.* [24] have proposed trust model that is integrated with CARE (Center for Advanced Computing Research and Education) resource broker. This trust model can support both grid and cloud systems [24]. The model computes trust using three main components namely, Security Level Evaluator, Feedback Evaluator and Reputation Trust Evaluator [20]. Both Shen *et al.* [25] and Shen and Tong [26] have analyzed the security of cloud computing environment and described the function of trusted computing platform in cloud computing.

Alhamad *et al*. [27] have proposed a SLA (Service Level Agreement) based trust model for cloud computing. The model consists of the SLA agents, cloud consumer module, and cloud services directory. A model called a multi-tenancy trusted computing environment model (MTCEM) for cloud computing has been proposed by Li *et al*. [28]. MTCEM has been proposed to deliver trusted IaaS to customers with a dual level transitive trust mechanism that supports a security duty separation function simultaneously [28]. Fu *et al*. [29] have studied the security issues associated with software running in the cloud and proposed a watermark-aware trusted running environment to protect the software running in the cloud. Ranchal *et al*.[30] have studied the identity management in cloud computing and proposed a system without the involvement of a trusted third party. Takabi *et al*. [31] have proposed a security framework for cloud computing consisting of different modules to handle security, and trust issues of key components. Noor and Sheng [32] proposed the "Trust as a Service" (TaaS) framework to improve ways on trust management in cloud environments. In particular, an adaptive credibility model that distinguishes between credible trust feedbacks and malicious feedbacks by considering cloud service consumers' capability and majority consensus of their feedbacks was introduced [32]. Reputation-based trust is emerging as a good choice to model trust of cloud service providers based on available evidence [32]. Many existing reputation based systems either ignore or give less importance to uncertainty linked with the evidence [32]. Pawar *et al.* [33] proposed an uncertainty model and define the approach to compute opinion for cloud service providers. Using subjective logic operators along with the computed opinion values, mechanisms to calculate the reputation of cloud service providers were proposed [33].

In order to protect the security of cloud entities and better practice cloud's objectives of providing low-cost and on-demand services, Li *et al.* [34] proposed a novel cloud trust transaction framework and also a new trust fuzzy comprehensive evaluation based cloud service discovery algorithm. Prajapati *et al.*[35] presented a formal trust management model based on the basics of the trust characteristics. The proposed model was capable to handle various cloud services access scenarios where entity has a past experience with the service or a stranger entity requesting to access the service without any identity or past interaction with the service [35]. The work defined the direct trust with a time-variant evaluation method and the recommended trust with a space variant evaluation method. Motivated by human nature, the model also has considered the reputation factor of trustor to calculate the direct trust [35] . The proposed approach also has used the satisfaction level to calculate recommended trust which is depends on service level agreements of the services resides in the cloud environment [35]. Trust has attracted extensive attention in social science and computer science as a solution to enhance the security of the system. Wu *et al.* [36] proposed a trust evaluation model based on D-S evidence theory and sliding windows for cloud computing. The timeliness of the interaction evidence as the first-hand evidence is reflected by introducing sliding windows [36]. In an open and dynamic environment of distributed system such as cloud technology trust plays a major role in determining the level of adoption of the technology. Trust is one of the most concerned obstacles for the adoption and growth of cloud computing.

Threats, risks and other security concerns are impeding the move to cloud environment by individuals and organizations. Due to the transfer of substantial part of control of activities to a third party concerns are generated about the trustworthiness of the technology. Most of the efforts at making cloud technology trustworthy are still at infant stage.

III. THEORETICAL BACKGROUND

The following concepts provide the theoretical basis for the solution proposed in this work with a view to achieving reliable and trustworthy cloud technology adoption model. These concepts are Expectancy Disconfirmation Theory (EDT) and Bayes' Theorem. The capability of Bayesian network to handle the complexity of the dynamism of the cloud environment and the perceived influence of service satisfaction in cloud computing usage informed the idea of conceptualizing a trust model based on the EDT Model and Bayesian Network.

Expectancy Disconfirmation Theory (EDT) is a theory for measuring customer satisfaction from perceived quality of products or services. EDT is a prominent theory from marketing that can predict and explain consumers' satisfaction with products or services. EDT has been applied most often in IT adoption or IT usage studies [37,38,39,40].

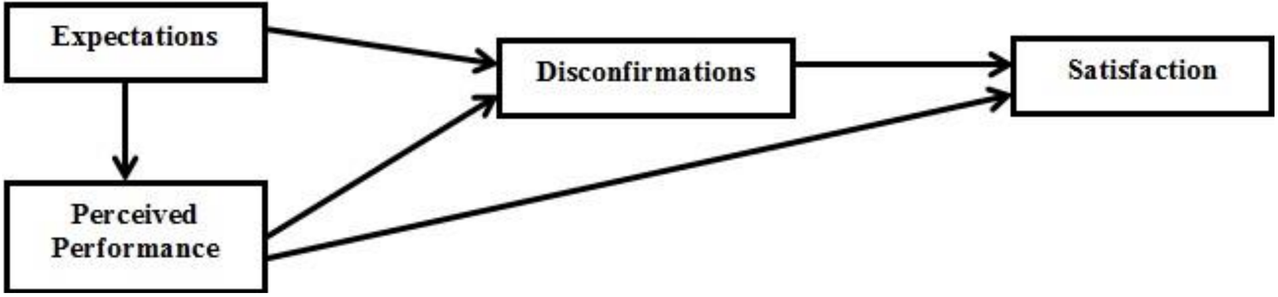


Fig. 1. Expectancy Disconfirmation Theory model [44]

The theory argues that 'satisfaction is related to the size and direction of the disconfirmation experience that occurs as a result of comparing service performance against expectations' [41]. It is a judgment that a product or service feature, or the product or service itself, provided (or is providing) a pleasurable level of expectation. This model consists of four components: expectations, perceived performance, disconfirmation, and satisfaction as shown in Fig. 1.

Expectations define the customer's anticipations about performance of products and services [42]. **Perceived performance** investigates the customer's experience after using products or services that can be better or worse than customer's expectation [43]. **Disconfirmation** is defined as the difference between the customer's initial expectation and observed actual performance [39].

The theory proposes that users first form expectations or belief probabilities of attribute occurrence. They then form post-usage perceptions about performance and a comparison between initial expectations and performance known as disconfirmation of expectations [44]. A positive disconfirmation means performance was better than expected, and a negative disconfirmation means performance was worse than expected. According to EDT, the better performance is, or the more positive the disconfirmation, the greater the satisfaction [44].

Bayes' Theorem is a mathematical formula used for calculating conditional probabilities. It figures prominently in *subjectivist* or *Bayesian* approaches to epistemology, statistics, and inductive logic. Subjectivists, who maintain that rational belief is governed by the laws of probability, lean heavily on conditional probabilities in their theories of evidence and their models of empirical learning [45]. Bayes' Theorem is central to these enterprises both because it simplifies the calculation of conditional probabilities and because it clarifies significant features of subjectivist position [45]. Bayesian Networks (BNs) provide a method for representing relationships between variables (called 'nodes' in the BN) even if the relationships involve uncertainty. They can be a useful modeling tool in situations where different types of variables and knowledge from various sources need to be integrated within a single framework [46, 47].

IV. PROPOSED MODEL

The proposed model is based on the marriage of EDT model and Bayesian network. The end result of being satisfied with a technology usage is that trust will be built around the technology and this will influence potential users. Hence EDT model's components are transformed into a Bayesian network nodes ending with a trust node and the probabilities dependencies among the various components are depicted in Fig. 2.

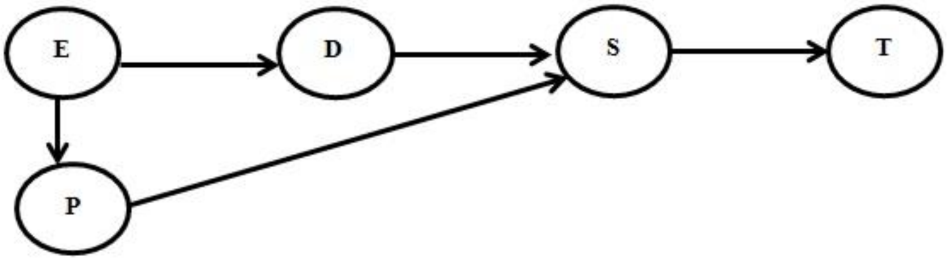


Fig. 2. Proposed Model Bayesian Network

The notation used for the model variables are as follows: E stands for expectation, P for perceived performance, D stands for disconfirmation, S stands for satisfaction and T stands for trust as shown in Fig. 2.

The model is formulated using the chain rules of conditional probability, the full joint probability distribution for the model is written as a product of the individual density functions, conditional on their parent variables as depicted in equation 1.

P(T,S,P,D,E)= P(T|S)*P(S|D,P)*P(P|E)*P(D|E)*P(E) (1)

Now suppose we use a cloud service with S level of satisfaction. Let S(T) be the event that the cloud service is trusted. Let P(S) be the event that the cloud service is satisfactory. The events S(T), that the cloud service is trusted, and S($\overline{T}$), that the service is not trusted, partitioning the set of all cloud services. Hence, by Bayes' Theorem, the probability that the cloud service is trusted, given that it is satisfactory is shown in equation (2).

$$P(T|S) = \frac{P(S \mid T)P(T)}{P(T \mid S)P(S) + P(T \mid \overline{S})P(\overline{S})} \qquad (2)$$

To apply this formula, we first estimate $P(S)$, the probability that the cloud service is satisfactory, as well as $P(\overline{S})$, the probability that the cloud service is not satisfactory. Without prior knowledge about the likelihood that the cloud service is not satisfactory, simplicity we assume that the cloud service is equally likely to be satisfactory as it is not to be satisfactory. That is, we assume that $P(S)$,= $P(\overline{S})$,= 1/2. Using this assumption, we find that the probability that the cloud service is equally likely to be satisfactory, is stated in equation (3)

$$P(T|S) = \frac{P(S \mid T)}{P(S \mid T) + P(S \mid \overline{T})} \qquad (3)$$

Determining a specific model, T, that best accounts for all the variations of cloud service usage can be accomplished by maximizing the level of cloud service satisfaction, S which according to Bayes' rule is stated in equation 4.

$$P(T|S) \alpha P(S|T)P(T) \qquad (4)$$

Determining the prior, $P(T|S)$, is full of uncertainty, then the model that maximizes the likelihood, $L(S|T)$ is chosen. The likelihood is proportional to the probability of observing the model T, treating the level of cloud service satisfaction, S as variables and the model T as fixed. The best estimator $\hat{S}$, is whatever value of $\hat{S}$ that maximizes the likelihood is shown in equation (5).

$$L(\hat{S}|T) = P(T|\hat{S}) \qquad (5)$$

Typically the parameter $\hat{S}$ maximized the likelihood of observing the model T.

Based on the proportional relationship as expressed in equation (5), the $\hat{S}$ that maximizes $L(\hat{S}|T)$ will also maximize $P(T|\hat{S})$ which is the probability of the observed model. $\hat{S}$ denotes the best satisfaction parameter for the model T.

The likelihood function in equation (5) then is expressed as log likelihood function as shown in equation (6) log likelihood expression

$$InL = InP(T|\hat{S}) \qquad (6)$$

V. Simulation, Result and Discussion

Experiments were conducted to investigate the effect of trust in cloud environment using transaction's response time as the evaluation criterion in two different settings. The response time in this regard refers to time lag between the instant of service request by the users to the instant of having the result of the service request. This depends on round trip time (RTT) and the users' load.

The experiment looked at the cloud transactions carried out without involving trust factor against the same kind of cloud transaction involving trust factor in order to investigate their difference. CloudAnalyst [48] useful to model and analyze large scale cloud computing development was used for the simulation. Considering the description and behaviour of throttled load balancing policy of the CloudAnalyst, the integration of trust capability in cloud environment was achieved using this policy while the round-robin load balancing policy, considering its behaviour was used for the cloud transaction without trust integration.

The first setting simulated cloud service transaction platform using CloudAnalyst. In this experiment there are 5,10,15,20, 25 and 30 groups of users with a cloud provider. These group of users were generated randomly using Poisson distribution by varying the number of users in a realistic manner.

The detailed parameter settings for the experiment are shown in Table 1. In this experiment, the numbers of users were varied while the number of the provider was fixed in order to observe the behaviour of the cloud environment without trust integration and cloud environment with trust integration in terms of the transaction's response time.

Table 2 shows the results obtained from the simulation. The graph in Fig. 3 shows the plot of the round trip time against the number of group of users. From the results, the cloud environment with trust integration ensures relatively quick response time than the cloud environment without trust integration with slow response time. This is evident from the percentage difference of the response time as shown in Table 2. It was observed that integration of trust into the cloud environment is accountable for the quick response time achieved in the simulated cloud transaction with trust integration.

TABLE I. SIMULATION PARAMETER SETTINGS FOR RESPONSE TIME WITH VARIED NUMBER OF USERS AND FIXED NUMBER OF PROVIDERS

Users	Provider	User Growth Factor	Request Growth Factor	Execution Instruction Per Length
5	1	10	10	100
10	1	10	10	100
15	1	10	10	100
20	1	10	10	100
25	1	10	10	100
30	1	10	10	100

TABLE II. RESPONSE TIME WITH VARIED NUMBER OF USERS AND FIXED NUMBER OF PROVIDERS IN CLOUD ENVIRONMENT WITHOUT TRUST AND WITH TRUST

Number of Users	Round Trip Time (ms) (without trust)	Round Trip Time (ms) (with trust)	Percentage Difference (%)
5	300.49	300.39	-0.03
10	300.40	300.58	0.06
15	300.70	300.16	-0.18
20	300.95	300.25	-0.23
25	300.37	300.23	-0.05
30	300.64	300.27	-0.12

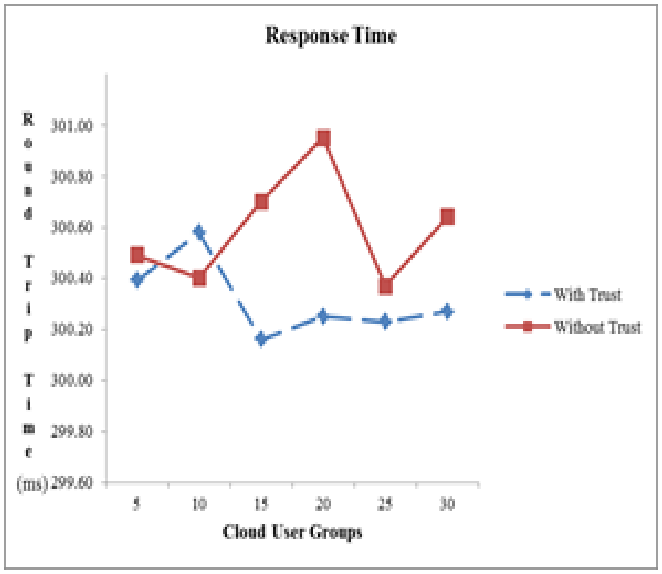


Fig. 3. Response time with varied number of users and fixed number of providers

In the second simulation experiment cloud transaction was also simulated using cloudAnalyst. Also throttled load balancing policy of the CloudAnalyst was used for the integration of trust capability in cloud environment while the round-robin load balancing policy, was used for the cloud transaction without trust integration. In this experiment there are 5,10,15,20, 25and 30 groups of cloud users with a varied number of cloud providers. This group of users was generated randomly using Poisson distribution by varying the number of users in a realistic manner. The detailed parameter settings for the experiment were shown in Table 3. In this experiment, the number of group of users were varied as well as the number of the providers in order to observe the behaviour of the cloud environment without trust integration and with trust integration in terms of the transaction's response time.

TABLE III. SIMULATION PARAMETER SETTINGS FOR RESPONSE TIME WITH VARIED NUMBER OF USERS AND VARIED NUMBER OF PROVIDERS

Users	Provider	User Growth Factor	Request Growth Factor	Execution Instruction Per Length
5	1	10	10	100
10	2	10	10	100
15	3	10	10	100
20	4	10	10	100
25	5	10	10	100
30	6	10	10	100

Table 4 shows the results obtained from the simulation. The graph in Fig.4 shows the plot of the round trip time against the group of users. From the results, cloud environment with trust integration ensures quicker response time than the cloud environment without trust integration with pretty much higher and unstable response time interval. This is evident from the percentage difference of the models' response time as shown in Table 4. It was observed that integration of trust into the cloud transaction is responsible for the quick response time of the cloud environment.

TABLE IV. RESPONSE TIME WITH VARIED NUMBER OF USERS AND VARIED NUMBER OF PROVIDERS IN CLOUD ENVIRONMENT WITHOUT TRUST AND WITH TRUST

Number of Users	Round Trip Time (ms) (without trust)	Round Trip Time (ms) (with trust)	Percentage Difference (%)
5	300.49	300.39	-0.033
10	300.65	300.49	-0.053
15	302.62	300.45	-0.717
20	300.88	300.87	-0.003
25	301.63	300.55	-0.358
30	301.13	300.61	-0.173

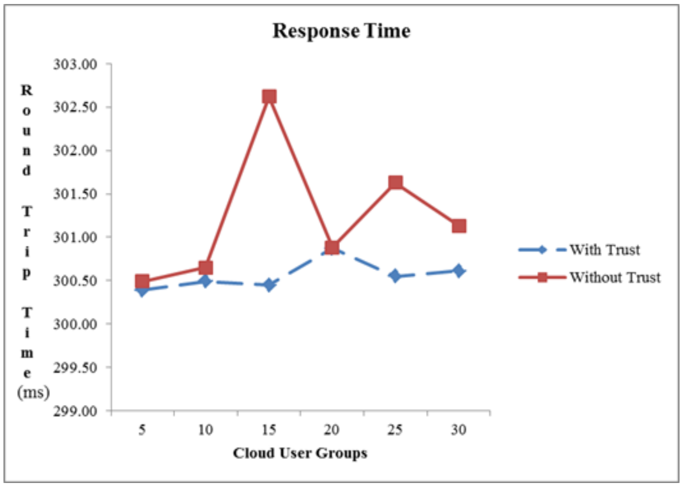


Fig. 4. Response Time with varied number of users and varied number of providers

In order to validate the results of the simulation of the cloud transaction with trust integration and the cloud transaction without trust integration, *t*-test was used. This was used to compare the simulation results [49, 50] for the response time of the cloud transactions in the two different simulation settings to determine if there is any significant difference between them or not as a result of trust integration in the cloud environment. From the simulation results, the cloud transaction with trust performed better than the cloud transaction without trust. In order to validate simulation results, paired difference *t* –test was performed on the percentage difference of the response time with six group sizes in the two different simulation settings for the cloud transaction with trust and the cloud transaction without trust. The Null hypothesis states that there is no significant difference between the cloud transaction with trust and the cloud transaction without trust while the alternative hypothesis states that there is significance difference between them.

The calculated *t* values for the response time in the two different experimental settings are -23.66, and -6.06, respectively as shown in Table 5.

TABLE V. CALCULATED T VALUES FOR THE WITHOUT TRUST AND WITH TRUST

Performance Metrics	**Calculated T Value**
Response Time (Experiment one)	-23.66
Response Time (Experiment two)	-6.06

The degree of freedom for the total group size of six (6) is five (5). Entering a *t* table with 5 degrees of freedom (df), at 95% confidence interval, the table t value is 2.02 (one-tailed, a significance level (α) of 0.05). The absolute calculated *t* values are higher than the tabled *t* value of 2.02 showing that the cloud environment with trust integration with lower response time is significantly different (p=0.05) from the cloud environment without trust integration as a result trust integration in it.

VI. CONCLUSION

The investigation carried out revealed that trust is required for achieving effective cloud technology adoption. The usage of cloud technology can lead to satisfaction, while satisfaction can aid in building trust in the technology and trust will lead to usage continuance intentions. Trust in a cloud service when properly addressed will positively affect cloud technology adoption. In accordance with the EDT model, satisfying the potential cloud users is not limited only to their expectation about cloud service. But also satisfying the cloud users from perceived performance is the first step that can attract the user's trust over offered cloud services. This line of research shows that trust in a cloud service can lead to perceived usefulness or other positive perceptions of the cloud services. Trust will positively influence the intention to use cloud services and the quality of prior experience will positively influence trust in cloud service provider's ability. The extent of prior experience will moderate the relationship between trust and intention to use cloud technology. Trust is a predicated factor needed for achieving higher rate of cloud adoption and facilitate effective usage of the cloud technology. Future work will consider parameters and factors that can help build and sustain trust in cloud environment for aiding fast cloud technology adoption. Developing techniques that demonstrates that the technology can be trust will also be considered.

REFERENCES

[1] Buyya, R., Yeo, C. S. and Venugopal, S., "Market oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities" *In* Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications (HPCC), 2008, Dalian, China.

[2] World Economic Forum, "Exploring The Future of Cloud Computing: Riding The Next Wave of Technology-Driven Transformation", Report of World Economic Forum In Partnership with Accenture 91-93 route de la Capite, 2010, CH-1223 Cologny/Geneva, Switzerland.

[3] Carcary, M., Doherty, E. and Conway, G. "The Adoption of Cloud Computing by Irish SMEs – an Exploratory Study" The Electronic Journal Information Systems Evaluation Vol.17 Issue 1 2014, (003- 014) , available online at www.ejise.com).

[4] John W. Rittinghouse and James F. Ransome Cloud Computing Implementation, Management, and Security CRC Press Taylor & Francis Group 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742 2010 by Taylor and Francis Group LLCCRC Press is an imprint of Taylor & Francis Group, an Informa business

[5] Buyya, R. 2013 Market-oriented cloud computing: Opportunities and challenges Enterprise Distributed Object Computing Conference (EDOC), 2013 17th IEEE International Vancouver, BC, Canada 9-13 Sept. 2013 pg. 3, IEEE

[6] Calheiros,R.N., Ranjan,R., Beloglazov, A., De Rose, C.A.F and Buyya, R. (2011) "CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms". *Software Practice Experience*, 41:23–50.

[7] Armbrust, M., Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I. and Zaharia, M. (2010). A View of Cloud Computing. Communications of the ACM Volume 53, Number 4 (2010), Pages 50-58 , USA.

[8] Buyya,R., Yeo, C., Venugopal, S., Broberg, J. and Brandic, I (2009). Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility. *Future Generation Computer Systems*, 25(6): 599-616.

[9] Khan, K.M. and Malluhi, Q. (2010). Establishing Trust in Cloud Computing, *IT Professional*, 12(5): 20 - 27.

[10] Sehgal, S. (2013) Road Towards Cloud Computing – What are the issues? – Part – I The Official blog of Simplilearn June 13, 2013 available at:// blog.simplilearn.com/it-service- management/road-towards-cloud-computing-what- are-the-issues-part-i

[11] McKnight, D.H., Cummings, L. L. and Chervany, N. L. (1996). Trust Formation in New Organaizational Relationships. MIS Research Center, Carlson School of Management, University of Minnesota.

[12] Ba, S. and Paylou, P. A. (2002). Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behaviour. *MIS Quarterly*, 26:243-268.

[13] Dasgupta, P. (2000) 'Trust as a Commodity', in Gambetta, D. (ed.) *Trust: Making and Breaking Cooperative Relations*, electronic edition, Department of Sociology, University of Oxford.

[14] Momani, M., Aboura, K., and Challa,S. (2007). RBATMWSN: Recursive Bayesian Approach to Trust Management in Wireless Sensor Networks, ISSNIP 2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information Processing Melbourne, Australia, pp. 347-352.

[15] Khan, K.M. and Malluhi, Q. (2010). Establishing Trust in Cloud Computing, *IT Professional*, 12(5): 20 - 27.

[16] Dingguo, Y., Nan, C. and Chengxiang, T. (2011). Research on Trust Cloud-Based Subjective Trust Management Model Under Open Network Environment. *Information Technology Journal*, 10: 759-768.

[17] Silas, S., Rajsingh, E.B. and Ezra, K. (2012). Efficient Service Selection Middleware using ELECTRE Methodology for Cloud Environments. *Information Technology Journal, 11: 868-875.*

[18] Hada, P.S., Singh, R. and Meghwal, M.M. (2011). Security Agents: A Mobile Agent based Trust Model for Cloud Computing. *International Journal of Computer Applications,* 36(12): 12 – 15.

[19] Li, B., Cao, B., Wen, K. and Li, R.(2011). Trustworthy Assurance of Service Interoperation in Cloud Environment. *International Journal of Automation and Computing,* 8(3): 297 – 308.

[20] Firdhous, M.,Ghazali, O. and Hassan, S. (2011). Trust Management in Cloud Computing: A Critical Review International Journal on Advances in ICT for Emerging Regions 2011 04 (02) : 24 - 36 critical review

[21] Sato, H., Kanai, A. and Tanimoto, Z. (2010). "A Cloud Trust Model in a Security Aware Cloud,"in 2010 10th IEEE/IPSJ International Symposium on Applications and the Internet (SAINT), Seoul, South Korea, pp. 121 - 124.

[22] Wang, T.F., Ye, B.S. Li, Y.W. and Yang, Y. (2010a). "Family Gene based Cloud Trust Model," in International Conference on Educational and Network Technology (ICENT), Qinhuangdao, China, pp. 540 - 544.

[23] Wang, T.F., Ye, B.S. Li, Y.W. and Zhu, L.S. (2010b). "Study on Enhancing Performance of Cloud Trust Model with Family Gene Technology," in 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), Chengdu, China, pp. 122 - 126.

[24] Manuel, P.D., Selve, T. and Abd-EI Barr, M.I. (2009). "Trust Management System for Grid and Cloud Resources," in 2009 First International Conference on Advanced Computing (ICAC 2009), December 13 – 15, 2009, Chennai, India, pp. 176 -181.

[25] Shen, Z., Yan, L.L.F. and Wu, X. (2010). "Cloud Computing System Based on Trusted Computing Platform," in 2010 International Conference on Intelligent Computation Technology and Automation (ICICTA), Changsha, China, pp. 942 - 945.

[26] Shen, Z. and Tong, Q. (2010). "The Security of Cloud Computing System Enabled by Trusted Computing Technology," in the 2010 Proceedings of the 2nd International Conference on Signal Processing Systems (ICSPS), Dalian, China, pp. 11-15.

[27] Alhamad, M., Dillon, T. and Chang, E. (2010). 'SLA-based Trust Model for Cloud Computing,' in 13th International Conference on Network-Based Information Systems, Takayama, Japan.

[28] Li, X.Y., Zhou, L.T., Shi, Y. and Guo, Y. (2010). "A Trusted Computing Environment Model in Cloud Architecture", in 2010 Ninth International Conference on Machine Learning and Cybernetics (ICMLC), July 11 - 14, 2010, Qingdao, China, pp. 2843-2848.

[29] Fu, J., Wang, C., Yu, Z., Wang, J. and Sun, J.G. (2010). 'A Watermark-Aware Trusted Running Environment for Software Clouds,' in Fifth Annual ChinaGrid Conference (ChinaGrid), Guangzhou, China, pp. 144 - 151.

[30] Ranchal, R., Bhargava, B., Othmane, L.B., Lilien, L., Kim, A.,Kang, M. and Linderman, M.(2010). "Protection of Identity Information in Cloud Computing without Trusted Third Party," in 29th IEEE International Symposium on Reliable Distributed Systems, New Delhi, India, 2010, pp. 1060-9857.

[31] Takabi, H., Joshi, J.B.D. and Ahn, G.J. (2010). "SecureCloud: Towards a Comprehensive Security Framework for Cloud Computing Environments", in 34th Annual IEEE Computer Software and Applications Conference Workshops, Seoul, South Korea, pp. 393 - 398.

[32] Noor, T.H. and Sheng, Q.Z. (2011). Trust as a Service: A Framework for Trust Management in Cloud Environments School of Computer Science The University of Adelaide, Adelaide SA 5005, Australia

[33] Pawar,P.S., Rajarajan, M., Krishnan Nair, S. and Zisman, A.(2012). Trust Model for Optimized Cloud Services: In Dimitrakos et al. (Eds.): International Federation for Information Processing 2012, AICT 374, pp. 97–112.

[34] Li, W., Ping,L., Qiu,Q. And Zhang, Q. (2012). Research on Trust Management Strategies in Cloud Computing Environment. *Journal of Computational Information Systems,* 8(4): 1757-1763

[35] Prajapati,S.K., Changder, S. and Sarkar, A.(2013). "Trust Management Model for Cloud Computing Environment", Proceedings of the International Conference on Computing, Communication and Advanced Network (ICCCAN 2013), March 15 – 17, India, pp. 1-5

[36] Wu, X., Zhang,R., Zeng, B. and Zhou, S. (2013). A trust evaluation model for cloud computing. International Conference on Information Technology and Quantitative Management (ITQM2013) Published by Elsevier B.V. Procedia Computer Science 17 (2013), pp. 1170 – 1177.

[37] Venkatesh, V. and Goyal, S. (2010). Expectation Disconfirmation and Technology Adoption: Polynomial Modeling and Response Surface Analysis1. *MIS Quarterly* 34(2): 281-303.

[38] Bhattacherjee, A., Perols, J. and Sanford, C. (2008). Information Technology Continuance: a theoretic extension and empirical test. *Journal of Computer Information Systems,* 49(1): 17-26.

[39] Bhattacherjee, A. and G. Premkumar (2004). Understanding changes in belief and attitude toward information technology usage: A theoretical model and longitudinal test. *MIS Quarterly,* 28(2): 229-254.

[40] Susarla, A., Barua, A. and Whinston, A. B. (2003). Understanding the Service Component of Application Service Provision: An Empirical Analysis of Satisfaction with ASP Services. *MIS Quarterly,* 27(1): 91-123.

[41] Ekinci Y. and Sirakaya E. (2004). 'An Examination of the Antecedents and Consequences of Customer Satisfaction'. In: Crouch G.I., Perdue R.R., Timmermans H.J.P., and Uysal M. *Consumer Psychology of Tourism,Hospitality and Leisure*. Cambridge, MA: CABI Publishing, pp. 189-202.

[42] Churchill, G. A., and Surprenant, C. (1982). An investigation into the Determinants of Consumer Satisfaction. *Journal of Marketing Research*, 19: 491–504.

[43] Spreng, R. A., MacKenzie, S. B., and Olshavsky, R. W. (1996). A Reexamination of the Determinants of Consumer Satisfaction. *Journal of Marketing,* 60: 15-32.

[44] Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, 17: 460–469.

[45] Erastus-Obilo, B. (2009). Reason Curve, Jury Competence, and the English Criminal Justice System: The Case for a 21st Century Approach Universal-Publishers

[46] Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference, San Mateo, California, Morgan Kaufmann Publishers.

[47] Jensen, F.V. (1996). *An Introduction to Bayesian Networks*, Springer, New York.

[48] Wickremasinghe, B., Calheiros, R.N. and Buyya, R. (2010). CloudAnalyst: A CloudSim-based Visual Modeller for Analysing Cloud Computing Environments and Applications. Available at http://www.cloudbus.org/cloudsim/

[49] Moore, D., and McCabe, G. (2006). *Introduction to the practice of statistics* 4th ed.New York: Freeman.

[50] Jackson, S.L. (2008). Research Methods and Statistics: A Critical ThinkingApproach: A Critical Thinking Approach 3rd ed. WADSWORTH Cengage Learning Inc., USA.

Proposal for Scrambled Method based on NTRU

Ahmed Tariq Sadiq
Computer Science Department
University of Technology
Baghdad, Iraq

Najlaa Mohammad Hussein
Computer Science Department
Baghdad University
Baghdad, Iraq

Suha Abdul Raheem Khoja
Electronic and communication
Engineering Department
Baghdad University
Baghdad, Iraq

***Abstract*—Scrambling is widely used to protect the security of data files such as text, image, video or audio files; however, it is not the most efficient method to protect the security of the data files. This article uses NTRU public key cryptosystem to increase the robustness of scrambling of sound files. In this work, we convert the sound file into text, and then scramble it in the following way: first, we encrypt the header of the sound file then, scramble the data of the file after the header in three stages. In each stage we scramble the data of the sound file and keep the original order of data in an array then, the three arrays are encrypted by the sender and sent with the encrypted header to the receiver in one file, while the scrambled data of the sound file is sent to the receiver in another file. We have tested the proposed method on several sound files; the results show that the time of encryption and decryption is reduced to approximately one-third, or less, compared to encrypting the file using NTRU.**

Keywords—NTRU; public key; cipher; sound; scramble; segment

I. Introduction

One of the classical encryption ciphers is transposition cipher. In transposition cipher, symbols of plaintext remain the same, but their original sequence is changed in a systematic way. Transposition ciphers are widely used before computer age. With the advent of technology the researchers have invented more complex ciphers like NTRU, which is a relatively new public key cryptosystem. NTRU was developed by three mathematicians J. Hoffstien, J. H. Silverman and J. Piper at the rump session of crypto in 1996. [1] NTRU (Nth degree truncated polynomial ring units) is the first secure public key cryptosystem not based on factorization or discrete logarithm problems. [2] It is a lattice-based public key cryptosystem; its security comes from the interaction of the polynomial mixing system with the independence of reduction modulo to relatively prime numbers.[3] The basic idea of NTRU is finding the shortest vector problem in a lattice. The base of NTRU operations are objects in a truncated polynomial ring R=z[x]/(xN-1) with convolution multiplication and all polynomials in the ring have integer coefficients and degree at most N-1. a = a0 + a1x + a2x2 + … + aN-2xN-2 + aN-1xN-1.[4]The main characteristic is that the polynomial multiplication is the most complex operation that is much faster than other asymmetric cryptosystems such as RSA, ElGamal and Elliptic curve cryptography.[5]

In this paper we introduce a new transposition (scrambling) method with the assistance of NTRU algorithm. In our method sound files will pass through more than one stage of scrambling. The result is a more complex permutation that is not easily reconstructed thus the sound files become significantly much more secure. [6]

II. Related Work

Sadiq, Hussein and Khoja proposed two methods to enhance the NTRU algorithm which they have used for encrypting sound files after converting the sound into text. In the proposed methods the message is encrypted one character at a time. Since NTRU encrypts only prime numbers, only the first 7 bits of each character are encrypted. In method I NTRU algorithm is enhanced by adding the result obtained from calculating a mathematical equation of one variable to the message and then the resulted encrypted bit is fed-back and added to the next bit of the message in the next step; this procedure is repeated for the subsequent bits of the message. In method II NTRU algorithm is enhanced by adding the subsequent states of LFSR (Linear Feedback Shift Register) to the subsequent bytes of the message. The proposed methods are tested on several sound files; the results show that the proposed methods maintain approximately the same original method encryption and decryption time while generating a more complex encryption.[7] Jaspreet Kaur and Er. Kanwal Preet Singh [8] used three different kinds of algorithms NTRU, RSA and RINGDAEL for speech encryption and decryption by first converting the speech into text, and then the text is converted into cipher text. The performance is analyzed for these three algorithms respectively. The parameters calculated are encryption, decryption, delay time, complexity, packets lost and security levels. In these three algorithms, encryption decryption and delay time are varying according to the number of bits per second. On the other hand, complexity and packets lost are approximately the same. There are no packets lost during transmitting and receiving the data. Also Jaspreet Kaur and Er. Kanwal Preet Singh [9] used three different kinds of techniques i.e. MD-5, SHA-2 and RINGDAEL for speech encryption, where the speech is first converted into text then the text is converted into cipher text. At the end, the performances of these three techniques are analyzed, respectively.

III. Ntru

A. *Prameters*

NTRU depends on 3 integer parameter (N, p, q), where N is a prime integer, p and q are relatively prime integers and q is larger than p. [10]

B. Key Generation

To generate the public key, two random polynomials f and g are chosen in the ring R with the restriction that their coefficients are small, usually in {-1, 0, 1}. Another symbol is imported here: L(d1, d2), which means a set of polynomials with d1 coefficients are 1, d2 coefficients are -1 and the rest are 0. f usually chosen from Lf (df, df-1) and g from Lg (dg, dg). Then fp (the inverse of f modulo p) and fq (the inverse of f modulo q) are computed with the property that: f*fp=1(mod p) and f*fq=1(mod q). If f doesn't have inverses, another f should be chosen. The pair of polynomials f and fp should be kept as the private key and the public key h can be computed by h = p * fq * g (mod q). [11]

C. Encryption

The plaintext m is a polynomial with coefficients taken mod p. A random polynomial r is chosen with small coefficients. The cipher text is

e = r * h + m (mod q). [12]

D. Decryption

To decrypt e, the polynomial a is computed first

a = f * e (mod q)

The coefficients of a must be chosen from the interval [-q/2, q/2]. Then the original message can be computed by

m = fq * a (mod q). [10]

IV. THE PROPOSED METHOD

In this method the sender and receiver agree upon four steps of scrambling to the sound, according to random arrays that will be generated and encrypted by the sender and sent to the receiver in a metadata file. The sender then sends highly scrambled sound file to the receiver. This scrambling hides all the information in the file and makes it very difficult to predict and or discover without needing to NTRU encrypting the metadata file.

The sender starts the encryption process by encrypting the header of the sound file and writes it as the first part of the metadata file.

The sound file is split into a number of segments. The segment length (abbreviated SL) indicates the number of bytes in one segment. This length is chosen by the sender and it can vary due to the application and the size of the sound file. SL is encrypted and written to the metadata file.

Then the sender uses the proposed method to start the first step of scrambling the generated segments, which consists of dividing the segments of the file into parts (namely "Data Blocks", DB).The number of data blocks is encrypted using NTRU and written to the metadata file. Data blocks are then scrambled among themselves in a random way, to follow the order of a random array (namely, "Pointers to Data Block", PTDB) which has a length equal to the number of the data blocks; where the indexes of this array are indicators to the new order of the data blocks and the values of this array are used as indicators to the original order of the data block. This array is encrypted using NTRU and written to the metadata file.

Blocks will now be processed one at a time. The length of data blocks (abbreviated LDB) varies and is chosen randomly at run time to hide it. LDB for each data block is encrypted using NTRU and written to the metadata file.

The block size is chosen to satisfy a maximum value restriction to reduce the number of multiplications for both the encryption and the decryption operations as well as to decreases the size of the resulting encrypted file. The block size should also satisfy a minimum, as a very small DB cannot ensure enough scrambling. These maximum and minimum lengths are specified by the sender.

The second step of scrambling is applied to segments inside each data block among themselves to follow the order of a random array (namely "Pointers to Segments", PTS) which has a length equal to the number of segments in that data block, where the indexes of this array are indicators to the new order of segments inside the data block and the values of this array are used as indicators to the original order of segments within the data block. These arrays (one per data block) are encrypted using NTRU and written to the metadata file.

The third step of this method requires that, for each data block, the bytes inside each segment be scrambled among themselves according to a random array (namely,” Pointers to Bytes within segment”, PTB), which has a length equal to the number of bytes in the segment (SL). Where the indices of this array are indicators to the new order of bytes inside each segment and the values of this array are used as indicators to the original order of segments within the data block. The method uses the same scrambling (i.e. the same PTB) for all segments within one data block. Thus, only one PTB is required per data block. PTB are encrypted using NTRU and are written to the metadata file.

The highly scrambled sound data is written to a data file one byte from each segment at a time according to PTS array

Finally, the sender sends to the receiver the two generated files which are: the data file containing the scrambled sound data and the metadata file which contain the encrypted information.

Example: a very small size sound file of 74 bytes is used in this example to illustrate the proposed method. The ASCII of data inside the file after the header is “34 56 128 5 133 22 200 215 99 122 21 78 37 152 163 23 213 183 172 224 174 53 34 183 89 156 172 211 94 11”, segment length is chosen to be 3 bytes, minimum Length of Data Block is 2 segments and maximum Length of Data Block is 6 segments.

The encryption process begins by encrypting the first 44 bytes, the header of the sound file using NTRU, then the four steps of scrambling are applied to the rest of the file, (the remaining 30 bytes of data). According to the method the file is divided into three data block, with the first block containing 3 segments, the second block containing 5 segments, and the last block containing 2 segments as illustrated in fig. (1)

Then the four stages of scrambling are applied as follows:

In the first step the order of the data block is changed where the second data block becomes the first, third data block becomes second and first data block becomes third, PTDB = [2, 3, 1], is used. This step is illustrated in fig. (2).

The second step in data scrambling changes the order of segments within each data block according to the block’s PTS arrays as illustrated in fig. (3).

A PTS is randomly created for each data block to be:

1) For the first Data Block, the PTS 1 is [2, 3, 1].
2) For second Data Block, the PTS 2 is [5, 3, 4, 1, 2].
3) For the third Data Block, the PTS 3 is [2, 1].

The third step of data scrambling changes the order of bytes within data segments as shown in fig. (4), (PTBs), which was randomly created to be:

1) For the first Data Block, PTB 1 is [2, 3, 1].
2) For the second Data Block, PTB 2 is [2, 1, 3].
3) For the third Data Block, PTB 3 is [3, 1, 2].

The fourth and last scrambling step is write to the resulted data file one byte form each segment at time and with accordance to the block's PTB, i.e. for data block one with PTB = [2, 3, 1], the second bytes from all segments in that block are written first, then the third bytes from all the segments in that block are written, lastly the first bytes from all segments are written to the file. This step is shown in fig. (5), while the contents of the metadata file is shown in fig. (6)

V. Experimental Results

In this work we convert the sound file into text; we can apply this method to any sound file after storing it in a text editor. This method is applied to the original NTRU algorithm [13, 14] (namely original method) and the proposed method.

We convert the sound file into text via ISO-8859-1: 8-bit single-byte coded graphic character sets - Part 1: Latin alphabet No. 1, is part of the ISO/IEC 8859 series of ASCII-based standard character encodings, which is intended for “Western European” languages [15].

We test the original and proposed methods on 25 wave sound files of sizes ranging from 10 KB to 1MB. The encryption and decryption time in seconds is computed for each one of the 25 files 25 times, and then average of computation is taken to enhance the accuracy of the computation.

Fig. (7) Displays the effect of file size on the time of encryption and decryption of the original method, Fig. (8) Displays the effect of file size on the time of encryption and decryption of the proposed method with SL 8, Fig. (9) Displays the effect of file size on the time of encryption and decryption of the proposed method with SL 12, Fig. (10) Displays the effect of file size on the time of encryption and decryption of the proposed method with SL 16, Fig. (11) Displays the relationship between file size and encryption time to the original and proposed methods and Fig. (12) Displays the relationship between file size and decryption time to the original and proposed methods.

VI. Conclusions and Future Work

The proposed method scrambles the sound file in a complex and efficient manner while reducing the time of encryption and decryption depending on SL. When SL equals 8, the time is reduced to approximately one-third of the time needed to encrypt and decrypt the same data using NTRU encryption method. Increasing SL to 12 and 16 had an effect of decreasing the time of encryption and decryption (up to 80% of the original encryption and decryption time).

For future research, we proposed a method that combines between NTRU algorithm and ECC algorithm to produce a new stronger cipher system that incorporates the advantages of the two algorithms.

References

[1] Ranjan, R., Baghel, A. S., Kumar, S., "Improvement of NTRU Cryptosystem", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 9, September 2012.

[2] Narasimham, C., Pradhan, J., "Evaluation of Performance Characteristics of Cryptosystem Using Text Files" Journal of Theoretical and Applied Information Technology, JATIT, 2008.

[3] Challa, N., Pradhan, J., "Performance Analysis of Public Key Cryptographic Systems RSA and NTRU" International Journal of Computer Science and Network Security. Vol.7 No. 8 August 2007, pp 87- 96.

[4] Kumar, R. G. V. S., Jumar, N. K., Sekhar, C. P., Numma, B. V. V. S., Kumar, V. B., "Modified Mutual Authentication and Key Agreement Protocol Based on NTRU Cryptography for Wireless Communications", International Journal of Computer Science and Network (IJCSN), Volume 1, Issue 4, August, 2012.

[5] Gupta, N., Ghosh, D., "Implementation of NTRU PKCS using Array for polynomials up to degree 147 with private key protection algorithm using XOR function only", IEEE 2008.

[6] Stallings, W., "Cryptography and Network Security Principles and Practice", Fifth Edition, Pearson, Boston, 2011.

[7] Sadiq, A. T., Hussein, N. M., Khoja, S. A., " Proposal for Two Enhanced NTRU", International Journal of Advanced Computer Science and Applications, Vol. 5, No. 5, 2014.

[8] Kaur, J., Singh, K. p. "Comparative Study of Speech Encryption Algorithms Using Mobile Applications", International Journal of Computer Trends and Technology, Vol.4, Issue. 7, Jully 2013, pp. 2346 - 2350.

[9] Kaur, J., Singh, K. p. "Speech to Text Encryption Using Cryptography Techniques", International Journal of Innovative Research and Development, Volume 2, Issue. 4, pp. 274 - 283, April 2013.

[10] Jha, R., Saini, A, K., "A Comparative Analysis and Enhancement of NTRU Algorithm for Network Security and Performance Improvement" International Conference on Communication Systems and Network Technologies, 2011.

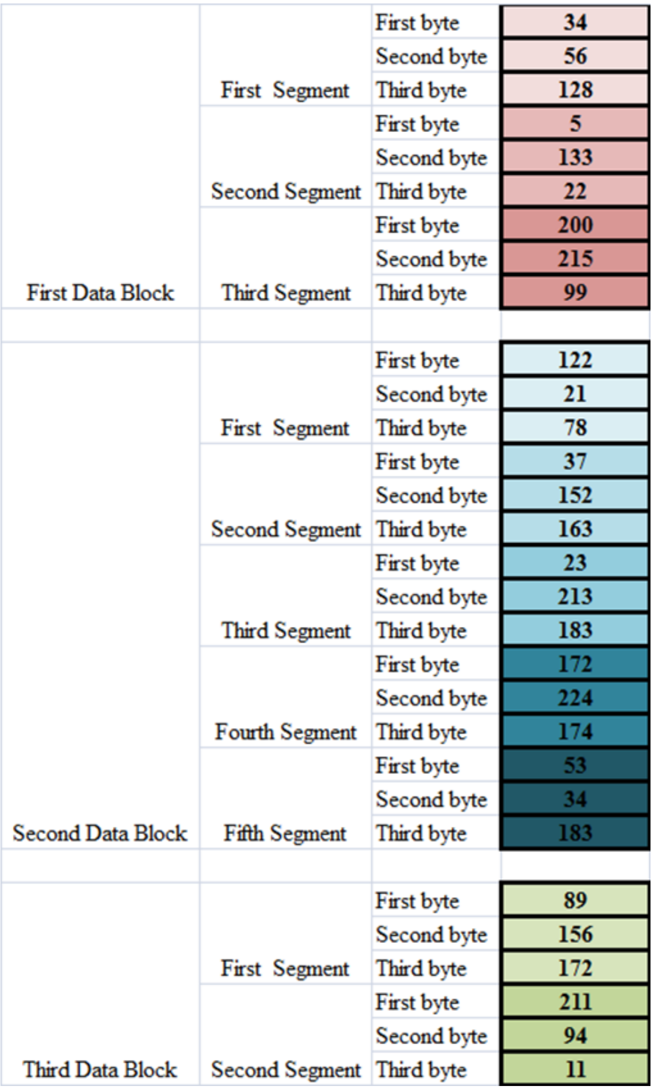
[11] Shen, X., Du, Z., Chen, R., "Research on NTRU Algorithm for Mobile Java Security" IEEE 2009, pp.366 - 369.

[12] Wei, S., Zhuo, Z., “Research on PKI Model Based on NTRU”, International Symposium on Electronic Commerce and Security, IEEE, 2008.

[13] O’Rourke, C. M., “Efficient NTRU Implementations” Thesis, April, 2002.

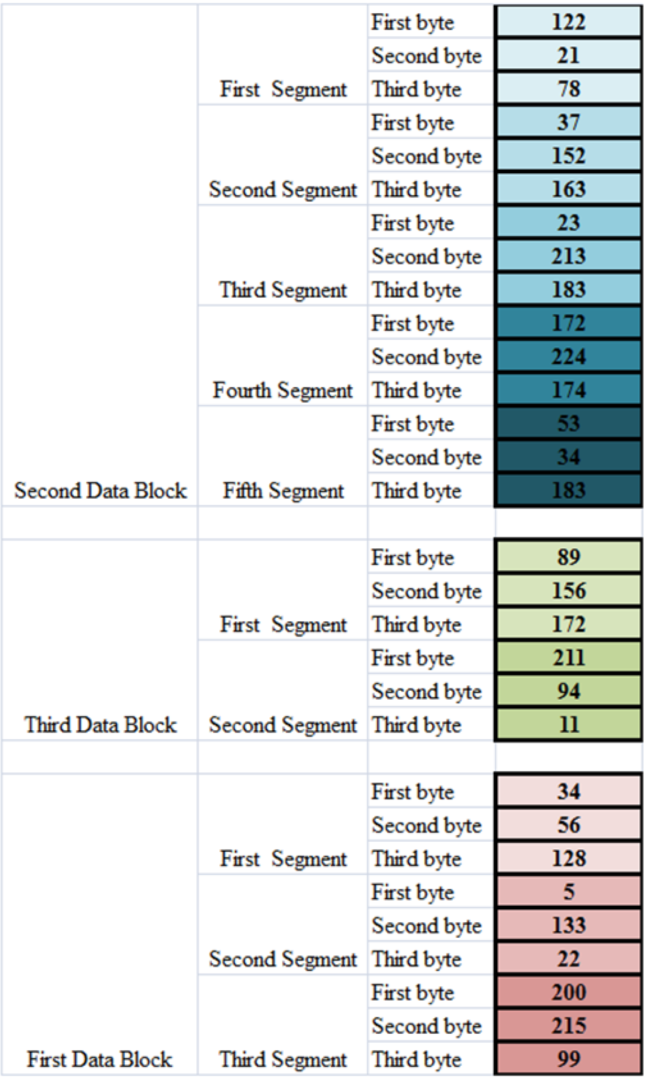
[14] Yadav, S. K., Bhardwaj, K.,“On NTRU Implementation: An Algorithmic Approach", Proceedings of the 4th National Conference; INDIA, 2010.

[15] ISO /IEC JTC 1/SC 2/WG 3 7bit and 8bit codes and their extension SECRETARIAT: ELOT, 1998.



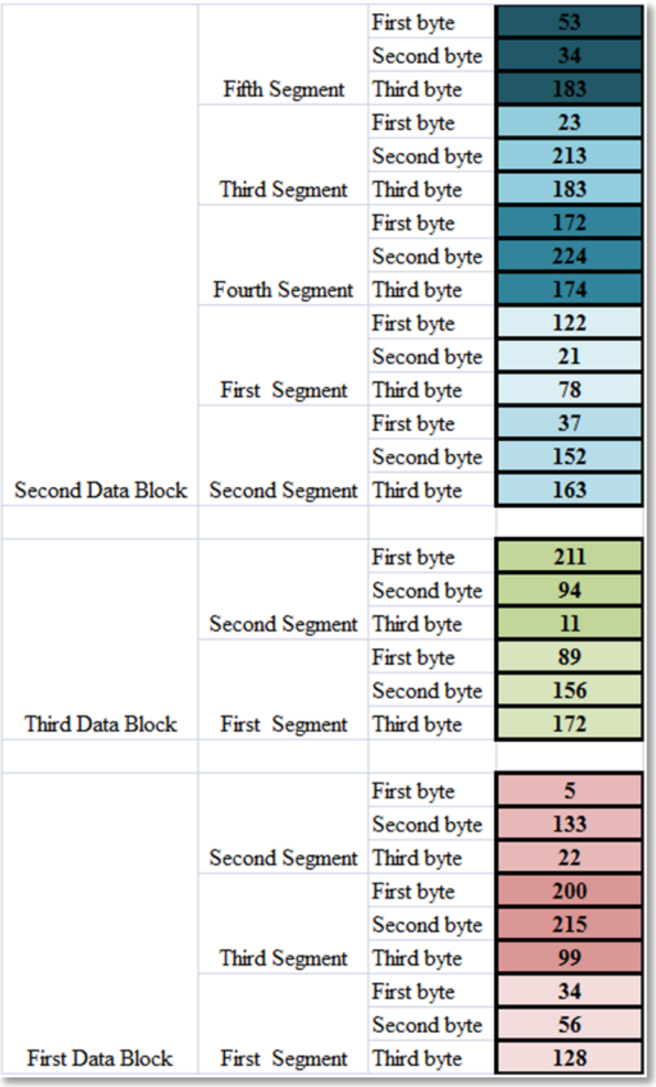
First Data Block	First Segment	First byte	34
		Second byte	56
		Third byte	128
	Second Segment	First byte	5
		Second byte	133
		Third byte	22
	Third Segment	First byte	200
		Second byte	215
		Third byte	99
Second Data Block	First Segment	First byte	122
		Second byte	21
		Third byte	78
	Second Segment	First byte	37
		Second byte	152
		Third byte	163
	Third Segment	First byte	23
		Second byte	213
		Third byte	183
	Fourth Segment	First byte	172
		Second byte	224
		Third byte	174
	Fifth Segment	First byte	53
		Second byte	34
		Third byte	183
Third Data Block	First Segment	First byte	89
		Second byte	156
		Third byte	172
	Second Segment	First byte	211
		Second byte	94
		Third byte	11

Fig. 1. Dividing the file in to segments and then to three data block



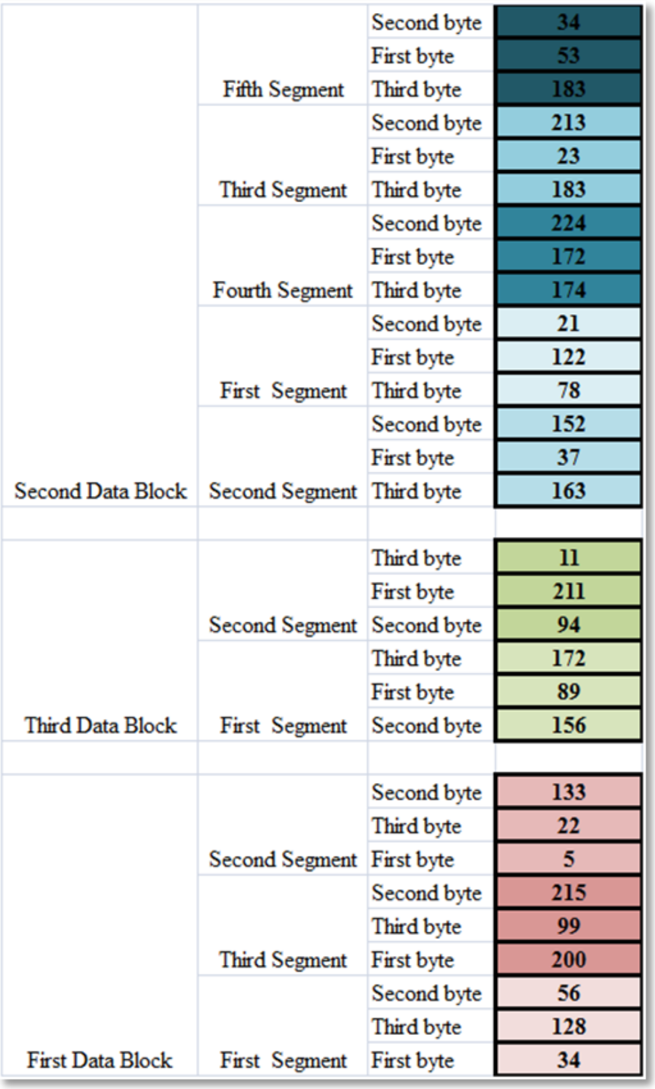
Second Data Block	First Segment	First byte	122
		Second byte	21
		Third byte	78
	Second Segment	First byte	37
		Second byte	152
		Third byte	163
	Third Segment	First byte	23
		Second byte	213
		Third byte	183
	Fourth Segment	First byte	172
		Second byte	224
		Third byte	174
	Fifth Segment	First byte	53
		Second byte	34
		Third byte	183
Third Data Block	First Segment	First byte	89
		Second byte	156
		Third byte	172
	Second Segment	First byte	211
		Second byte	94
		Third byte	11
First Data Block	First Segment	First byte	34
		Second byte	56
		Third byte	128
	Second Segment	First byte	5
		Second byte	133
		Third byte	22
	Third Segment	First byte	200
		Second byte	215
		Third byte	99

Fig. 2. The first step of scrambling (scrambling the dbs)



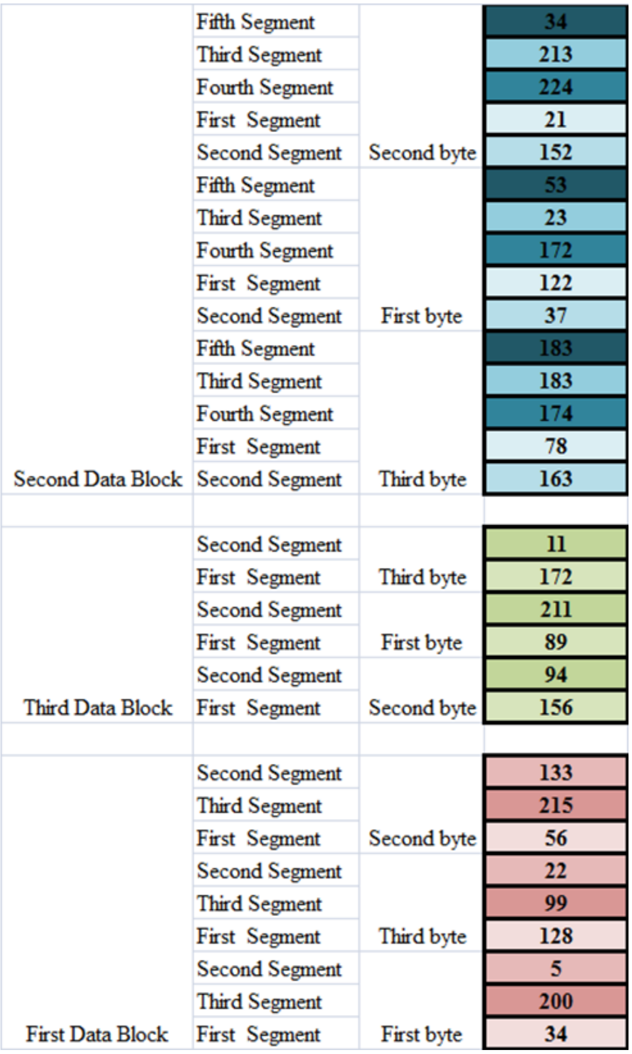
Second Data Block	Fifth Segment	First byte	53
		Second byte	34
		Third byte	183
	Third Segment	First byte	23
		Second byte	213
		Third byte	183
	Fourth Segment	First byte	172
		Second byte	224
		Third byte	174
	First Segment	First byte	122
		Second byte	21
		Third byte	78
	Second Segment	First byte	37
		Second byte	152
		Third byte	163
Third Data Block	Second Segment	First byte	211
		Second byte	94
		Third byte	11
	First Segment	First byte	89
		Second byte	156
		Third byte	172
First Data Block	Second Segment	First byte	5
		Second byte	133
		Third byte	22
	Third Segment	First byte	200
		Second byte	215
		Third byte	99
	First Segment	First byte	34
		Second byte	56
		Third byte	128

Fig. 3. The second step of scrambling (scrambling segments within dbs)



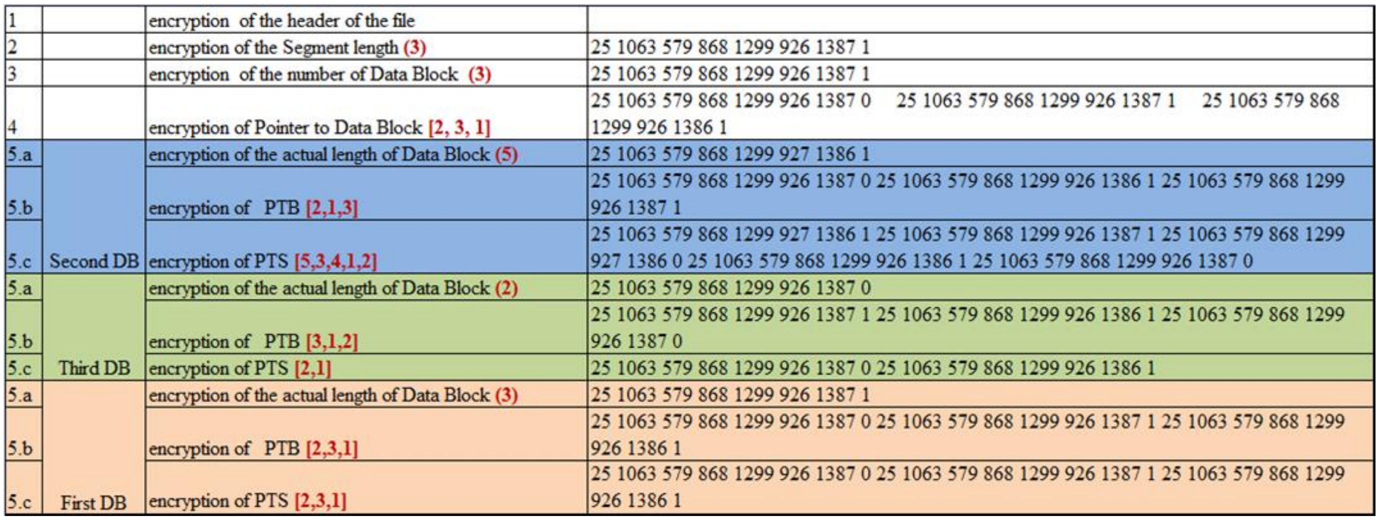
Second Data Block	Fifth Segment	Second byte	34
		First byte	53
		Third byte	183
	Third Segment	Second byte	213
		First byte	23
		Third byte	183
	Fourth Segment	Second byte	224
		First byte	172
		Third byte	174
	First Segment	Second byte	21
		First byte	122
		Third byte	78
	Second Segment	Second byte	152
		First byte	37
		Third byte	163
Third Data Block	Second Segment	Third byte	11
		First byte	211
		Second byte	94
	First Segment	Third byte	172
		First byte	89
		Second byte	156
First Data Block	Second Segment	Second byte	133
		Third byte	22
		First byte	5
	Third Segment	Second byte	215
		Third byte	99
		First byte	200
	First Segment	Second byte	56
		Third byte	128
		First byte	34

Fig. 4. The third step of scrambling (scrambling bytes within segments)



Second Data Block	Fifth Segment	Second byte	34
	Third Segment		213
	Fourth Segment		224
	First Segment		21
	Second Segment		152
	Fifth Segment	First byte	53
	Third Segment		23
	Fourth Segment		172
	First Segment		122
	Second Segment		37
	Fifth Segment	Third byte	183
	Third Segment		183
	Fourth Segment		174
	First Segment		78
	Second Segment		163
Third Data Block	Second Segment	Third byte	11
	First Segment		172
	Second Segment	First byte	211
	First Segment		89
	Second Segment	Second byte	94
	First Segment		156
First Data Block	Second Segment	Second byte	133
	Third Segment		215
	First Segment		56
	Second Segment	Third byte	22
	Third Segment		99
	First Segment		128
	Second Segment	First byte	5
	Third Segment		200
	First Segment		34

Fig. 5. The fourth step of scrambling which results in the order of data that will be written to the data file



1		encryption of the header of the file	
2		encryption of the Segment length (3)	25 1063 579 868 1299 926 1387 1
3		encryption of the number of Data Block (3)	25 1063 579 868 1299 926 1387 1
4		encryption of Pointer to Data Block [2, 3, 1]	25 1063 579 868 1299 926 1387 0 25 1063 579 868 1299 926 1387 1 25 1063 579 868 1299 926 1386 1
5.a	Second DB	encryption of the actual length of Data Block (5)	25 1063 579 868 1299 927 1386 1
5.b		encryption of PTB [2,1,3]	25 1063 579 868 1299 926 1387 0 25 1063 579 868 1299 926 1386 1 25 1063 579 868 1299 926 1387 1
5.c		encryption of PTS [5,3,4,1,2]	25 1063 579 868 1299 927 1386 1 25 1063 579 868 1299 926 1387 1 25 1063 579 868 1299 927 1386 0 25 1063 579 868 1299 926 1386 1 25 1063 579 868 1299 926 1387 0
5.a	Third DB	encryption of the actual length of Data Block (2)	25 1063 579 868 1299 926 1387 0
5.b		encryption of PTB [3,1,2]	25 1063 579 868 1299 926 1387 1 25 1063 579 868 1299 926 1386 1 25 1063 579 868 1299 926 1387 0
5.c		encryption of PTS [2,1]	25 1063 579 868 1299 926 1387 0 25 1063 579 868 1299 926 1386 1
5.a	First DB	encryption of the actual length of Data Block (3)	25 1063 579 868 1299 926 1387 1
5.b		encryption of PTB [2,3,1]	25 1063 579 868 1299 926 1387 0 25 1063 579 868 1299 926 1387 1 25 1063 579 868 1299 926 1386 1
5.c		encryption of PTS [2,3,1]	25 1063 579 868 1299 926 1387 0 25 1063 579 868 1299 926 1387 1 25 1063 579 868 1299 926 1386 1

Fig. 6. The contents of the metadata file

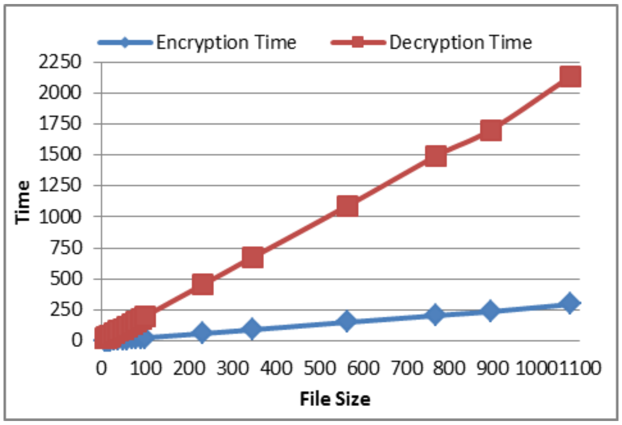


Fig. 7. The effect of file size on the time of encryption and decryption of the original method

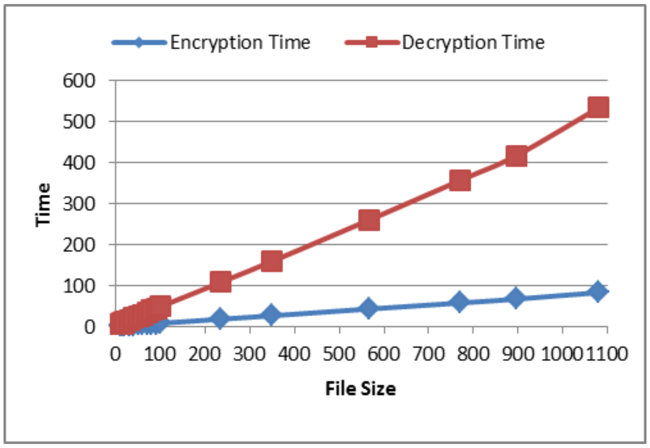


Fig. 8. The effect of file size on the time of encryption and decryption of the proposed method with sl 8

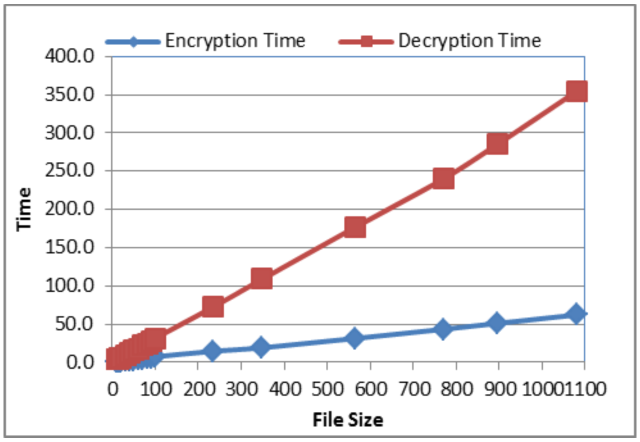


Fig. 9. The effect of file size on the time of encryption and decryption of the proposed method with sl 12

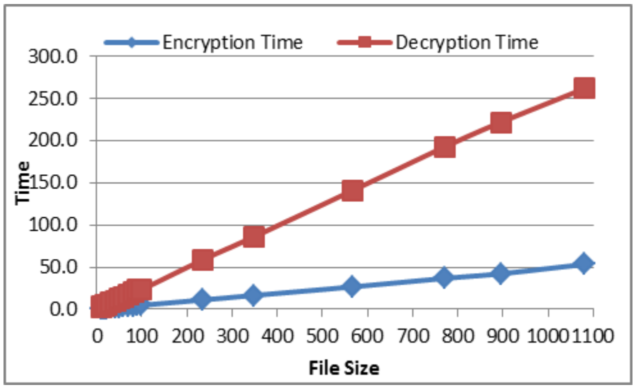


Fig. 10. The effect of file size on the time of encryption and decryption of the proposed method with sl 16

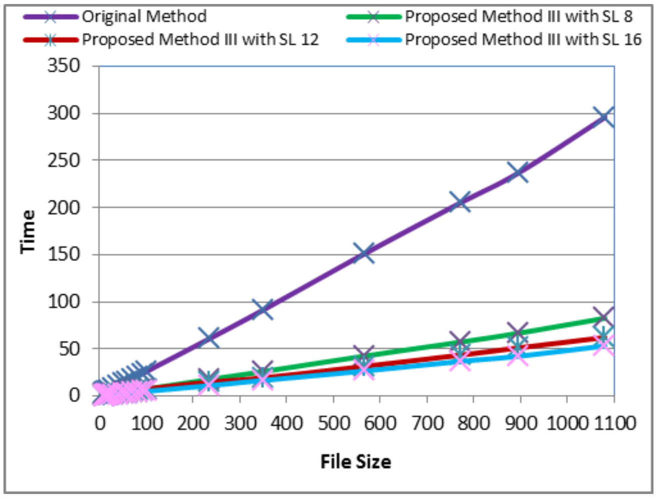


Fig. 11. Compression for the effect of file size on time of encryption between the original method and the proposed method

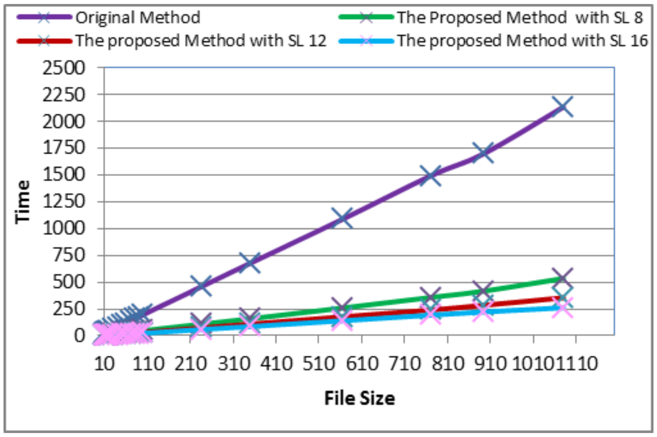


Fig. 12. Compression for the effect of file size on time of decryption between the original method and the proposed method

Network Efficiency – Optimized Automaton Approach

K. Thiagarajan,
Department of Mathematics,
PSNA College of Engineering and Technology,
Dindigul, India

N. Subashini,
Department of Mathematics,
SRM University,
Chennai, India

M. S. Muthuraman,
Department of Mathematics,
PSNA College of Engineering and Technology,
Dindigul, India

***Abstract*—A sperner's grid is thought of a finite state system, where in the model gives rise to an optimal network through characterization of paths .the automation graphs of the various states gives rise to different groomable light paths in network.**

Keywords*—*Automaton; Network; Efficiency; Characterization

I. Introduction

In recent years with the use of different kinds of communication, the network has gained popularity. Automaton traffic engineering is an effective solution to control network conjestion. Automaton traffic engineering comprises of scientific principles thus providing optimal characterization in network.

A network is represented by a set of nodes, inks between the nodes interconnecting them Destination nodes refer to traffic entering or leaving a node, transit nodes refer to nodes were no traffic can enter or leave the node.

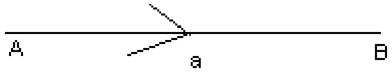
A. Network via Automata

A finite state system represents a mathematical model of a system with certain input. The model finally gives a certain output. The input given to the machine is processed by various states; these states are called intermediate states. This intermediate state leads to final state.

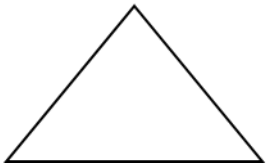
A finite state Automata or simply a network consists of directed graphs composed of states and arcs. We start with a single initial state and could have any number of final states. A path is observed as a sequence of arcs or links from start state to final state.

B. Network through Directions

Consider a directed graph as D= (V (D), X (D),χ_D),Consisting of nonempty set V (D), X (D) being disjoint from V (D) and χ_D a function which associates V (D) and X (D). The set V (D) is called the vertex set of D and X (D) its link set Elements of V (D) are taken as nodes of D and elements of X (D) are called links of (D).



a= (A, B) is a link of directed graph D then 'a' is said to join A and B. We consider only links through directions.



Consider a network as a triangular grid having nodes 0, 1, 2. Nodes could be taken within this grid. We consider only one node taken inside this grid. This single node can be either of 0, 1 or 2. In this paper we consider a single node within the grid. Considering a node within the grid, then the edged would be in order 0-1, 1-2, 2-0. We consider '0' as a single node taken inside the grid, the direction 0 to 1 is taken as traffic in and 1 to o as out traffic. Similar aspects could be thought with regards to the other two namely 1 and 2. Edges with similar nodes are taken as ignored links.

If 'n' is the number of nodes taken on the side of triangular grid having end nodes 0 and 1, the number networks having 010 is 2n+1, the number of nodes having nodes 01010 is 'n' and the number having nodes 0101010 is (n-1).

The maximum of 3 would exits as combination of 01, as we consider a triangular grid. Considering 010, 01010, 0101010 we would have a graph of P_2, P_4, and P_6 where P_K refers to directions in the graph K being even. Finite Automaton is obtained for each of the graphs of 010, 01010, and 0101010. The finite automaton associated with a directed graph is called a transition graph. The vertices of the graph correspond to the states.

C. Nomenclature

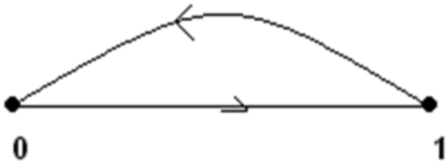
q_{i-} State
α- Number of directions towards the starting node of a path
C-Total number of networks
S- Number of directions towards the ending. Node of a path
$\overline{p}$ - Efficiency

D. Note

1) Efficiency the ratio of the number of directions, towards the starting node with respect to the different between the total number of the networks of networks and number of directions, towards the ending node of the path .

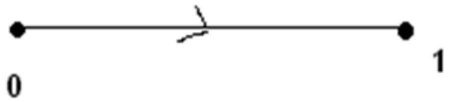
2) Self loops are considered as ignored only in graphical representation self looping is considered in state diagrams.

II. Generation and Efficiency of 010 Network



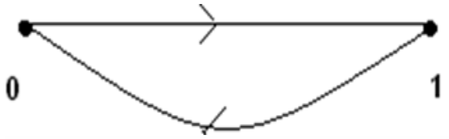
The automation for the above graph has the starts q_0, q_1 as described below

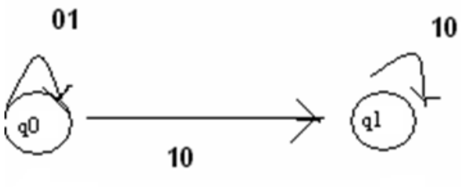
A. State-q_0



q_0 is the initial state infers the path 0 to 1 & is in itself under 01

B. State-q_1

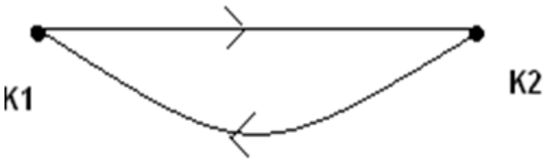




q_1 is the state inferring 1 to 0 under the state q_0. Thus q_0 goes to q_1 under 10

$$\begin{array}{c|cc} & 01 & 10 \\ q_0 & q_0 & q_1 \\ q_1 & --- & q_1 \end{array}$$

q_0 goes to itself under 01 & q_1 is in itself under 10

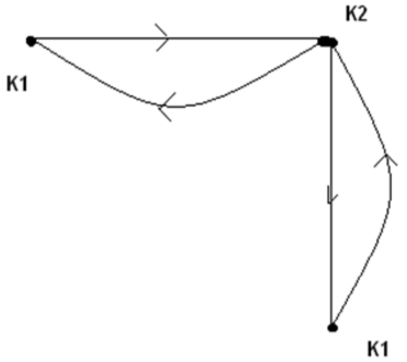


K1 represents the node 1 & K2 represents the node 0

C. Characterization of the path K2 to K1

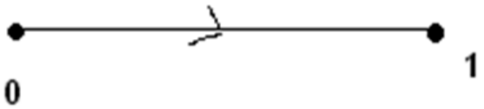
C=2
α→K2 = 1
S→ K1 = 1
$\overline{p}$ = 1

III. Generation and Efficiency of 01010



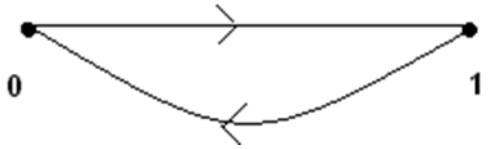
K1 represents 1; K2 represents 0; K3 represents 1 the above graph represents 0 to 1; 1 to 0 & 0 to 1 & 1 to 0 thus giving the networks 01010. The automation for the above graph has the states q_0, q_1, q_2, and $q_{3.}$

A. State-q_0



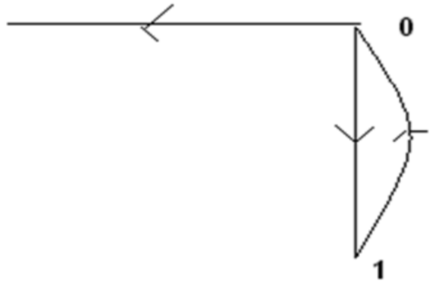
q_0 is the initial state infers the path 0 to 1 and is in itself under 01.

B. State-q_1



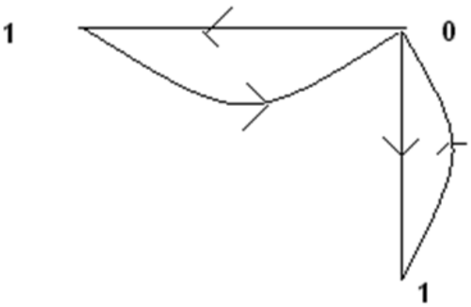
q_1 is the state inferring 1 to 0 under the state q_0. Thus q_0 goes to q_1 under 10, q_1 remains in itself under 10.

C. State-q_2

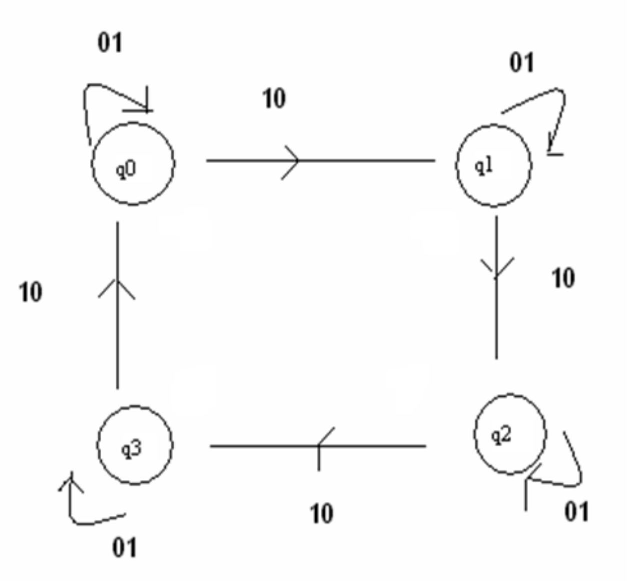


q_2 is a state inferring 0 to 1 under the state q_1, thus q_1 goes to q_2 under 01, q_2 remains in itself under 10.

D. State-q_3



q_3 is the state inferring 1 to 0 under the state q_2, Thus q_2 goes to q_3 under 10,q_3 remains in itself under 10.



The matrix of the states with regards to the path 01&10

$$\begin{array}{c|cc} & 10 & 01 \\ q_0 & q_0 & q_1 \\ q_1 & q_1 & q_2 \\ q_2 & q_2 & q_3 \\ q_3 & q_3 & q_0 \end{array}$$

Efficiency of the network 01010

C=4

E. Characterization of the path K1 to K2

C=4
α→K1 =2
S→ K2=2
$\overline{p} = 1$

F. Characterization of the path K2 to K3

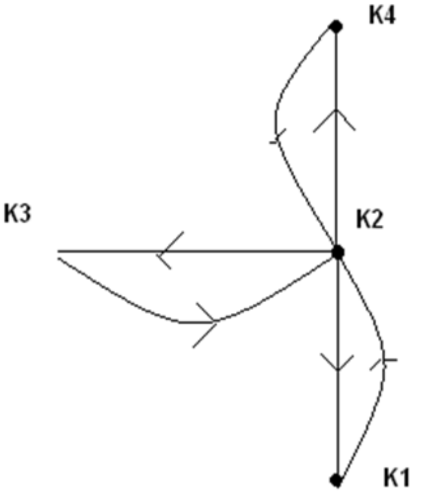
C=4
α→K2 =2
S→ K3=2
$\overline{p} = 1$

G. Characterization of the path K2to K1

C=4
α→K2 =2
S→ K1=2
$\overline{p} = 1$

H. Characterization of the path K3to K1

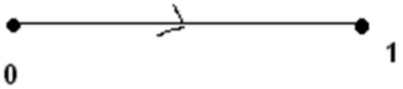
C=4
α→K3 =2
S→ K1=2
$\overline{p} = 1$



IV. Generation and Efficiency of 0101010 Network

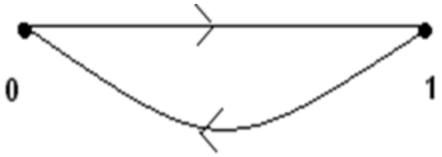
K1 Represents 1; K2 represents 0; K3 Represents 1; K4 Represents 1.The automation for the above graph has the states q_0, q_1, q_2, q_3, and q_4.

A. State-q_0



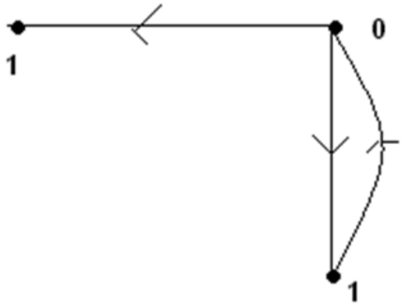
q_0 is the initial state infers the path 0 to q_0 is in itself under 01.

B. State-q_1



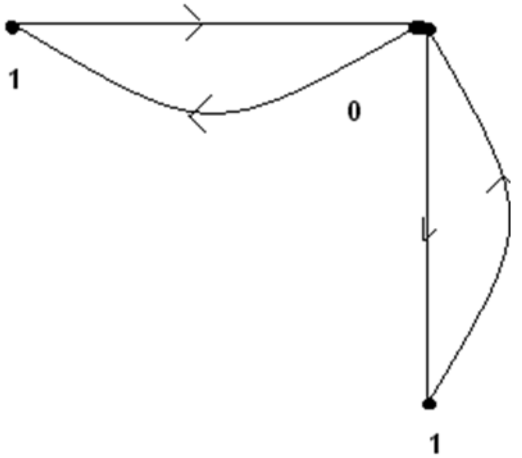
q_1 is the state inferring 1 to 0 under the state q_0. Thus q_0 goes to q_1 under 10, q_1 remains in itself under 10.

C. State-q_2



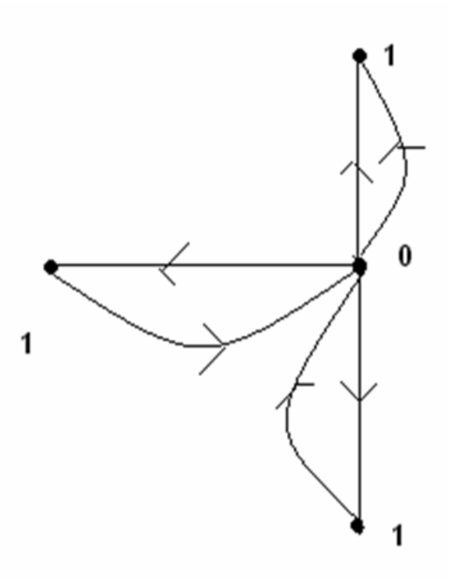
q_2 is a state inferring 0 to 1 under the state q_1, thus q_1 goes to q_2 under 01,q_2 remains in itself under 10.

D. State-q_3



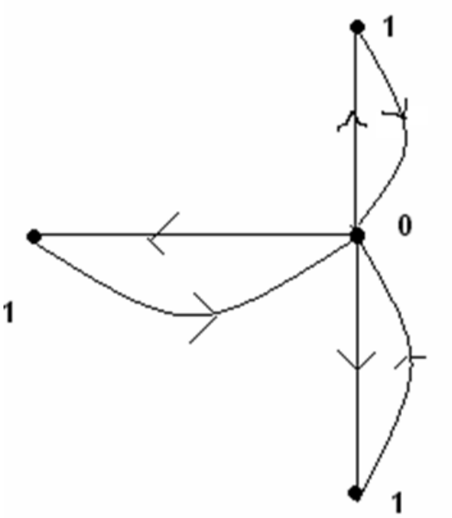
q_3 is the state inferring 1 to 0 under the state q_2, Thus q_2 goes to q_3 under 10,q_3 remains in itself under 10.

E. State-q_4



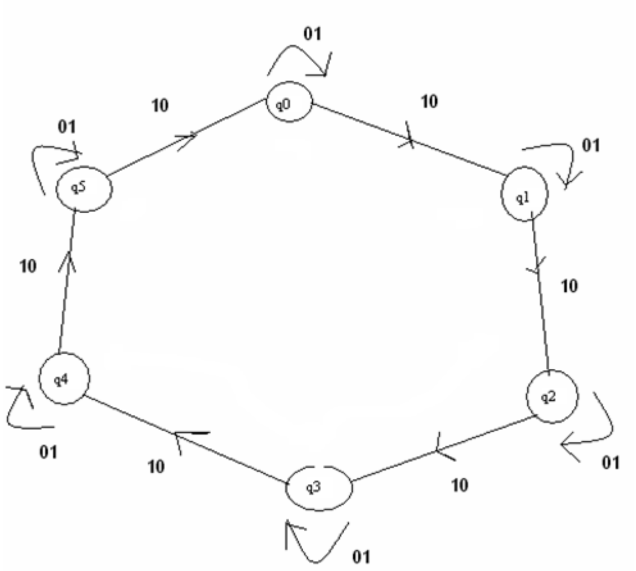
q_4 is a state inferring 0 to 1 under the state q_3, thus q_3 goes to q_4 under 01,q_4 remains in itself under 10.

F. State-q_5



q_5 is a state inferring 0 to 1 under the state q_4, thus q_4 goes to q_5 under 01.q_5 remains in itself under 10.

The matrixes of the states are as fallows



$$\begin{array}{c|cc} & 01 & 10 \\ \hline q_0 & q_0 & q_1 \\ q_1 & q_1 & q_2 \\ q_2 & q_2 & q_3 \\ q_3 & q_3 & q_4 \\ q_4 & q_4 & q_5 \\ q_5 & q_5 & q_0 \end{array}$$

Efficiency of the network 0101010

G. Characterization of K1 to K2

K2 to K1
K3 to K2
K2 to K3
K2 to K4
K4 to K2

C=6
α→K1=3
α→K2=3
α→K3=3
α→K4=3
S→K1=3
S→K2=3
S→K3=3
S→K4=3

H. Characterization of K1 to K2

K2 to K1
K3 to K2
K2 to K3
K2 to K4
K4 to K2

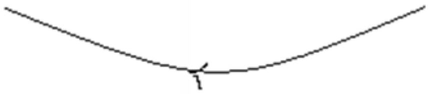
In the of the above case we obtain $\overline{p} = 1$

I. Result

We observe from the automation of the graphs 010; 01010; 0101010 the state q_i ; i=1,2,3,4 goes to q_i+1

In the Graphical representation we observe

$K_1 \rightarrow K_2 \rightarrow K_3 \rightarrow$--------------$\rightarrow K_{n-1} \rightarrow K_n$



In general $K_i \rightarrow K_{i+1}$ under the path 01

V. Conclusion

We observe that in the graphical representation of the graphs 010; 01010; 0101010; we observe in the characterization of the path K_1 to K_2 &K_2 to K_1 efficiency is 1.In the graphical representation of 01010 & 0101010 we observe the characterization of the path K_2 to K_3 &K_3 to K_2 the efficiency is 1.Both the networks are of the source and destination groom able light paths, Wherein the termination is at the node 0 & could be routed towards other nodes, & can also originate from other nodes. When we consider efficiency we observe that destination is at the node 0 & 1 according to the consideration of the path automata is applied to paths of a network there by giving ideas for better groom ability

Acknowledgment

The authors would like to thank Dr. Ponnammal Natarajan, Former Director – Research, Anna University- Chennai, India and currently an Advisor, (Research and Development), Rajalakshmi Engineering College, Dr. K. Sarukesi former vice chancellor in Hindustan University-Chennai, India, Professor E.Sampath Kumar Acharya and Professor L.Pushpalatha, University of Mysore, Mysore, for this initiative ideas and fruitful discussions with respect to the paper's contribution.

Reference

[1] Graph Theory by Frank Harary, Addison Wesley Publishing Company, Third Printing, October (1972).

[2] Graph Theory With Applications To Engineering And Computer Science By Narsingh Deo, PHI Publications, New Delhi (1995).

[3] Special Study in (1,2)- Distance Coloring and (1,2) -Distance Graphs. K.Thiagarajan, L.Pushpalatha, Ponnamal Natarajan. Ultra Scientist Of Physical Sciences, Vol.19 (3) M.731-733(2007).

[4] Special Study in (2, 2) Distance Coloring and (2,2) -Distance Graphs (for paths and cycles). K.Thiagarajan, Ponnamal Natarajan. Ultra Scientist Of Physical Sciences, Vol. 19 (3) M.734-736(2007).

[5] Introduction to Automata Theory, Languages, and Computation(3rd Edition Pearson Publication) -John E. Hopcroft, Rajeevmotwani, Jeffrey D. Ullman.

Authors Profile

K. Thiagarajan completed his Ph.D from University of Mysore. He is currently working as Associate Professor in the Department of Mathematics, P.S.N.A.College of Engineering and Technology, Dindigul, Tamilnadu, India. He has totally 18 years of experience in teaching. He has attended and presented research articles in 37 National and International Conferences and published one National journal and more than 50 International journals, Currently he is working on Big Data and web mining through automata and set theory. His area of specialization is 'Coloring of Graphs and DNA Computing.

Ms. R. Subashini have completed M .sc, B Ed, M phil. and she is having 10 years teaching experience in various engineering colleges. Currently she is working as a Assistant Professor in Department of Mathematics at SRM University, Kattankulathur, Chennai. She have Participated and presented two papers in national and international conferences at SRM and Sathyabama university. Currently she is working in optimization of networks.

M.S.Muthuraman (B.Sc., M.Sc., M.Phil., in Mathematics)is currently working as an Associate Professor ,Department of Mathematics, PSNA College of Engineering and Technology,Dindigul, Tamilnadu, India. He is the author of 25 papers published in reputed journals and in proceedings of conferences, with many references from other researchers. His interests include Fuzzy Algebra, Discrete Mathematics.

HTCSLQ : Hierarchical Tree Congestion Degree with Speed Sending and Sum Costs Link Quality Mechanism for Wireless Sensor Networks

Mbida Mohamed [1]
Phd student
Department of Emerging Technologies Laboratory (LAVETE)
Faculty of Sciences and Technology Hassan 1st University
Settat, Morocco

Ezzati Abdellah [2]
Professor
Department of Emerging Technologies Laboratory (LAVETE)
Faculty of Sciences and Technology Hassan 1st University
Settat, Morocco

***Abstract*—Wireless Sensor Network (wsn) performances have progressed over the last few years, aiming at expanding the lifetime nodes. Among the important studied parts on wsn is the congestion degree that can be identified by several algorithms such as HTAP (Hierarchical Tree Alternative Path), which is the target in the current study. It is important to mention that besides the variations observed between the four HTAP deployments of this classical algorithm (detailed on the rest of paper) , there is a possibility that other factors, such as energy efficiency, network lifetime are affected by the node displacements in the HTAP phases (in case of the route change caused by the isolated dead nodes to the sink) , which give the idea to reduce this undesirable problem by designing a new algorithm called Hierarchical Tree Congestion degree with Speed sending and Sum costs link Quality (HTCSLQ) by reducing the energy consumption , according to the choice of the lower Congestion degree , optimal Speed sending and Sum costs link Quality values from the source to destination.**

Keywords*—HTAP (Hierarchical Tree Alternative Path); Cd (congestion degree); Ss(speed of sending); SCVQL (Sum of costs variable Link quality); Wireless Sensor Network (wsn)

I. Introduction

Wsn nodes is an embedded system that links between sensors/actuators with different communications mechanisms , each node is characterized by autonomy and establish a collaboration with their neighbors in order to form the network and transmit data to the sink , for this reason it is important to use the management system in queues of wsn.

The queues management Causes problems in sequencing level of data, among those factors which affect negatively the performance of wsn is the congestion (the target of this study). There are many protocols in the literatures designed to control congestion wsn , as the HATP(Hierarchical tree Alternative Path) protocol is most study and is designed to check and avoid the congestion under four differents layouts node placement (explained in the related work of this article). In this article, a new congestion control protocol HTCSLQ are designed , which will corrects the overconsumption energy of problems displacements nodes compared to the execution of classic phases HTAP.This research requirement focuses on modeling a new Algorithm Congestion on WSN, with optimized energy consumption. The paper is organized as follows : Section II presentation of related work Section III description of HTCSLQ mechanisms to avoid the problems of displacements nodes. Section 4 demonstrates the proposed dynamic HTCSLQ algorithm. Section 5 provides simulation and comparative study with the classical HTAP using the Jprowler simulator, and Section 6 the conclusion.

II. Related Work

In the literature HTAP is distributed and scalable framework with any type of network in order to reduce the level of congestion and make a safe transmission in each scenario , and designed to minimize packet Losses.This technique is based on the creation of alternative paths between the source and the sink, this establishment does not always make only the use of nodes in initial shorter path , its linear randomized choice of next hops (random selection of Congestion degree value in each elected path nodes) (Figure 1) involves and keep communication between nodes.

HTAP algorithm consists of four fundamental parts (Figure 2) :

- Flooding with level discovery functionality (FD): Through this step, each node locate its neighbors and renew its neighbor table, and every node is placed in levels between the source and the sink.
- Alternative Path Creation Algorithm (APC) (boolean congested or not): In goal to avoid congestion, every congested node receiver is sending a notification packet congestion to inform the sender. Accordingly the sender node stops the transmission of packets to the congested node and finds another least congested node in its neighbor table in aim to maintain the transmission of data, and the creation of new roots is done to the sink.

The Hierarchical Tree Algorithm (HT): A hierarchical tree is created between the nodes, and became connected using the exchange of packets handshake ,this packets contains also the state of congestion. This two algorithms combined build the Hierarchical Tree Alternative Path (HTAP) algorithm.

- Handling of Powerless (Dead Nodes):

The HTAP algorithm in case of the nodes going to lose its energy and directly isolated from the network and the tables of its neighbor nodes are updated.

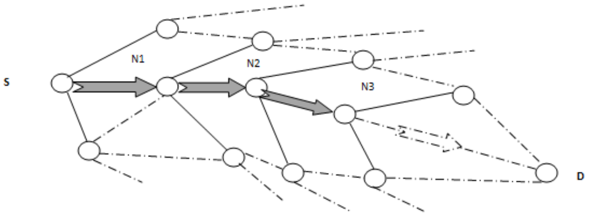


Fig. 1. Example of randomized and linear choice HTAP

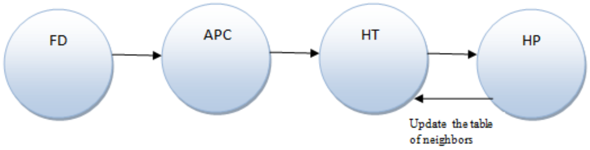
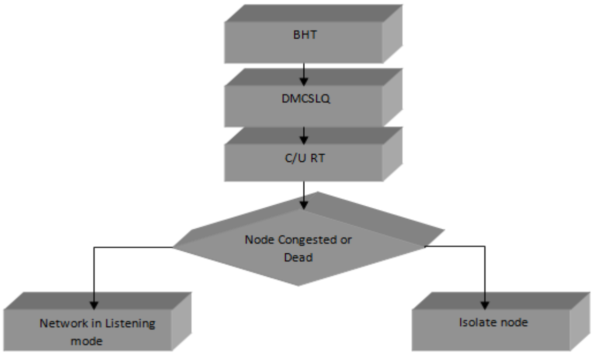


Fig. 2. HTAP algorithm functioning

III. The Theorical View of HTCSLQ

HTCSLQ Algorithm are designed in order to be adaptable with any kind of network purpose is to widen the network life time and to increase its during the transmission data, The following phases describes its functioning (Figure 3).



- BHT : Building of Hierarchical tree
- DMCSLQ : Detection and measurement of Cd , SS , SCVLQ of each node
- C/U RT   : Construction or updating every root table with the three factors

Fig. 3. Architecture of HTCSLQ building Network

A. Congestion deggree mechanism

The calculation of Congestion Degree is the ratio between Local packet inter-arrival time and Local packet inter-service time. The measurement of the value of congestion Degree specify the rate of congestion of the network. the following formula presente the theorical calculation :

Congestion Degree (Cd) = Ts/Ta Where Ts = local packet inter-service time , Ta = local packet inter-arrival time

Algorithm 1 : Cd calculation

```
Input :
Ts : local packet inter-service time
Ta : local packet inter-arrival time
Tree of nodes : TN
Node : n
Output:
Congestion degree value : CDV
for each ni ε TN do
for each cdi of ni
Cdvi<==Tsi/Tai
End
End
```

B. Speed of sending calculation

When a node has a packet to be sent to the sink, it has to calculate the speed of the available nodes. Based on this estimation, the nodes which have the ability to send a packet according to the suitable speed. In order to do this, the sender is connected to sink node , for example the node i with coordinates (x i , y i) and the destination with (x d , y d) , the progress of node j with location (x j , y j) is found by the projection of point j into line connecting i and d and is presented as P ij (Figure 4). The value of P ij is calculated with the formula : P ij= D ij* $cos\ \alpha$

Where D ij is the distance between node i and j and α is the angle between nodes ij and id. These are computed as:

D ij = $\sqrt{(x\,i\ -\ xj)2+\ (y\,i-\ y\,j)2}$ and $\alpha = \arctan|m1-m2|/|1+m1m2|$

in which m1 and m2 are the sloop of line ij and line id respectively

m1= yj-yi/xj-xi, m2=yd-yi/xd-xi

The speed for node j is found as: Speed ij = P ij /delay ij

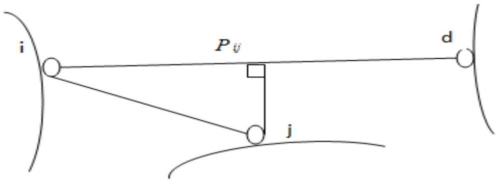


Fig. 4. Geometrical speed of sending calculation

The following algorithm describe the calculation :

Algorithm 2 : SS calculation of jth node

```
Input :
Pnj : progress of node j
according to n .
Delay nj  : time to send
between node n and j .
Tree of nodes : TN

Output:
Speed of Sending value :
SSV
```

```
for each nj ε TN do
for each ni ε TN do
for each pij of nj
SSVij<== P ij /delay ij
End;
End;
End;
```

C. Sum of Costs variable link quality calculation

In this part the value of Link quality is calculated , in case : if there is three nodes connected (T " source node" and M and N and R "final receiver") , such as each node has its rooting table which contains the sum of costs variable link quality (SCVLQ) from the source to the sink , however this technique makes the Costs dynamic in the rooting table , as example if the node N receive the packet (the simplified structure of packet is described in figure 6) from M , which contains the sum of costs part determined from the source to destination , its compared with the calculation of local sum costs to the same destination (at node N) , if it finds that the value of the SCVLQ of node N is smaller than the SCVLQ at node M , so it changes the value on the packet (figure5) , else it changes the SCVLQ rooting table at node N , however this mechanism provides a good transmission of packets with a best quality of links between nodes (NB : This technique is used in sending packet phase , in this case all nodes will changes its SCVQL rooting table to the SCVQL packet after the choice of next hop explained in the rest of paper).

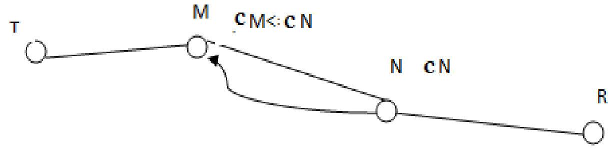


Fig. 5. Diagram of dynamic SCVLQ

The next algorithm illustrate the mechanism of SCVLQ:

```
Algorithm 3 : SCVLQ calculation of jth node
Input :
Count Cost X : sumation of costs between source and sink in routing table in Xth node .
Count Cost Y : sumation of costs between source and sink in roting table of Yth node
Tree of nodes : TN
Minimal LQ cost : MLQC
Output:
Packet to next hop : PNH
for each ni ε TN do
for each count cost i of ni
curent count costs i<== count cost i
End;
for each ni ε TN do
if( curent count costs i > curent count costs i+1)
curent count costs i <== curent count costs i+1
End;
Minimal LQ cost <= curent count costs i;
send PNH(Minimal LQ cost);
End;
```

Delimiter Of onset packet	ID	ASN	ADN	SCVLQ	DATA	....	Delimiter Of the end of packet

ASN : Adress Source of Node

ADN : Adress Destination of Node

SCVLQ :Sum of costs variable Link Quality

Fig. 6. General structure of packet exchanged between nodes with ASVLQ

D. HTCSLQ

This algorithm are designed to correct the classical Algorithm of HTAP congestion , however it reduces the factors effects ,such as energy efficiency, network lifetime which are affected by the node displacements , thus reduces more level of congestion better than the HTAP . The folowing algorithm present the functioning of HTCSLQ protocol using the 3 mechanisms :

```
Algorithm 4 : HTCSLQ
Input :
    Congestion degree : Cd
    Speed of sending : Ss
    Sum of costs variable of Link Quality : SCVLQ
    Tree of nodes : TN
    Optimal elected next node : OENN
Output :
    Packet to next optimal hop: PNOH
for each ni ε TN do
for each nj ε TN do
if ( Cdj <Cdj+1 || Ssj<Ssj+1 && SCVLQj< SCVLQj+1)
Send PNOH(OENNij) ;
End ;
End ;
End;
```

HTCSLQ also allows to quantify and measure the 3 variables (Cd, Ss, SCVLQ) of each node (Figure 7) , after that it build the root from the source to the final destination, and also make choice of the elected next hops , the source send the data to this node and fill the SCVQL part of packet by the current SCVQL selected , however the HTAP allows just the detection of congestion , in order to isolate the nodes from the network.

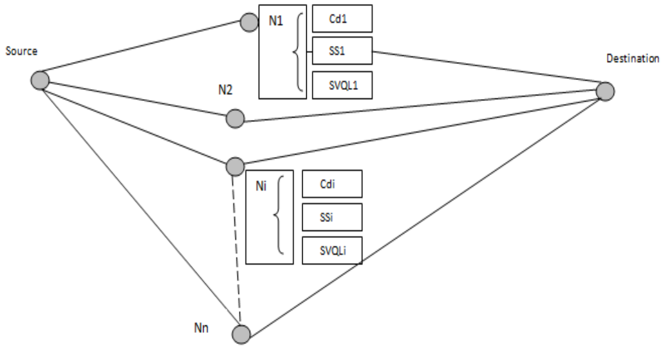


Fig. 7. Diagram HTCSLQ Measurements of the 3 mechanisms

IV. Simulation and Results

After a series of randomized simulations HTCSLQ and HTAP with the same model Tree (from the root to the last leaf node of network) and settings (Table 1) , on JProwler (Figure 8) the congestion control and characteristics of HTCSLQ algorithm is experimented with the simulation space which includes the deployment of the nodes in a 1000m x 1000m grid.

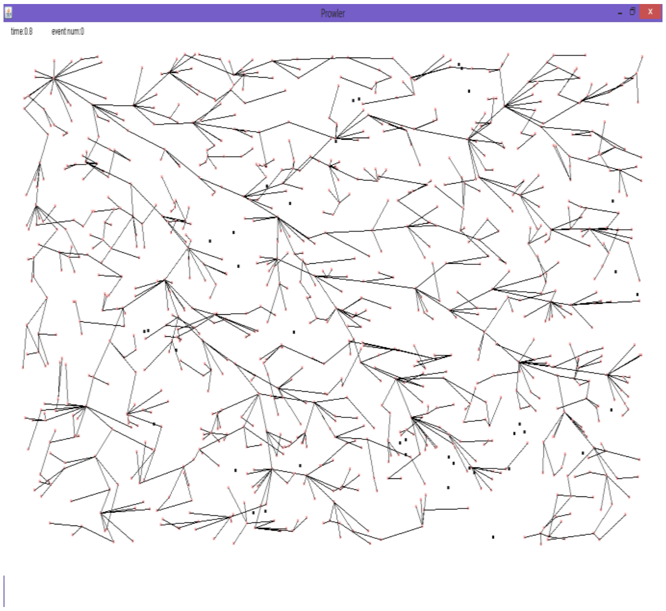


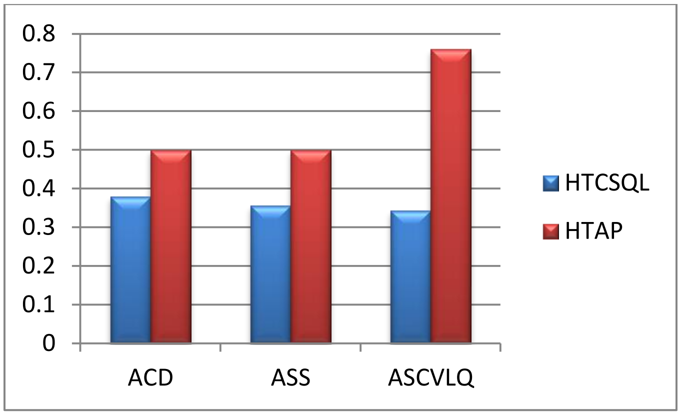
Fig. 8. Exemple of HTCSLQ Hirarchical tree in WSN

TABLE I. Parametres of Simulations

Simulation time	20000 ms
Max Radio Strength	100
Type of Network	Grid
Distibution node	uniform
Radio strength cut off	1/30
Send transmission Time	960 ms
Noise variance	0.025
MaxAllowednoise on sending	5
Receiving Start SNR	4.0
Corruption SNR	2.0
Number of nodes	1000

A. Experimental HTCSLQ mechanisms study

In this part the Congestion degree, Speed of sending and sum of costs variable link quality of HTCSLQ is evaluated and compared with the HTAP algorithm , using by reference the encoding Java in JProwler Simulator:



Indications:

ACD (coefficient)	: Average of congestion degree
ASVLQ real = [ASVQL * 10] :	Average of Sum of costs variable Link quality
ASS (kbits/s)	: Average of Speed of Sending

Fig. 9. Comparative performance between HTCSLQ and HTAP

According to the comparative study (Figure 9) the average congestion degree from the same source to the destination in the network of the algorithm HTCSLQ is optimal compared to the HTAP because the first detect the nodes which have a lower congestion then the average expected , however the HTAP detect only the nodes congested and isolate its . Also it is found that the ASS and ASCVLQ reached an average lower than has the HTAP , in order to find an optimal speed of data transmission and network stability with a minimum sum of costs . The life time of nodes becomes long because the HTCSLQ has the possibility to choose the nodes which are more efficient than other based on the 3 mechanisms (Cd , SS ,SCVLQ) in the case of rebuilding the network and isolation of congested or dead nodes .

B. Most congested HTCSLQ and HTAP nodes

Depending on the simulation under Jprowler , a function is encoded for showing the most congested nodes higher than the experimental average (cd = 0.6) (using the same configuration of Table 1) , the following graphs represent the results for these two algorithms:

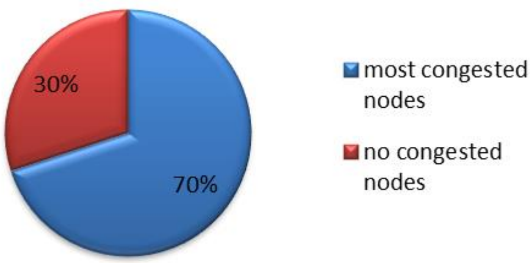


Fig. 10. Pourcentage of most HTAP congested nodes

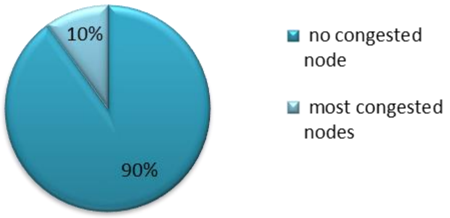


Fig. 11. Pourcentage of most HTCSLQ congested nodes

By Reference of successive simulations in the HTAP and HTCSLQ network on the same source to the destination, the linear and randomized choice of HTAP saturates the same nodes used during all simulations, however the HTCSLQ avoids pressure on the same node and uses the three mechanisms in goal to choose the best and optimal root in each simulation.

According to the two figures (Figure 10,11) the percentage of HTCSLQ most congested nodes which can be chosen as root from the same source to the destination (from the root to the end of the tree structure in the current simulations), reached a probability of 10% , as long as the HTAP reached 70% (Figure 10 and 11) , and this led to deduce that the algorithm HTCSLQ during the reconstruction of the optimal root in case of every simulation from the same source to destination gives a longer life time to the network and energy efficiency of each node .

V. Conclusion and Future Work

The HTCSLQ algorithm has succeeded in reducing the effects caused by the displacement of nodes (phase of isolation dead nodes) compared to the Classical HTAP , as said before the functioning of this Algorithm is resumed in two parts : measurement of the 3 variables(Cd, Ss,SCVQL) in all networks nodes and choices of this optimal values in order to establish the less expensive root in terms of energy efficiency, network lifetime to the sink . The future work includes a study of fault tolerance integration and measurements , to design a new algorithm that can be compared with the performances of the current algorithm (HTCSLQ), and see if the integration and the optimal choice of this value in every nodes give more equilibrium energy and life time in the network.

References

[1] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva, “Directed Diffusion for Wireless Sensor Networking,” ACM/IEEE Transactions on Networking, vol. 11, no. 1, pp. 2–16, February 2002.

[2] M. Ishizuka and M. Aida, “Performance Study of Node Placement in Sensor Networks” In Proceedings of the 24th international Conference on Distributed Computing Systems

[3] C. Wang, K. Sohraby and B. Li, “SenTCP: A Hop-by-hop Congestion Control Protocol for Wireless Sensor Networks,” In Proceedings of IEEE INFOCOM 2005 (Poster Paper), Miami, Florida, USA, Mar. 2005.

[4] S.-J. Park, R. Vedantham, R. Sivakumar, and I. F. Akyildiz, “A scalable approach for reliable downstream data delivery in wireless sensor networks,” In Proceedings of ACM MobiHoc’04, May 24-26, 2004, Roppongi, Japan.

[5] J. Kang, Y. Zhang, and B. Nath, “End-to-End Channel Capacity Measurement for Congestion Control in Sensor Networks,” In Proc. of the Second International Workshop on Sensor and Actor Network Protocols and Applications (SANPA 2004), August 2004

[6] C. Y. Wan, S. B. Eisenman, and A. T. Campbell, “CODA: Congestion Detection and Avoidance in Sensor Networks”, In Proceedings of the ACM SenSys, November 2003.

[7] S. Tilaky, N. B. Abu-Ghazalehy and W. Heinzelmanz, “Infrastructure Tradeoffs for Sensor Networks,” In Proc. of the 1st ACM international Workshop on Wireless Sensor Networks and Applications (WSNA '02), pp.49-58, Atlanta, GA, September 2002.

[8] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva, “Directed diffusion for wireless sensor networking,” IEEE/ACM Transactions on Networking, vol. 11, no. 1, pp. 2–16, 2003.

[9] A. Cerpa, J. Elson, M. Hamilton, J. Zhao, D. Estrin, and L. Girod, “Habitat monitoring: Application driver for wireless communications technology,” in SIGCOMM LA ’01: Workshop on Data communication in Latin America and the Caribbean. New York, NY, USA: ACM, 2001, pp. 20–41.

[10] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva, “Directed diffusion for wireless sensor networking,” IEEE/ACM Transactions on Networking, vol. 11, no. 1, pp. 2–16, 2003.

[11] Cheng Wang, Changjun Jiang, Yunhao Liu, Senior Member, IEEE, Xiang-Yang Li, Senior Member, IEEE, and Shaojie Tang , IEEE TRANSACTIONS ON COMPUTERS, VOL. 63, NO. 6, JUNE 2014

[12] Ye, F.; Chen, A.; Lu, S.; Zhang, L. Scalable solution to minimum cost forwarding in large sensor networks. In Proceedings of 10th International Conference on Computer Communications andNetworks.

[13] Wang, X.; Bi, D.; Ding, L.; Wang, S. Agent collaborative target localization and classification in wireless sensor networks. Sensors 2007, 7, 1359–1386.

[14] Chen, J.; Li, S.; Sun, Y. Novel deployment schemes for mobile sensor networks. Sensors 2007,7, 2907–2919.

[15] Xu, Y.; Heidemann, J.; Estrin, D. Geography-informed energy conservation for ad hoc routing. In Proceedings of the 7th Annual ACM/IEEE International Conference on Mobile Computing and Networking.

[16] S. Yao, B. Mukherjee, S.J.B. Yoo, and S. Dixit, “All-Optical Packet-Switched Networks: A Study of Contention Resolution Schemes inan Irregular Mesh Network with Variable-Sized Packets,” Proceedings, SPIE OptiComm 2000, Dallas, TX, pp.235-246, Oct. 2000.

[17] K.K. Ramakrishnan and Raj Jain, “A Binary Feedback Scheme for Congestion Avoidance in Computer Networks,” ACM Transactions on Computer Systems, vol.8, no.2, pp.158-181, May 1990.

[18] Guru P.V. Thodime, Vinod M. Vokkarane, and Jason P. Jue The University of Texas at Dallas, Richardson," Dynamic Congestion-Based Load Balanced Routing in Optical Burst-Switched Networks ",USA 2013

[19] Kamal Kumar Sharma, Dr. Harbhajan Singh and Dr. R.B Patel “A Hop by Hop Congestion Control Protocol to Mitigate Traffic Contention in Wireless Sensor Networks” in Proceedings of International Journal of Computer Theory and Engineering, Vol.2, No.6, pp 1793-8201, December, 2010.

[20] Basaran, kyoung-Don Kang, Mehmet H. Suzer “Hop-by-Hop Congestio Control and Load Balancing in Wireless Sensor Networks”, in proceedings of 2010 IEEE 35th conference on Local Computer Networks, pp. 448-455, 2010.

[21] Chieh-Yih Wan, Shane B. Eisenman,Andrew T. Campbell,” Energy-Efficient Congestion Detection and Avoidance in Sensor Networks “,in proceedings of ACM Transactions on Sensor Networks, Vol. 7, No. 4, Article 32,pp.32.1-32.31 2011.

[22] Babak Namazi*, Karim Faez** Department of Electrical Engineering, Amirkabir University of Technology," Energy-Efficient Multi-SPEED Routing Protocol for Wireless Sensor Networks", International Journal of Electrical and Computer Engineering (IJECE)Vol.3,No.2,April2013

Efficient Proposed Framework for Semantic Search Engine using New Semantic Ranking Algorithm

M.M.El-gayar[1]
P.H.D Student,
Information Technology Department,
Faculty of Computer and
Information Science,
Mansoura, Egypt

N.Mekky[2]
Assistant Professor,
Information Technology Department,
Faculty of Computer and
Information Science,
Mansoura, Egypt

A. Atwan[3]
Associate Professor,
Information Technology Department,
Faculty of Computer and
Information Science,
Mansoura, Egypt

***Abstract*—The amount of information raises billions of databases every year and there is an urgent need to search for that information by a specialize tool called search engine. There are many of search engines available today, but the main challenge in these search engines is that most of them cannot retrieve meaningful information intelligently. The semantic web technology is a solution that keeps data in a readable format that helps machines to match smartly this data with related information based on meanings. In this paper, we will introduce a proposed semantic framework that includes four phases crawling, indexing, ranking and retrieval phase. This semantic framework operates over a sorting RDF by using efficient proposed ranking algorithm and enhanced crawling algorithm. The enhanced crawling algorithm crawls relevant forum content from the web with minimal overhead. The proposed ranking algorithm is produced to order and evaluate similar meaningful data in order to make the retrieval process becomes faster, easier and more accurate. We applied our work on a standard database and achieved 99 percent effectiveness on semantic performance in minimum time and less than 1 percent error rate compared with the other semantic systems.**

Keywords—Semantic Search Engine; Ontology; Semantic Ranker; Crawler; RDF;SPARQL

I. Introduction

There is a huge amount of data stored on the Internet that is only useful and helpful if accessed as information, not as pure data. To access information from Internet, we need a ‘smart or intelligent’ search facility. Search engines are the tools to help users to find data from the huge warehouse of web pages. To extract data, most of the search engines use syntax-based search or full-text search methods. Full-text searching is a technique whereby a computer program matches terms in a search query with terms embedded within individual documents in a database [1]. An important issue in full-text searching technique is that “Because full-text searching relies on linguistic matching—matching a word or phrase in a search query with the same word or phrase in a document in the database being searched—it is subject to failure when a variant term exists and is not matched” [2].

Recent syntax based search engines use various techniques to solve the limitations of a syntax-based search such as page ranking and content score [3]. To get web pages ranked by the search engines, website developers use a method called Search Engine Optimization (SEO). Keywords and meta-tags are the main tools used for SEO. These methods enhance the factor of user friendliness and increase the chances of more accurate results, but these are not the ultimate solution. Data searched by a syntax-based search engine has some limitations, including high recall with low precision (e.g. thousands of results in response to one or few keywords).

A semantic web is optimized solution to these challenges. Semantic Web can be defined as some documents linked in such a way so that the data becomes readable and understandable in a meaningful way [4]. One way of viewing this semantic web is that it is a concept of utilizing the Internet in such a way that searching the World Wide Web returns results relevant to the meaning of the search query. On the Semantic information is illustrated via a new W3C model called the Resource Description Framework (RDF). Semantic Search system is a search system for the Semantic Web. Existing Web sites can be utilized by both individuals and computers to trace exactly and gather information available on the Semantic Web. Ontology is the most significant conception used in the semantic web infrastructure, and RDF(S) (Resource Description Framework/Schema) and Web Ontology Languages (OWL) that used to represent ontologies.

In recent years, the Resource Description Framework (RDF) has become a popular protocol for storing web-based data with well-defined meanings,that usedto link data to improve semantic meaning has widened the scope of this protocol. While RDF data is routinely used by many organizations (e.g. gov.uk and bbc.co.uk) its potential to improve semantic searches is now of interest to the database and Internet research communities.[5]

The Semantic Web will maintain more professional discovery, computerization and reuse of data and offer some support for combinational problem that cannot be solved with existing web techniques. At this time of research on semantic web search systems are not in the beginning stage, but unlike the traditional search systems such as Google, Yahoo, and Bing (MSN) and so forth still lead the present markets of search systems.

In this manuscript, we suggest a new semantic ranking algorithm based on HTML parsing and crawling algorithm to handle the search engine challenges discusses in Section-3. Experimental outcomes display that the recommended technique is more efficient to retrieve and sort a large amount of data from huge datasets or data warehouses. The manuscript is categorized as follows Section two (2) discuses related work of semantic systems, and Section three (3)discuses the global challenges that face search engines. Section four (4) described the several component of the recommended framework and recommended ranking algorithm. In Sect. 5, experiments results and analysis will be presented and in the last section conclusions and references will be given.

II. Related Works

Information recovery and retrieval by searching on the web is not a fresh idea but has different problems when it is evaluated to general information retrieval. Dissimilar search systems return different search results due to the differences in indexing and searching process. Google, Yahoo, and Bing have been out there which holds the queries after developing the keywords. They only search information given on the web page, recently, some research group's start distributing results from their semantics-based search engines. Many novel search engines have been developed for the data Web. Most of these systems are focused on RDF document search like (d'Aquin, Baldassarre, Gridinoc, Sabou, Angeletou, & Motta, 2007; Oren, Delbru, Catasta, Cyganiak, Stenzhorn, &Tummarello, 2008) or ontology search like (Ding, Pan, Finin, Joshi, Peng, &Kolari, 2005). Recall that an RDF document serializes an RDF graph; an ontology, as a schema on the data Web, defines classes and properties for describing objects. Although both RDF document search and ontology search are essential for application developers, they can hardly serve ordinary Web users directly. Instead, object-level search is in demand and dominates all other Web queries (Pound, Mika, & Zaragoza, 2010).

Swooglesystem that described in W3Cand Duckduckgohavesome limitations especially regarding user experience, time of query response and storage capacity. As shown in figure 1, Swoogle's architecture can be broken into four major components: SWD discovery, metadata creation, data analysis, and interface.

Swoogle architecture is data centric and extensible: different components work on different tasks independently. Swoogle offers the some services such as search SW terms and documents, i.e. URIs that have been defined as classes and properties and Provide metadata of SW documents and support browsing the Semantic Web. But, Swoogle has some limitations such as poor indexing of documents and long response time of query.[6]

Another example of semantic search engine is Semantic Web Search Engine (SWSE). Following traditional search engine structural design, SWSE contains of crawling, improving data, indexing process and an interface for search to retrieve an information; unlike traditional search engines, SWSE works over RDF Web data tightly also known as Linked Data which implies unique challenges for the system design, architecture, algorithms, implementation and user interface. [7]

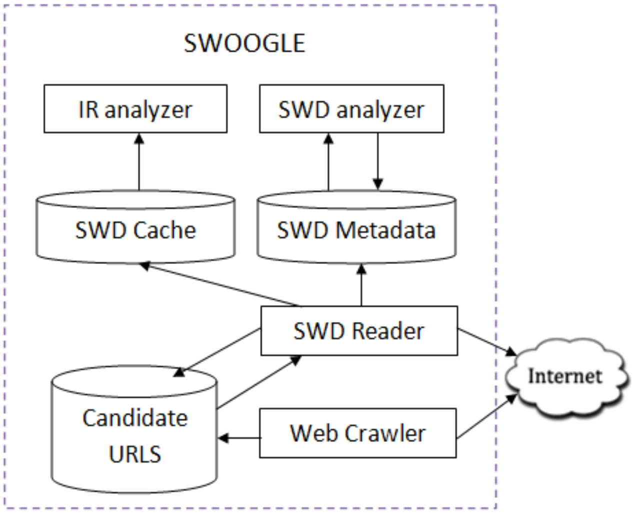


Fig. 1. Swoogle architecture

The sophisticated system design of SWSE loosely follows that of conventional HTML search engines. Figure 2 details the pre-runtime architecture of SWSE system, viewing the mechanism involved in realizing a local index of RDF Web data agreeable for search. Like usual search systems, SWSE includes modules for crawling, ranking and indexing data; on the other hand, there are also factorsspecially designed for treatment RDF data, namely the consolidation module and the reasoning module. The high-level index building process is as follows:

- The crawler recognizes a set of seed URIs. Results analysis for keyword query Bill Clinton Fig. 2. Focus analysis for entity Bill Clinton recover a large set of RDF data from the Web;
- The consolidation module tries to and identical (i.e., equivalent) identifiers in the data
- The ranking module achieves links-based analysis over the crawled data and gains scores indicating the significance of individual factors in the data (the ranking module also considers URI redirections encountered by the crawler when performing the links-based analysis);
- The reasoning modulematerializes new data which is implied by the natural semantics of the input data (the reasoning module also requires URI redirection information to assess the trustworthiness of sources of data);
- The indexing module organizes an index which supports the information retrieval tasks required by the user interface.

But, SWSE has some limitations such as poor ranking of documents because the ranking stage is coming before the indexing stage. Ranking technique is coming independently with data indexed in dataset.

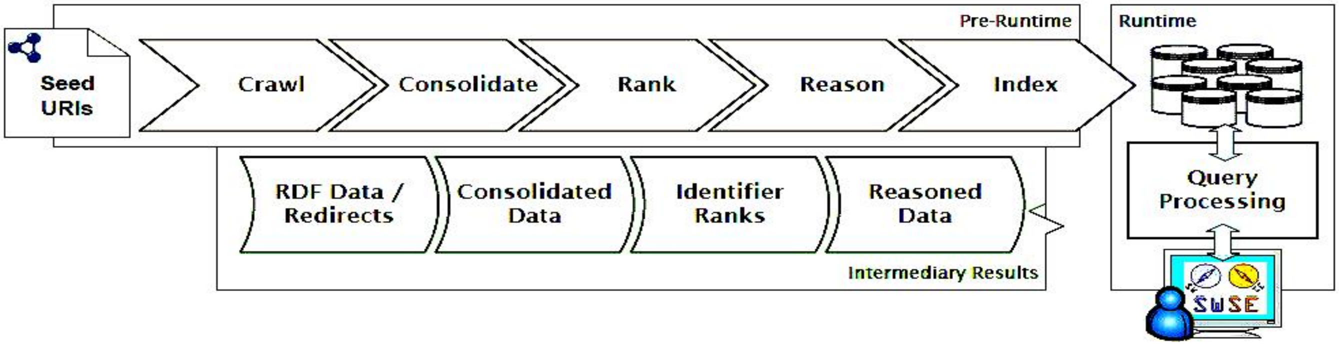


Fig. 2. SWSE Architecture

Another solution model of semantic search engine called Falcons Object Search [8] which firstly is a keyword-based object search engine. For each discovered object, the system constructs an extensive virtual document consisting of textual descriptions extracted from its concise RDF description. Then an inverted index is built from terms in virtual documents to objects for supporting basic keyword-based search. That is, when a keyword query arrives, based on the inverted index, the system matches the terms in the query with the virtual documents of objects to generate a result set. Unfortunately this model is not interested to rank these objects according to query.

This paper investigates some concepts on how the semantic web might be queried in the context of semantic search engines and proposes a framework that facilitates an effective search over the semantic web. Firstly the various factors that influence the search experience over the Internet will be reviewed. Secondly the semantic core technologies necessary to perform a basic search over the Internet will be described - that is RDF and RDF Query Language (SPARQL). Thirdly the academic and social impact of this work is clarified. Finally a proposed framework for a complete search experience for a semantic search engine is presented.

III. Challenges for Semantic Web Search Engine

A semantic web search engine should be able to search data over the Internet with maximum precision and accuracy and should be able to link related data. Semantic search engine should consider the following criteria: user experience, efficiency (performance and associated time) , ranking process, scalability, and cost effectiveness.

A. *User Experience*

A friendly user interface is the mainly significant feature that will increase the user experience. Search engines such as Yahoo, Bing and especially google have all been through a number of enhancements in order to give end-users with the best potential user experience. Even if the results are incomplete or sometimes not accurate due to syntax-only based search algorithms, end-users still remain to these search engines with good user experience. Enhancements to the end-user interface of a semantic query search engine needs important development so that poor input representation of a query will automatically suggest corrections for spelling mistakes and poor grammar, and of course find the best matched results with a high accuracy.

B. *Efficiency*

An efficient semantic search engine’s performance depends upon the size of data to be matched, the request time to server or database and the associated response time. For a semantic web query the finishing time also depends on factors such as delays caused by looking up URL (Uniform Resource Locators) [9], indexing large-scale of data [5], and dealing with query termination and broken links problems[10]. Some smart semantic search systems cannot illustrate their important performance in developing precision and lowering recall. In Ding’s semantic flash system, the source of the search system is based on the top-50 returned results from Google that is not a semantic search engine, which could be low precision and high recall [11].

C. *Ranking Process*

The main idea of the Semantic web search engine is to retrieve the most relevant (most precision) and accurate results in response to a query. Ranking process such as page rank algorithm is a method of rating web pages so that the web pages with the highest ranking are presented at the top of a list of search results [12]. This is a challenging task given that there are “more than 12.3 billion web pages in the World Wide Web” at the time of writing [13], and a single user query on a search engine may return millions of results. It is consequently critical that the search engine can sort and rank the retrieved documents effectively in order of either relevancy or authenticity. There are a number of techniques used by search engines to rank page results. Page ranking techniques can organize results in order of relevance, significance and content score [13].

D. *Scalability*

Scalability within the perspective of a search engine is the ability of a system to handle a hurriedly growing amount of data. Relational database management systems have frequently shown that they are very efficient with the structure of relational data but not scale well [1]. However scalability for data in a semantic web presents additional challenges because of the open source of the RDF protocol.

E. *Cost Effectiveness*

A high-quality search system must give a solution that is cost effective. Due to the open source of RDF, the queries can be quite expensive while dealing large data sets. For efficient data retrieval, search engines use indexing techniques at the cost of additional storage. As a semantic search engine processes an open structure like RDF, complex queries can be very expensive to process. Only a few solutions have been proposed to solve this issue (e.g. caching process) or use other technique in indexing process. Some efforts have been made to introduce cost effective search algorithms such as the SPARQL as search technique[5]. In section 4 which discuss the proposed framework that handle and overcome these challenges with suitable user interface like Google search engine that overcome user experience. In addition to standard crawler model, indexing algorithm to overcome scalability and cost effectiveness. Finally we introduce the proposed ranking algorithm that considered as main contribution of this paper to overcome ranking and efficiency problems.

IV. Proposed Framework

Proposed Framework is designed in a modular fashion and logically composed of two separate phases, the online phase (retrieving phase - which user deal directly with the server) and offline phase (Crawling, indexing and ranking phase - which server deal not directly with any users). This section describe in details the two phases that described in figure 3 and figure 4.

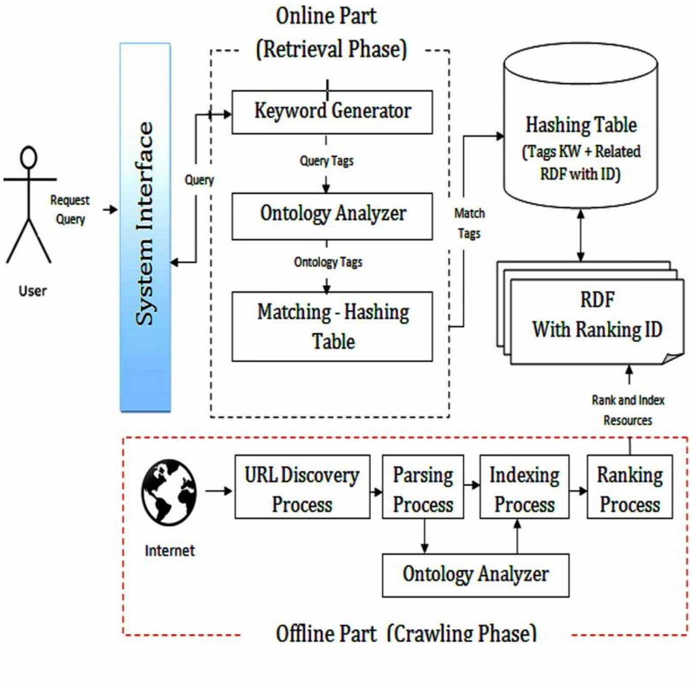


Fig. 3. Logical Architecture of Proposed Framework

A. *Offline Phase (Crawling Phase)*

In this module, three steps are used. The First Step is crawling process that contains two sub-steps, the first sub-step is URL Discovery Process (pre-crawling process) and the second sub-step is officially crawling process. The second step is indexing process that contains also two sub steps, the first sub-step is parsing HTML document and extract useful information, however the second sub-step is officially indexing process. The third step is semantic ranking process using proposed semantic ranking algorithm that discussed in algorithm #2.

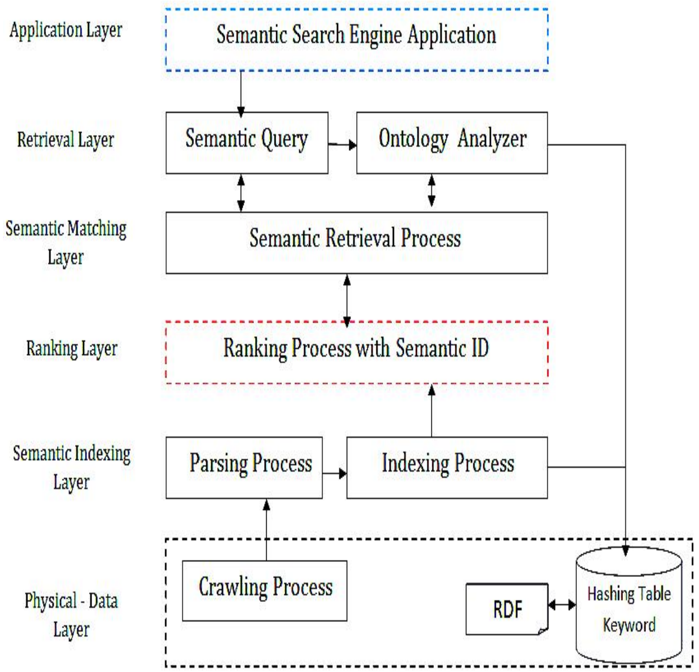


Fig. 4. A Layer Diagram of Proposed Framework

1) *Crawling Process*

The First Step, discover URLs from World Wide Web by using URL Discovery Process and crawl these URLs to find related keywords and meta data about queries with most frequently used in pages with heights page rank algorithm. In crawling module, crawling algorithm is used that discussed in algorithm #1.

```
ALGORITHM#1: Page_Crawling
INPUT : URL
OUTPUT : arrayOfKeywords, Position, Title, URLs
Procedure Page_Crawling
Begin
B_url = getBaseURL (URL)
P = Download(URL)
Urls = ExtractOutgoingURL in p with B_url
ForeachUrls as Url do
    Position = ParsePosition(Url)
    IF pageTitle is not Null
        Title = ExtractPageTitle(Url)
    IF pageMetaTags is not null
        arrayOfKeywords[Url][Title][Position] = ExtractPageMetaTags(url)
    Else
        Continue
 End Foreach
IF  arrayOfKeywords =  null
  Page_Crawling(Urls)
End IF
End Procedure
```

Algorithm.1. Crawling Algorithm

Since our architecture is currently implemented to index RDF/XML, we would feasibly like to maximizethe ratio of HTTP lookups which result in RDF/XML content; i.e., given the total HTTP lookups as L, and the total number of downloaded RDF/XML pages as R. In order to reduce the amount of HTTP lookupswasted on non-HTML/ RDF/XML content, we implementthe following heuristics:

a) firstly, we blacklist non-http protocol URIs;

b) secondly, we queue URIs with common extensions that are highly unlikely to return RDF/XML/HTML/PDF and we blacklist the extensions of images or videos like (jpg, gif, AVI, MKV, , etc.)

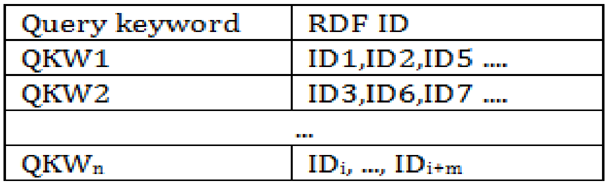
c) thirdly, we check the returned HTTP header and only retrieve the content of URIs reporting Content-type: HTML or application/rdf+xml

2) *Indexing Process*

The first sub-step in this process is parsing HTML document to extract useful information such as meta data, title, time stamp, author, keywords and related URLs to be crawled again. The second sub-step is indexing results from the crawling process in the first step into database. But with huge data cannot be indexed into relational database because of scalability, storage, sorting, semantic and retrieval issues. Due to semantic, we use RDF instead of relational database, but we face the unstructured problem. So, we can use hybrid dataset from relational and RDF (based on XML). Relational dataset used for storing hash tables (such as table shown in figure 5.a and figure 5.b) that contains keywords and related RDF ID. RDF used for storing all data that extracted from parsing process.

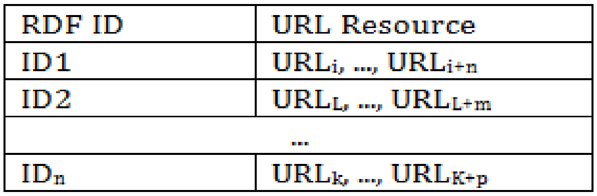
TABLE I. DESCRIPTION OF 8-BITS CHILD DATA OF SEMANTIC ID

Priority	**Title**	**Meta Tag**	**Duplication in text**	**Body URL**	**Header URL**	**Left or Right URL**	**Footer URL**
Priority Of text > 10	If Title exits	If Meta tag exits	If Duplication in text	If URL in bodyof the page	If URL in headerof the page	If URL in left or rightof the page	If URL in footer of the page



Query keyword	RDF ID
QKW1	ID1,ID2,ID5 ....
QKW2	ID3,ID6,ID7 ....
...	
QKW_n	ID_i, ..., ID_{i+m}

Fig. 5. a. Hash table of Query Keyword



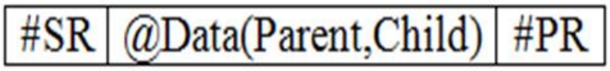
RDF ID	URL Resource
ID1	URL_i, ..., URL_{i+n}
ID2	URL_L, ..., URL_{L+m}
...	
ID_n	URL_k, ..., URL_{K+p}

Fig. 5. b - Hash table of URL Resources

3) *Semantic Ranking Process*

The idea of developing ontology-based annotations for information is not a fresh idea; semantic search system would consider keyword impression and would return a page only if keywords (or synonyms, homonyms, etc.) are founded within the page and linked to associate concept. The success is measured by the "predictability" that the user would have guessed such anrelationship exists.

The ranking strategy assumes that given a query "Q", and a page "p", it is possible to build a query sub-graph $G_{Q,p}$exploiting the information available in page annotation according to ranking ID from the data stored in RDF. Ranking algorithm uses global ID that consist of Semantic ID, Child Data, Parent Data and page rank as shown in figure 6.



#SR	@Data(Parent,Child)	#PR

Fig. 6. 32 bits Semantic ID

Ranking algorithm can rank data according to parse and index step that be implemented in the first phase as the following:

- Child partthat called (URL Resource) consist of(8 bits) 2 digits hexadecimal - with in right 4 bits for position and left 4bits for data as discussed in algorithm#3 (Child_Data_Part_in_Semantic_Rank). This ID expresses the child RDF data from another parent RDF Data.

- Parent Partconsidered as (Semantic Id of parent URL)Semantic ID for parent URL - (16 bits) 4 digits hexadecimal. This ID expresses the parent RDF data of the child RDF data in the above.

- SR (Semantic Rank) in figure 5 used as a header (4 bits) 1 digit hexadecimal as discussed in algorithm#2 (Create_Semantic_ID_of_Child_URL). This ID is global semantic ID that express of all these IDs and can be retrieved faster than others.

- PR (Google Page Rank) that famous ranking algorithm that used in Google according to page that crawled in the first phase. This ID uses values from 0 to 10 - (4 bits) 1 digit hexadecimal

B. *Online Phase (Retrieving Phase)*

In this module as shown in figure 7, three steps are used. The First Step is keyword generator process that used to split the query request into distinct words. The second step is ontology analyzerprocess that help system to recommend tags to query keywords where tags will have associated ontologies. The retrieving of ontologies from an online store or library in order to tag a word is a lengthy process that will have a cost in terms of efficiency. The third step is matching process that used to match the query tags against the keywords stored in hashing table. After matching process, system can detect the related RDF with related URL resources.

TABLE II. DESCRIPTION OF 4 BITS SEMANTIC RANK (SR)

(PR Child + SR Parent) > 15	Relevance feedback > 10	Description >= 4	Position >= 4

ALGORITHM#2: Create_Semantic_ID_of_Child_URL

INPUT : 8bits Child_Data, Child_URL

OUTPUT : Semantic ID

```
1:  Procedure Create_Semantic_ID_of_Child_URL
2:  Begin
3:  Create 4bits Semantic Rank of Child_URL (SR) from step 4 to step 15
4:  Get 4bits of Child_Data from algorithm #3 Child_Data_Part_in_Semantic_Rank
5:  If Right 4bits of Child_Data hexadecimal number greater or equal than 4
6:        Set least significant bit of 4bits Semantic Rank (SR)  to 1
7:  End If
8:   If Left 4bits of Child_Data hexadecimal number greater or equal than 4
9:        Set second bit from least significant bit of 4bits Semantic Rank (SR)  to 1
10:   End If
11: If Relevance feedback of users greater than 10
12:       Set third bit from least significant bit of 4bits Semantic Rank (SR)  to 1
13: End IF
14: If (Semantic Rank of Parent URL + Page Rank of Child URL) greater than 15
15:       Set most  significant bit of 4bits Semantic Rank (SR)  to 1
16: End If
17: Create Semantic ID from step 17 to step
18: most significant 4bits is Semantic Rank
19: Left-Middle 8bits is child_data
20: Right-Middle 16bits is Parent_Semantic_ID
21: Least significant 4bits is page rank number in binary of child URL
22: Return Semantic ID of child URL
23: End Procedure
```

Algorithm.2. Creation of Semantic ID from Child URL

ALGORITHM#3: Child_Data_Part_in_Semantic_Rank

INPUT : Child_URL, arrayOfKeywords, child_Url_position, Page_Title, Ontology_text

```
1:  Procedure Child_Url_Semantic_Rank
2:  Begin
3:  Create Right_four_ bits of child data from step 4 to step 12
4:  If child_Url_position in  Footer_position
5:        Set Right_four_ bits of child data to 0001 binary = 1 in hexadecimal
6:  Else If child_Url_position in  Left_position or Right_position
7:        Set Right_four_ bits of child data to 0010 binary = 2 in hexadecimal
8:   Else If child_Url_position in  Header_position
9:        Set Right_four_ bits of child data to 0100 binary = 4 in hexadecimal
10:  Else If child_Url_position in  Body_position
11:       Set Right_four_ bits of child data to 1000 binary = 8 in hexadecimal
12: End IF
13: Create Left_four_ bits of child data from step 14 to step 25
14: If duplication of keyword in text exists
15:       Set least significant bit in Left_four_ bits of child data to 1
16: End If
17: If arrayOfKeywords not equal null
18:       Set the second bit of least significant bit in Left_four_ bits of child data to 1
19: End If
20: If Page_Title not equal null
21:       Set the third bit of least significant bit in Left_four_   bits of child data to 1
22: End If
23: If arrayOfKeywords exits in Ontology_text
24:       Set most significant bit in Left_four_ bits of child data to 1
25: End If
26: Return 8bits of Child Data part in Semantic ID
27: End Procedure
```

Algorithm.3. Child Data Part in Semantic Rank

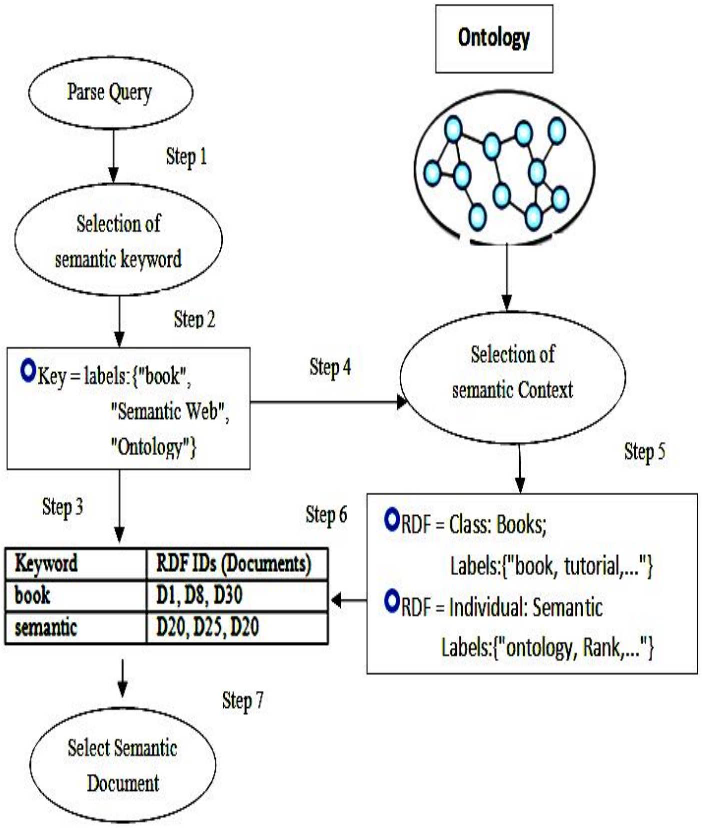

Ontology
Parse Query
Step 1
Selection of semantic keyword
Step 2
Key = labels:{"book", "Semantic Web", "Ontology"}
Step 4
Selection of semantic Context
Step 5
Step 3
Step 6
RDF = Class: Books; Labels:{"book, tutorial,..."}
RDF = Individual: Semantic Labels:{"ontology, Rank,..."}
Keyword
RDF IDs (Documents)
book
D1, D8, D30
semantic
D20, D25, D20
Step 7
Select Semantic Document

Fig. 7. Flow Diagram of Online Retrieval Phase

V. DATA SET AND EXPERIMENT RESULTS

A. Machine Specifications Used in Testing

The machine specifications are Corei7 CPU, 2GB RAM, 500GB Hard Disk and Windows 7. The Software specifications are Apache Server (localhost) with PHP version 5.3 and MYSQL Database version 5.5.

B. Data Collection

A standard assessment data gathering should be not influenced towards any exacting system or towards aexact domain, as our objective is to assess general idea entity search over RDF data. Therefore, we needed a collection of documents that would be a realistically large estimation to the amount of RDF data accessible 'live' on the Web and that contained related information for the queries, while concurrently of a size that could be convenient by the resources of a research groups. We chose the 'Billion Triples Challenge' (BTC) 2011 data set, a data-set created for the Semantic Web Challenge in 2011 as displayed in table 3.

TABLE III. BILLION TRIPLE CHALLENGE 2011 DATASET

Description	Billion Triples Challenge
Author	Andreas Harth
Size	20GB gzipped
Download	http://km.aifb.kit.edu/projects/btc-2011/

C. Query sets

The Semantic Search Challenge comprised two tracks. The Entity Search track is identical in nature to the 2010 challenge. However, we created a new set of queries for the entity search task based on the Yahoo! Search Query Tiny Sample v1.0 dataset. We selected 10 queries which name an entity explicitly and may also provide some additional context about it.

D. Proposed System Evaluation

Table 4 is datasheet that describe the result of retrieval process after crawling process of 10 samplesfrom proposed system. Datasheet in table 4 shows summation of total results for each query included relevant result (performance) and irrelevant (error) result. Figure 8 shows the performance and error chart of the retrieval process.

Table 5 is datasheet that describe total time of retrieval process in seconds of 10 samples from proposed system. Figure 9 shows the time chart of the retrieval process.

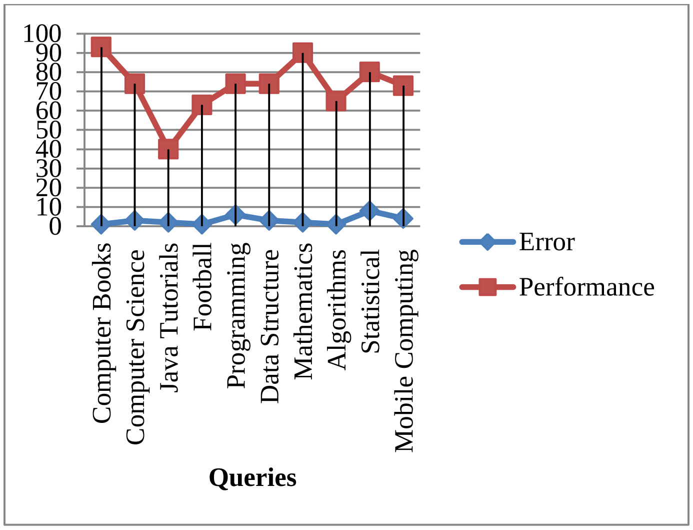


Fig. 8. Performance and Error Chart for Retrieval Process

TABLE IV. RELEVANT AND IRRELEVANT RESULTS OF RETRIEVAL PROCESS AFTER CRAWLING PROCESS ON 10 QUERIES AS SAMPL

10 QueriesSamples	Crawler Result After ontology Analyzer	Retrieval Error (Irrelevant Result)	Retrieval Performance (Relevant Result)
Computer Books	94	1	93
Computer Science	77	3	74
Java Tutorials	42	2	40
Football	64	1	63
Programming	80	6	74
Data Structure	77	3	74
Mathematics	92	2	90
Algorithms	66	1	65
Statistical	88	8	80
Mobile Computing	77	4	73

TABLE V. TIME FOR EACH QUERY OF RETRIEVAL PROCESS ON 10 QUERIES AS SAMPLE

10 Queries Samples	Total Time of Crawling and Filtering in ms
Computer Books	30
Computer Science	50
Java Tutorials	20
Football	40
Programming	60
Data Structure	40
Mathematics	40
Algorithms	40
Statistical	60
Mobile Computing	50

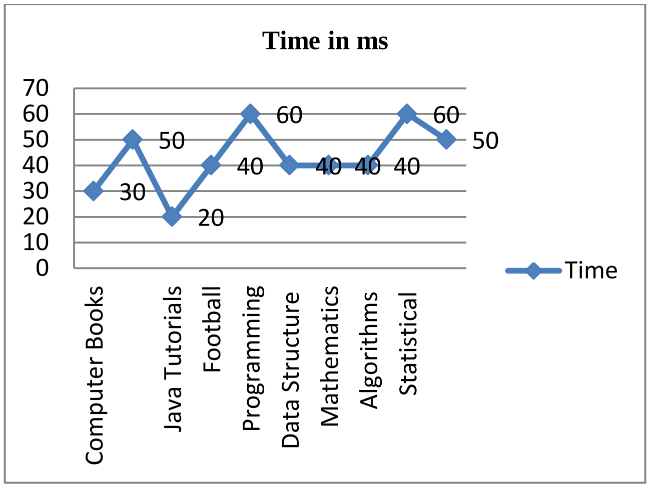


Fig. 9. Time Chart for Retrieval Process

VI. Conclusion

The topic of the semantic search engine has attracted large interests both from industry and research with resulting variety solutions in different tasks. There is no standardized framework that helps to monitor and stimulate the progress in this field. In this paper, Four standard tasks of semantic search engine are discussed including crawling, indexing, ranking and finally retrieving task.

We focus on ranking phase that considered as the main contribution of this paper. New ranking algorithm is produced to rank similar meaningful data after indexing phase. In addition to, data retrieval process become faster, easier and more accurate. The performance achieved with 99 percent relevant results in maximum time 60 ms and 1 percent only for irrelevant results. The proposed framework and ranking algorithm can be further developed for future use in detecting more accurate semantic information from social networks in a short time.

References

[1] J. Beal, “Weaknesses of Full text search”, The Journal of Academic Librarianship, vol. 34, Number 5, 2008, pp. 438-444.

[2] J. Beal, “Geographical research and the problem of variant place names in digitized books and other full-text resources” Library Collections, Acquisitions, and Technical Services, vol. 34, Issues 2–3, 2010, pp. 74-82.

[3] M. P. Selvan, C. A. Sekar and P. A. Dharshini, “Survey on Web Page Ranking Algorithms”, International Journal of Computer Applications, vol. 41, no.19, Published by Foundation of Computer Science, March 2012.

[4] T. Berners-Lee, “Weaving the Web”, pp. 2-5. The Original Design and Ultimate Destiny of the World Wide Web by its Inventor. Harper: San Francisco, 1999.

[5] L. Chang, W. Haofen, Y. Yong and X. Linhao, “Towards Efficient SPARQL Query Processing on RDF Data”, Tsinghua Science & Technology, vol. 15, issue 6, Dec. 2010, pp. 613-622.

[6] Tim Finin, Yun Peng, R. Scott, Cost Joel , " Swoogle: A Search and Metadata Engine for the Semantic Web " University of Maryland Baltimore County, Baltimore MD 21250, USA, 2011.

[7] Aidan Hogan and Andreas Harth and Jürgen Umrich and Sheila Kinsella and Axel Polleres and Stefan Decker: "Searching and Browsing Linked Data with SWSE: the Semantic Web Search Engine." In JWS 9(4), 2012.

[8] Gong Cheng and Yuzhong Qu: "Searching Linked Objects with Falcons: Approach, Implementation and Evaluation." In IJSWIS 5(3), 2009.

[9] O. Harting and F. Huber, F., “A main memory index structure to query linked data”. Proc. 4th Linked Data On the Web (LDOW11), March 2011.

[10] O. Harting and J. Freytag, J “Foundations of traversal based query execution over linked data”, Proc. 23rd ACM conference on Hypertext and social media (HT '12). ACM, New York, NY, USA, 2012, pp. 43-52.

[11] D. Ding, J. Yang, Q. Li, L. Wang, and W. Liu, "Towards a flash search engine based on expressive semantics," in Proceedings of WWW Alt.'04 New York, 2004, pp. 472-473.

[12] G. P. Schneider and J. Evans, “New perspectives on the Internet”. Ohio : South-Western, 2012.

[13] M. P. Selvan, C. A. Sekar and P. A. Dharshini, “Survey on Web Page Ranking Algorithms”, International Journal of Computer Applications, vol. 41, no.19, Published by Foundation of Computer Science, March 2012.

Self-Healing Hybrid Protection Architecture for Passive Optical Networks

Waqas A. Imtiaz
Department of Electrical Engineering
IQRA National University
Peshawar, Pakistan

M. Waqas
Department of Electrical Engineering
IQRA National University
Peshawar, Pakistan

P. Mehar
Department of Electrical Engineering
IQRA National University
Peshawar, Pakistan

Yousaf Khan
Department of Electrical Engineering
IQRA National University
Peshawar, Pakistan

***Abstract*—Expanding size of passive optical networks (PONs) along with high availability expectation makes the reliability performance a crucial need. Most protection architectures utilize redundant network components to enhance network survivability, which is not economical. This paper proposes new self-healing protection architecture for passive optical networks (PONs), with a single ring topology and star-ring topology at feeder and distribution level respectively. The proposed architecture provides desirable protection to the network by avoiding fiber duplication at both feeder and distribution level. Moreover, medium access control (MAC) controlled switching is utilized to provide efficient detection, and restoration of faults or cuts throughout the network. Analytical analysis reveals that the proposed self-healing hybrid protection architecture ensures survivability of the affected traffic along with desirable connection availability of 99.9994 % at minimum deployment cost, through simple architecture and simultaneous protection against failures.**

Keywords—passive optical network; protection; star-ring topology; reliability; CAPEX

I. INTRODUCTION

Passive optical network (PON) consists of a long feeder fiber (FF) between an optical line terminal (OLT) at the central office (CO) and remote node (RN) at the subscriber premises. RN contains an 1: N optical coupler (CPR), which connects N optical network units (ONUs) through dedicated distribution fibers (DFs). PONs are anticipated to solve the last mile bottleneck between high-speed core/metropolitan networks and access domain, owing to its significant advantages, like high subscribers count, minimum capital expenditure (CAPEX) due to shared FF, less operational expenditure (OPEX) due to passive components between OLT and ONU, and support for high data rates [1].

Efficient operation of PONs requires high availability along with fault detection and restoration for smooth transmission of data between OLT and subscriber premises [1-3]. Therefore, it is imperative for the network operators to develop simple and efficient protection architecture, which is reliable and economical for the common end user. Evolution of PON protection architectures began with ITU-T G.983.1 standards in the form of type A, B, C and D schemes. Type A, and B concentrates on protecting the feeder level only, while type C, and D protects the entire PON by duplicating components throughout the network. Consequently, type A, and B fail to provide the desirable availability since no protection at provided at the distribution level. Whereas, type C and D provides high reliability performance, but unfortunately they require duplication of entire PON, which significantly increase the overall capital expenditure (CAPEX) for the network that is shared among limited number of subscribers [2]. Moreover, it is also observed that use additional, redundant, components to provide the necessary protection reduce the network reliability [4].

Ring-based protection schemes including single and double ring architectures have been proposed to negate the issue of fiber duplication in PONs. In [5], a single ring based fiber is placed instead of the long FF, which changes into tree topology with two active branches in case of failure. However, the un-protected star shaped distribution network reduce its availability, and hence, feasibility for the access domain [3]. A similar single ring-star architecture is proposed in [6], with the use of relatively simple components. However, it also fails to provide the necessary protection to ONUs, which significantly reduces its reliability for access domain.

Ring based fault detection and restoration architecture for PONs is proposed in [7-8], which utilize CPRs for high availability of the network along with ring extension between multiple RNs. However, single CPR per ONU introduce serious power budget issues, which reduces the ability of single ring based PONs to support large number of subscribers [3]. Tree and star-ring architectures are proposed in [9-10] for protection at both feeder and distribution levels. However, duplication of FF between OLT and RN significantly increases the CAPEX, which is not viable for the common end user in access domain.

This paper proposes new hybrid protection architecture for PONs with a ring-star-ring topology, to ensure ubiquitous transmission at both feeder and distribution fibers. Moreover, MAC controlled switching add an efficient self-healing capacity to the proposed architecture. Operation of the proposed self-healing hybrid protection architecture is thoroughly analyzed for failures or cuts at both feeder and distribution fiber followed by reliability analysis through availability modeling technique in [3]. It is observed that the proposed system efficiently restores the flow the traffic in case of failure at both feeder and distribution fibers, along with the provision of desirable availability, 99.999%, through simple architecture and simultaneous protection against failures.

II. Protection Architecture

The proposed self-healing hybrid protection architecture for PONs is shown in Fig. 1. All services are originating from OLT at the CO, which consists of a transmitter and receiver module. Output of OLT is connected to an optical circulator (OC_{oc}), which splits the traffic for access domain and OLT receiver module respectively. Traffic from OC_{oc} is fed into an erbium doped fiber amplifier (EDFA), which amplifies the optical signal to meet the power budget requirements of the proposed architecture. Output of EDFA is split into two identical paths through $1:2$ CPR_{co}. Path “a” of CPR_{co} extends the FF in the clockwise (CW) direction, while path “b” in the counter clockwise (CCW) direction. Path “b” also contains a 1:2 optical switch (OS_{co}), which activates in case of failure. Port 1 of OS_{co} is connected to the ground, while port 2 extends the FF in the CCW direction. Under normal mode of operation, OS_{co} is at position 1, and OLT handles the traffic through path “a”. In case of failure, OLT medium access control (MAC) layer flips the switch towards port 2, which immediately restores the flow of traffic. MAC controlled switching eliminates the needs of extra arrangement at OLT. Moreover, efficient algorithms can significantly eliminate delays and unnecessary switching in case of failures.

Feeder ring (FR) is formed by connecting both paths through a $2:2$ passive CPR_{x1} at RN_x as shown in Fig. 1, where port 1 and port 3 of CPR_{21} in RN_2 connect both paths of the FF to form a single ring topology at the feeder level. While port 2 and port 4 of CPR_{21} extends the FR into the access arena through another $2:M$ CPR_{x2}. CPR_{22} is used to connect M number of ONUs with the FR through dedicated DFs as shown in Fig. 1. If X represents the number of RNs, then the proposed system can support up to $X \times M$ number of ONUs.

Each ONU contains a $1:2$ CPR with three ports as shown in Fig. 1. Port 1 is used to connect each ONU with its dedicated DF, while port 2 is fed into 2:1 OS_{onu}. Port 3 of the CPR is used to form a ring at the distribution level (DR), which ensures ubiquitous supply of traffic in case of failure at the DFs through an efficient star-ring topology. By default, OS_{onu} is at position 1 and all traffic is delivered through DFs. In case of failure at the DF, MAC layer of effected ONU changes the switch position, which immediately converts the flow of affected traffic from DF to the backup DR. Output of the OS_{onu} is fed into an OC_{onu}, which distributes the traffic in both ONU receiver and transmitter modules.

III. Protection Analysis

A. Feeder ring failure

It is assumed that the proposed PON is under normal mode of operation and all traffic is delivered through path “a”. If failure occurs at point “f” as shown in Fig. 2. OLT will not receive any traffic from ONUs placed beyond the point of failure (PoF) and vice versa. Therefore, OLT will initiate its recovery mechanism and send DISCOVERY GATE (DG) packets to check the status of FR in terms of ONUs registration. If OLT receive registration messages from one or all ONUs of each RN, it will continue its normal operation through path “a”.

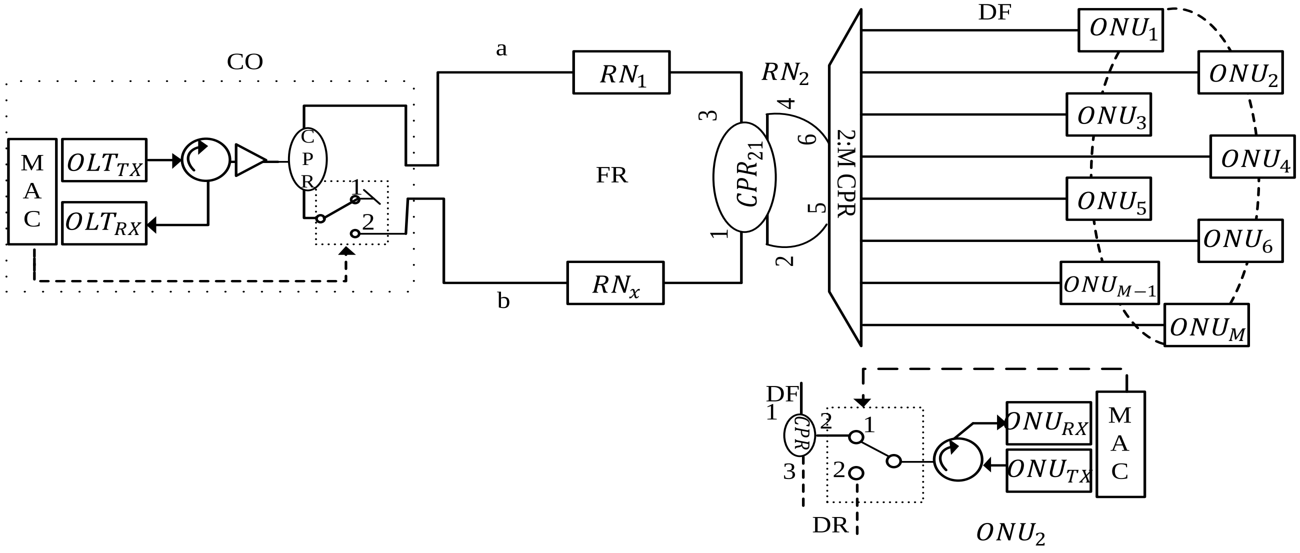


Fig. 1. Proposed self-healing hybrid PON

In case of negative registrations, OLT confirms any failure or cut across the FR through several atempts. On repeated negative registrations, OLT MAC layer will perform the switching operation, which converts the single ring topology into tree topology with two active branches as shown in Fig. 2.

Path "a" will deliver traffic to ONUs before the PoF, while path "b" will connect OLT with ONUs beyond the PoF.

It must be observed that ONUs will also initiate their recovery mechanism in case of failure at point "f". In order to avoid any unnecessary switching the distribution level, the number of attempts for fault detection at the ONUs must be twice as compared to the number of attempts at OLT.

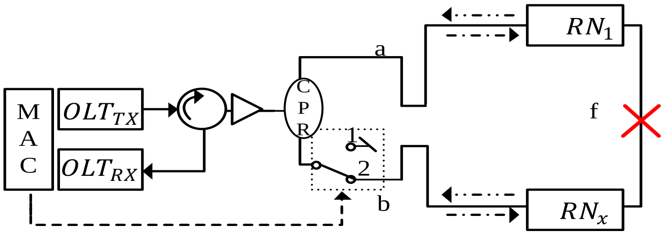


Fig. 2. Fault detection and recovery at FR

B. *Distribution fiber failure*

It is assumed that the OS_{onu} is at position 1, and ONU_2 sends and receive all traffic through the DF . If a failure or cut occurs at point "c", ONU_2 will stop to receive any traffic from the OLT as shown in Fig. 3. At the same time, OLT will cease to receive any upstream traffic from the disconnected ONU_2. Thus, ONU_2 will initiate its recovery mechanism, and send DG packets to check the status of DF. In case of successful registration messages from OLT, ONU_2 will continue its normal operation through its respective DF. If no registration messages are received from OLT, ONU will attempt to confirm the fault or cut by sending several DG packets.

In case of successive negative registrations, ONU_2 MAC layer will flip the switch position to connect ONU_2 with the backup DR. Thus, all traffic will immediately transfer from the faulty DF, to ONU_3 DF through DR between ONU_2 and ONU_3.

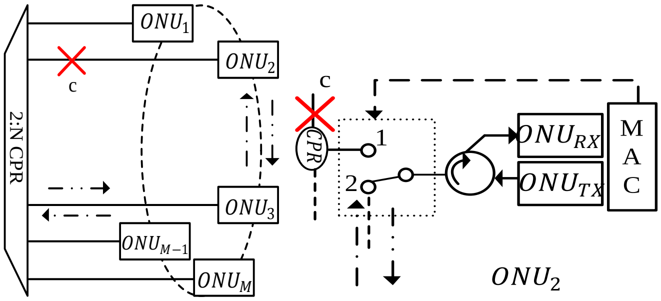


Fig. 3. Fault detection and restoration at DF

IV. Analytical Analysis

A. *Network capcity*

It is assumed that each RN_x supports M number of ONUs, then the total ONUs N can be written as:

$$N = X \times M \tag{1}$$

If L represents the length of the fiber, P_{CO} represents the power losses at the CO, P_{FR} represents the power losses at the FR including RNs, P_{RX} is the power drop at the distribution network, and R_{sen} represents the receiver sensitivity, then the downstream power budget can be written as:

$$P_{CO} = P_T - P_{OLT} - P_{OC} + P_{EDFA} - P_{CPR} - P_{OS} \tag{2}$$

$$P_{FR} = -\alpha L - XP_{CPR} - 10log_{10}(M) \tag{3}$$

$$P_{RX} = -\alpha L - P_{CPR} - P_{OS} - P_{OC} - P_{ONU} \tag{4}$$

Efficient operation of the proposed protection architecture requires that:

$$P_{CO} + P_{FR} + P_{RX} \geq R_{sen} \tag{5}$$

Now, rearranging (5) can determine the number of RNs, which can be written as:

$$X \leq P_T - R_{sen} - 16.25 - 10log_{10}(M) + P_{EDFA}/3 \tag{6}$$

Table 1 shows the description and specifications of optical components in Eq. (2)-(4). Total number of ONUs, in the proposed PON are shown in Fig. 4, in relation to M and R_{sen} at $P_T = 0\ dBm$, and $P_{EDFA} = 25\ dB$. It can be observed that the proposed protection architecture can support a large number of subscribers, owing to the use of 2:M CPR at RNs. This significantly reduce the power budget issue and allow the system to support maximum ONUs within power budget threshold.

TABLE I. Optical Components Description and Values

Symbols	Description	Values
P_T	Optical source power	$0\ dBm$
P_{OLT}	Power loss at OLT	$3\ dB$
R_{sen}	Receiver sensitivity	$-20 \sim -24\ dBm$
P_{sa}	Power loss at switch	$0.5\ dB$
P_{edfa}	Amplifier gain	$25\ dB$
P_{oc}	Circulator loss	$0.25\ dB$
P_{onu}	Power loss at ONU	$2\ dB$
α	Propagation loss	$.25\ dB/km$
$10log_{10}(S)$	Coupler loss (s=2)	$3\ dB$

B. *Reliability Analysis*

This section evaluates and compares reliability of the proposed architecture with protection schemes in [2], and [6] while using availability modeling technique in [3], and [11]. Figure 5 shows the reliability block diagrams (RBDs) of the protection architectures, where network components (including fibers) are arranged in series and parallel combination. Series arrangement refers to the unprotected components of the network. While, parallel combination represents the protected components of the network [3].

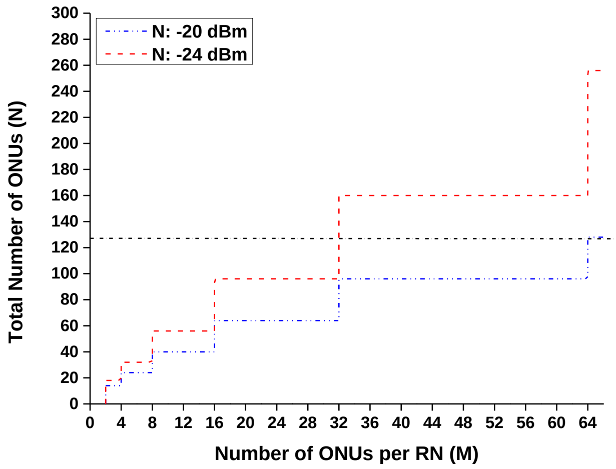
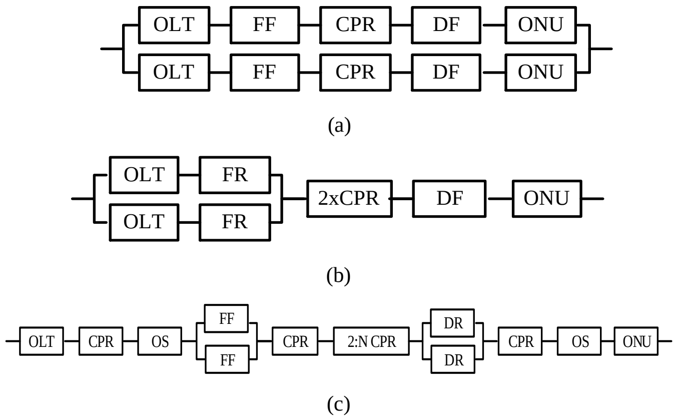


Fig. 4. Total number of ONUs at different values of R_{sen}



(a)

(b)

(c)

Fig. 5. Reliability block diagrams of different protection schemes: (a) scheme in [2] (b) scheme in [6] (c) proposed architecture

The characteristic parameter for each component in RBDs is the asymptotic unavailability (U_i) where i represent a component in the PON. Hence, the system availability with m components is given by:

$$A = 1 - \sum_{i=1}^{m} U_i \qquad (7)$$

Based on RBDs, availability of considered protection architecture can be written as:

$$A_{[c]} = 1 - [(U_{OLT} + U_{FF} + U_{CPR} + U_{DF} + U_{ONU}) \times (U_{OLT} + U_{FF} + U_{CPR} + U_{DF} + U_{ONU})] \qquad (8)$$

$$A_{[6]} = 1 - [(U_{OLT} \times U_{OLT}) + (U_{FR} \times U_{FR})] + (2 \times U_{CPR}) + U_{DF} + U_{ONU}] \qquad (9)$$

$$A_P = 1 - [U_{OLT} + U_{CPR} + U_{Os} + (U_{FF} \times U_{FF}) + U_{CPR} + U_{2:NCPR} + (U_{DF} \times U_{DF}) + U_{CPR} + U_{Os} + U_{ONU}] \qquad (10)$$

Connection aavailability is calculated for 20 km fiber at the feeder level and 5 km fiber at the distribution level, while using system unavailability parameters in Table 2 [3][11]. Figure 6 shows that type C protection scheme in [2] provides higher connection availability followed by our proposed architecture, 99.9994 %, which is above the minimum criteria for desirable connection availability (5 nines) [3]. While the similar ring-star topology in [6] fails to provide the desired reliability, due to its lack of protection at the distribution level.

Feasibility of the proposed architecture is further analyzed over CAPEX analysis, through components cost and analysis technique in [3]. Following parameters are adapted for fair comparison of the considered scheme. Tree based PONs contain 20 km FF and 5 km DF, hybrid ring-star PONs contains 20 km FR, 5 km DF and 1 km fiber between adjacent ONUs, single PON with 16 ONUs is considered in both tree and hybrid PONs, and two RNs are considered in hybrid ring-based PONs with eight ONUs each. Moreover, cost of fiber burying is ignored due to high variation. Cost comparison reveals that that our scheme provides desirable connection availability with minimum deployment cost of 2556 USD. Whereas, type C protection scheme duplicates the entire PON from OLT and ONU, which significantly increase their cost to 4218 USD per subscriber as compared to the proposed self-healing hybrid protection architecture.

TABLE II. UNAVAILABILITY AND COST OF DIFFERENT SYSTEM ELEMENTS

System Elements	Unavailability
OLT (Tx and Rx)	$5.12e^{-7}$
ONU (Tx and Rx)	$1.54e^{-6}$
Optical Circulator	$3e^{-7}$
Optical Switch	$1.2e^{-6}$
RN (2:2 CPR)	$3e^{-7}$
RN (2:N CPR)	$7.2e^{-7}$
Fiber (/Km)	$1.37e^{-5}$

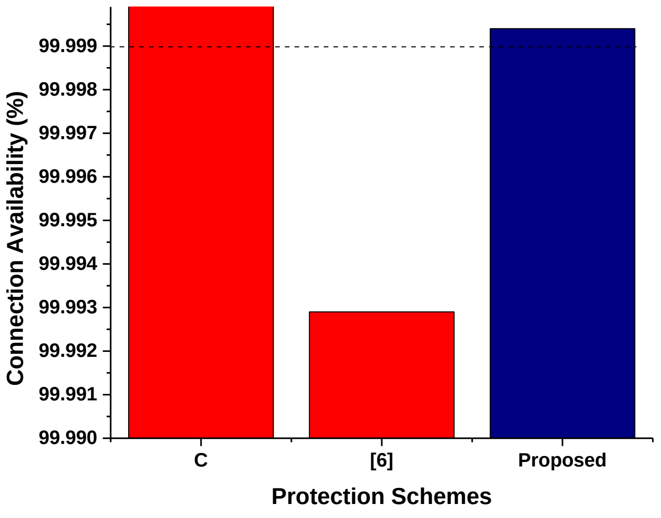


Fig. 6. Relaibility analysis of proposed scheme in comparison with different solutions

V. CONCLUSION

Efficient and economical protection architectures are characterized as an important factor for the adaption of PONs in access domain. This paper proposes new protection architecture for PONs, which ensures network survivability along with reduction in deployment cost. The proposed architecture consists of a single ring topology at the feeder level and a star-ring topology at the distribution level, which efficiently detects and restores the flow of affected traffic through MAC controlled switching. Moreover, it is observed through numerical analysis that the proposed self-healing hybrid protection architecture provides the desirable connection availability with no fiber duplication and hence less deployment cost as compare to the existing solutions.

References

[1] C. Lam, Passive optical networks. Amsterdam: Elsevier/Academic Press, 2007.

[2] E. Wong, 'Survivable architectures for time and wavelength division multiplexed passive optical networks', Optics Communications, vol. 325, pp. 152-159, 2014.

[3] L. WOSINSKA, J. CHEN and C. LARSEN, 'Fiber Access Networks: Reliability Analysis and Swedish Broadband Market', IEICE Trans. Commun., vol. 92, no. 10, pp. 3006-3014, 2009.

[4] B. Lee, 'Simple ring-type passive optical network with two-fiber protection scheme and performance analysis', Optical Engineering, vol. 46, no. 6, p. 065002, 2007.

[5] Fu-Tai An, D. Gutierrez, Kyeong Soo Kim, Jung Woo Lee and L. Kazovsky, 'SUCCESS-HPON: A next-generation optical access architecture for smooth migration from TDM-PON to WDM-PON', IEEE Communications Magazine, vol. 43, no. 11, pp. S40-S47, 2005.

[6] X. Zhao, X. Chen, X. Fu, “A novel protection switching scheme for PONs with ring plus tree topology” Proc. SPIE 6022, Network Architectures, Management, and Applications III, 60223H December 05, 2005, doi:10.1117/12.636269.

[7] C. Yeh and S. Chi, 'Self-Healing Ring-Based Time-Sharing Passive Optical Networks', IEEE Photonics Technology Letters, vol. 19, no. 15, pp. 1139-1141, 2007.

[8] P. Lafata and J. Vodrážka, 'Experimental Verification of Passive Optical Network With Ring Topology', *Microwave and Optical Technology Letters*, vol. 55, no. 9, pp. 2201-2205, 2013.

[9] Y. Qiu and C. Chan, 'A novel survivable architecture for hybrid WDM/TDM passive optical networks', Optics Communications, vol. 312, pp. 52-56, 2014.

[10] C. Yeh, C. Chow and Y. Liu, 'Self-protected ring-star-architecture TDM passive optical network with triple-play management', Optics Communications, vol. 284, no. 13, pp. 3248-3250, 2011.

[11] M. Zhu, W. Zhong and S. Xiao, 'A Survivable Colorless Wavelength Division Multiplexed Passive Optical Network With Centrally Controlled Intelligent Protection Scheme', J. Opt. Commun. Netw., vol. 4, no. 10, p. 741, 2012.

Computer Vision for Screening Resistance Level of Rice Varieties to Brown Planthopper

Elvira Nurfadhilah
Department of Computer Science
Bogor Agricultural University
West Java, Indonesia

Aunu Rauf
Department of Plant Protection
Bogor Agricultural University
West Java, Indonesia

Yeni Herdiyeni
Department of Computer Science
Bogor Agricultural University
West Java, Indonesia

Rahmini
Plant Protection
Indonesian Center for *Rice Research*
West Java, Indonesia

Abstract—**Brown planthopper is one of the most important insect pest that threatens the stability of national rice production in Indonesia. One of the efforts to save rice production is by using brown planthopper resistant variety. Currently the determination approach is still conventional based on Standard Seedboxes Screening Test from IRRI with assistance of experienced experts in the scoring process resistance level.In this study, a prototype of application system to predict resistance levels by image color approach was developed. The method consists of collecting images data, preparation process (background and objects segmentation), and determination of area proportion which has been infected (sick and dead) and healthy, based on 'A' value from CIELab color space laboratory. According to proportion value distribution, the rule of rice resistance to brown planthopper assessment based on image was developed. The rule is mostly similar with IRRI standard rules. All of images were assessed based on the rule and then the model was developed with an error rate of 17.02%.**

Keywords—brown planthopper; color extraction; resistance; standard seedboxes screening test

I. Introduction

Brown planthoppers's latent pestsare difficult to detect, yet their presence had always been a threat to the stability of national rice production.Brown planthopper is a rice-specific herbivore and sucks the phloem sap of rice plants through its stylet mouthpart [1]. Moreover, the brown planthopper attacks may undirectly transfer three lethal viruses for paddy plants, namely the ragged stunt virus,grassy stunt virus type 1, and grassy stunt virus type 2. The symptoms of brown planthoppers attack on individual hills of plants include yellowing leaves, followed by drying plants that look burnt / hopperburn[2].

In the effort to save the rice production, many possibilities of pest control are available, including using pest-resistant varieties, natural enemies, cultivation method (planting timing, irrigation, etc), and insecticides [3]. One of the importants aspects of pest control is using planthopper pest-resistant varieties.

The one of the important tasks in overcoming pests is using pest resistant varieties. Indonesian Center for Rice Research is one the centers under the Ministry of Agriculture which focuses on obtaining superior pest-resistant varieties by testing them against various brown planthopper biotypes. Cultivar screening for planthopper resistancy using greenhouse screening used in IRRI is the Standard Seedboxes Screening Test(SSST).

Currently, resistance level using SSST is done manually by experienced experts in resistance level scoring process. Digital image based system prediction is a new approach in screening and scoring the variety resistance level against BPH.

According to Madhogaria [4], the separation between sick and healthy areas can be done by classifying the RGB value using SVM classification. Several experiments have been done to seperate the leaves areas which have been infected with sickness spot with the healthy leaves area, by segmenting the leaves which have been detected sick using the R component from RGB, A from CIELab, H from HSV and Cr from YCbCr with Otsu threshold. From the research, the best result was obtained from using the A component from CIELab [5]. Another research [6] has been done to measure the infection severity, by using the component A of CIELab color space on paddy hills images, to differentiate the infected plants from the healthy plants, then looking for the interval through diagram box plot. The measurement accuracy obtained in the experiment was 70.83%.

In this experiment, the ratio of healthy area against the infected area on plant images in seedboxes was calculated using the A component in CIELab color space with multi threshold Otsu. Then classification was done on the damage areas (sick + dead areas) using the interval threshold against the total plant area to classify the ratio of sick areas. The results can be classified into 6 categories, score 0 (Highly Resistant), score 1 (Resistant**)**, score 3 (Moderate Resistant), score 5 (Moderate Susceptible), score 7 (Susceptible) and score 9 (Highly Susceptible).

II. Material and Method

A. System Framework

Figure 1 shows the flows of research method. The research method consists of image collection, pre-processing (background and object segmentation), and determination of attack level based on the attacked plant areas.

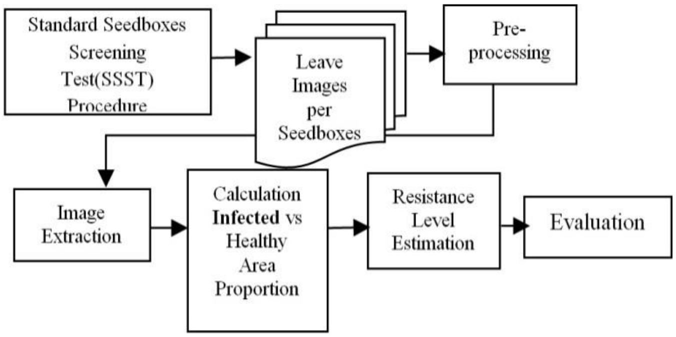

Standard Seedboxes Screening Test(SSST) Procedure
Leave Images per Seedboxes
Pre-processing
Image Extraction
Calculation Infected vs Healthy Area Proportion
Resistance Level Estimation
Evaluation

Fig. 1. The System Framework

B. Standard Seedboxes Screening Test (SSST)

SSST is a method to score the resistance level of each variety against planthoppers by giving several planthopper pairs, then measuring the level of pest growth and its effect on the variety. This method is commonly used to screen the greenhouses in Asia. More than 60,000 entries / year were evaluated in one greenhouse in IRRI. Whereas the procedures to obtain the image data during resistance level scoring using SSST from [6] can be seenin Figure 2 and Table 1.

Table 1 is the standard guidance in manual scoring done in Indonesian Center for Rice Reseach in scoring the paddy plants resistance level against brown planthopper pests, time to scoring when *susceptible check*(TN1) varieties 90% dead.

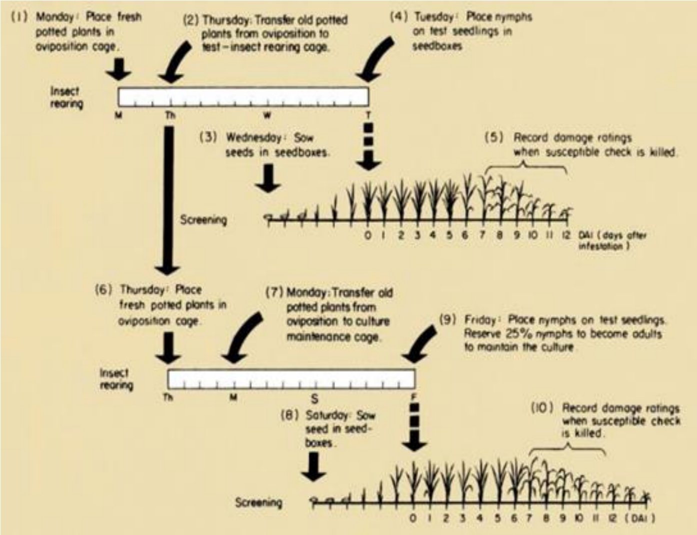

(1) Monday: Place fresh potted plants in oviposition cage.
(2) Thursday: Transfer old potted plants from oviposition to test–insect rearing cage.
(4) Tuesday: Place nymphs on test seedlings in seedboxes
Insect rearing
M
Th
W
T
(3) Wednesday: Sow seeds in seedboxes.
(5) Record damage ratings when susceptible check is killed.
Screening
0 1 2 3 4 5 6 7 8 9 10 11 12 DAI (days after infestation)
(6) Thursday: Place fresh potted plants in oviposition cage.
(7) Monday: Transfer old potted plants from oviposition to culture maintenance cage.
(9) Friday: Place nymphs on test seedlings. Reserve 25% nymphs to become adults to maintain the culture
Insect rearing
Th
M
S
F
(8) Saturday: Sow seed in seed-boxes.
(10) Record damage ratings when susceptible check is killed.
Screening
0 1 2 3 4 5 6 7 8 9 10 11 12 (DAI)

Fig. 2. Sample schedule of sowing seeds and brown planthopper pest investing 7]

TABLE I. Greenhouse Scoring Guidance According to 2014 IRRI Standard[8]

Symptom Score	Symptom	Criteria
0	No injury	Highly Resistant
1	Very slight injury	Resistant
3	First and 2nd leaves of most plants partially yellowing	ModerateResistant
5	Pronounced yellowing and stunting or about 10-25% of the plants wilting or dead and remaining plants severely stuned or dying	Moderate Susceptible
7	More thanhalfofthe plants wilting or dead	Susceptible
9	All plants dead	Highly Susceptible

C. Pre-processing Image

There were some differences in lighting and contrast at the time the picture was taken. Therefore, enhancement was carried out by performing auto brightness to the picture manually.After that, thresholding was performed between object and the background using the Blue component of RGB color space, as in Figure 3, assuming that 80% of B value is the object and the remaining 20% is the background value. Explanation about the threshold is depicted in Blue screen histogram in Figure 3. On the plant, the frequency was much lower in comparison with the background hence was not shown clearly in the graph.

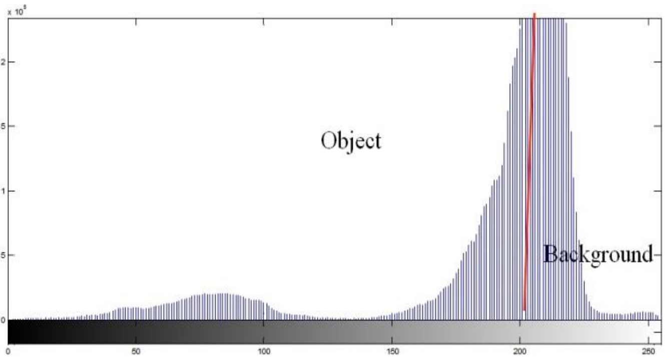


Fig. 3. Sample blue screen histrogram of sample image

D. Image Color Transform

In rice plant, sick/ healthy leaf can be different by color.The color component ‘A’ from CIELab was used to separate the healthy, sick and dead leaves. The color component ‘A’ shows the changing in color from green to red with range of value from 0-255. Healthy plants have more green, whereas sick plants have more yellow to red color components, and dead colors tend to have red to brown colors. Healthy, sick and dead areas may be separated using these color components.

The plant image color space was changed from RGB to CIELab using algorithm [5]. Whereas the formula used was as the following:

$$A=1.4749 \times (0.2213 \times R - 0.3390 \times G + 0.1177 \times B) + 128 \quad (1)$$

E. Multilevel Threshold Otsu

Multilevel Threshold Otsu[9] selects a global threshold value by maximizing the separability of the clusters in ‘A’ levels. Assuming that an image can be represented in L ‘A’ levels (0,1, . . . , L-1). The number of pixels at level i is denoted by fi; then, the total number of pixels equals $N = f_0 + fi + \ldots + fi_{-1}$ .For a given ‘A’ level image, the occurrence probability of ‘A’ level i is given by:

$$p_i = \frac{f_i}{N}, \quad p_i \geq 0, \quad \sum_{i=0}^{L-1} p_i = 1 \qquad (2)$$

If an image is segmented into K clusters (C_0, Ci,...,C_{K-1}) , K-1 thresholds (t_0, t_1, . . . , t_{K-2}) must be selected. The cumulative probability μ_k and mean ‘A’ level for each cluster C_k are respectively given by:

$$w_k = \sum_{i \in C_k} p_i \quad dan \quad \mu_k = \sum_{i \in C_k} i \cdot p_i / w_k \, , k \in \{0, 1, \ldots, K-1\} \qquad (3)$$

Therefore, the mean intensity of the whole image μ_k and the between-class variance σ^2_B are respectively determined by:

$$\mu_T = \sum_{i=0}^{L-1} i \cdot p_i = \sum_{k=0}^{K-1} \mu_k \cdot \omega_k \qquad (4)$$

And

$$\sigma^2{}_B = \sum_{k=0}^{K-1} \omega_k (\mu_k - \mu_T)^2 = \sum_{k=0}^{K-1} \omega_k \mu^2{}_k - \mu^2{}_T \qquad (5)$$

Hence, the optimal thresholds ($t*_0$, $t*_1$, . . . , $t*_{K-2}$) can be determined by maximizing the between-class variance as:

$$\{t_0^*, t_1^*, \ldots, t_{K-2}^*\} = \underset{0 \leq t_0 < \ldots < t_{K-2} < L-1}{\overset{\text{arg}}{}} \max\{\sigma^2{}_B(t_0, t_1, \ldots, t_{K-2})\} \; (6)$$

We used 2 optimal thresholds to separate healthy, sick and dead leave areas.

F. Infected Area Ratio

After segmentation of healthy, sick and dead leave areas, The number of pixels identified as healthy, sick and dead area were calculated against the total plant areas excluding the background, using the following formula:

$$D = \frac{\text{Pi}}{\text{Pi} + \text{Ps}} \qquad (7)$$

D = Damage leaves ratio on seedboxes image

Pi = The number of infected leave parts on seedboxes image (in pixels)

Ps = The number of healthy leave parts on seedboxes image (in pixels)

Where Damage Area (Pi) = Sick Area + Dead Area

G. Resistance Level Estimation

Determination of superior varieties resistance level using the proportions of healthy and damage (sick or dead) leave areas and ratio classifications using threshold interval for damage area proportions may be classified into 5 categories, namely score 1 (Resistant), score 3 (Moderate Resistant), score 5 (Moderate Susceptible), score 7 (Susceptible) and score 9 (Highly Susceptible). The severity level may be determined using interval method based on value distribution of infected area proportion [5].

III. Experimental Results

A. Data Colected

Data used were obtained from direct observation when the susceptible check variety (TN1) were dead almost 90 %. The data were captured using Macro Digital Camera Canon EOS 550D. The images were captured from seedboxes with white paper background. There were 10 tested varieties and repeated 6 times. 1 control and 5 which were investigated were planthopper.

Based on scoring results, score distributions were not even.

Score 0 → 100 images
Score 1 → 20 images
Score 3 → 90 images
Score 5 → 230 images
Score 7 → 100 images
Score 9 → 30 images
Score 1 or 3 → 20 images
Score 7 or 9 → 10 images
Total: 600 images

Since there were difference of scoring for 30 images, only 570 images were used.

B. Pre-processed Image

Prior toimage processing, pre-processing must be done to obtain optimum results. Images with white background have higher blue value than plant images, hence may be used to separate background from object (plants). Figure 4 may be used to view the process more clearly.

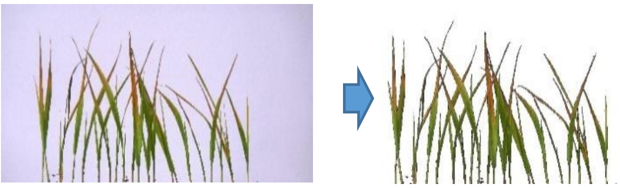


Fig. 4. Background separation from plant objects

Figure 4 shows an illustration of image segmentation with threshold value of 70-90% from Blue value distribution. The threshold value depends on image quality (contrast/ brightness etc.).

C. Image Extraction

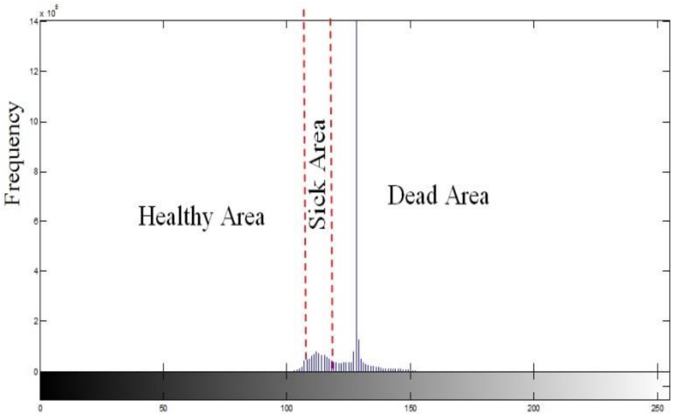


Fig. 5. Sample histogram of A screen from CIELab

Classification of healthy and sick (yellow to hopperburn) areas was then performed on the processed image. An illustration of image extraction can be seen in Figure 5 and 6.

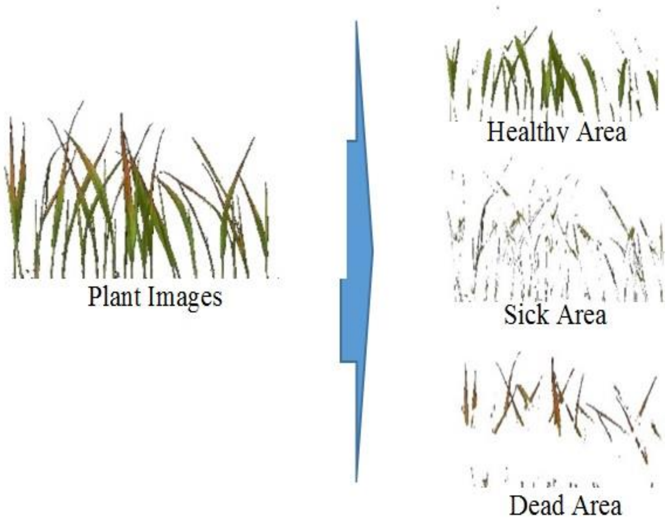


Fig. 6. Separation between healthy, sick and dead area

In this Multilevel Otsu, 16 clusters were used with interval between clusters 16.

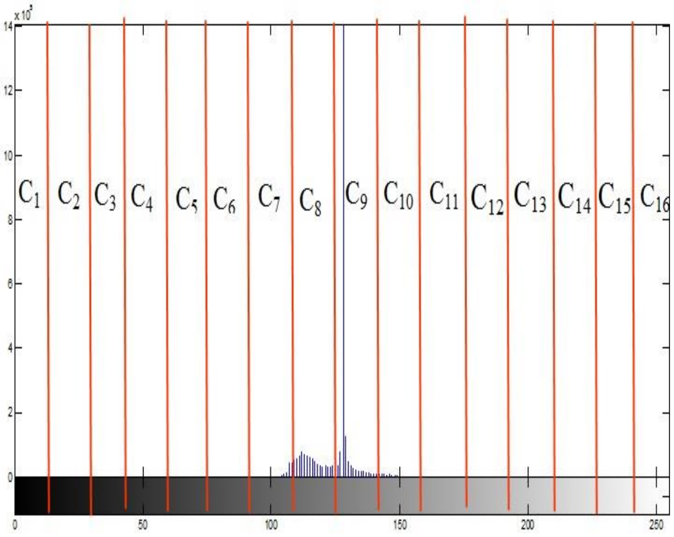


Fig. 7. Sample screen histogram of A divided into16 clusters

Each cluster's σ^2_B was then calculated. Then 2 clusters with maximum values were chosen from the 16 clusters. For those 2 clusters with maximum σ^2_B, the threshold value which satisfy the maximum σ^2_Bvalues between clusters was found. The illustration this process may be seen in Figure 7.

Each image has different threshold, depending on lighting, contrast and A-value distribution. The mean threshold was found in Cluster 8 (112-127) and Cluster 9 (128-143). Whereas the frequency distribution may be seen in Table 2 and Figure 8.

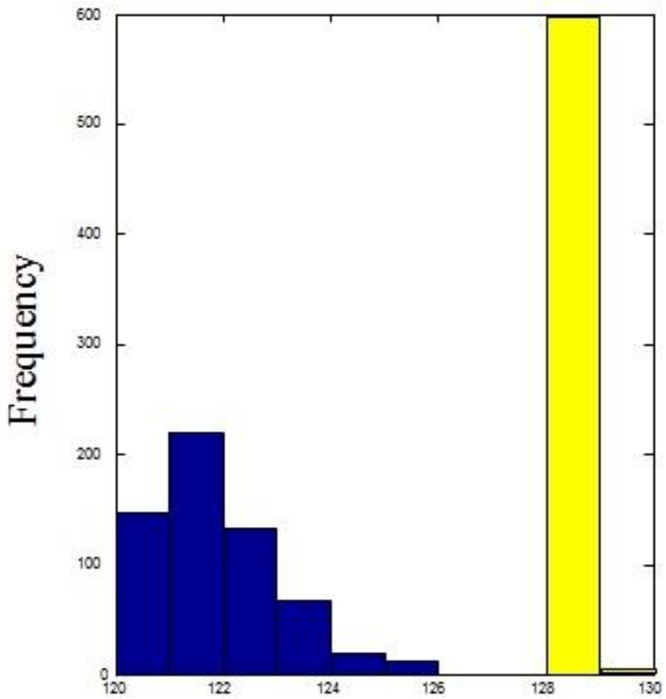


Fig. 8. Frequency Distribution of Threshold Histogram Treshold

D. Resistance Level Classification

Number of 570 images were taken from 57 seedboxes varieties which have been infected by brown planthopper pests in this experiment, using seedboxes modification. Three seedboxes was not used, which had different values among experts. From the 570 images used, experts measured that 100 images were classified as score 0 (resistant). Furthermore, 20 images were classified as score 1 (resistant), 90 images as score 3 (Moderate resistant), 230 images as score 5 (Moderate susceptible), 100 images as score 7 (susceptible) and 30 images as score 9 (highly susceptible).

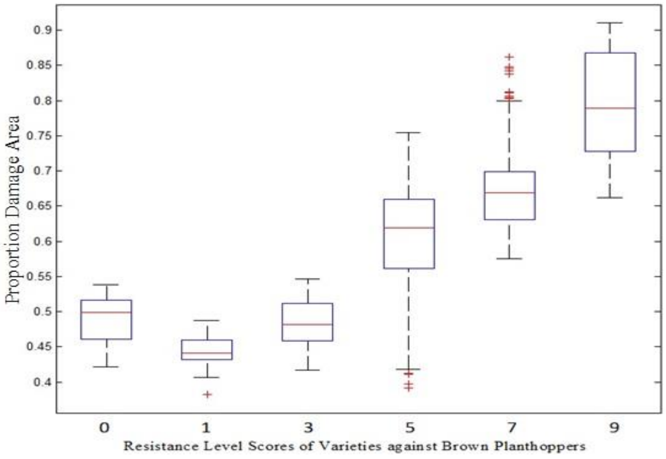


Fig. 9. Boxplotof six resistance levels

Boxplot on Figure 9 which illustrates data distribution of six resistance level categories shows the distribution of each category. Boxplot for the six classes particularly for score 3 and 5 show quite high classification error. This is due to the overlapping between mean ratio from score 3 and 5. Based on the ratio value distribution of sick area, the resistance scoring rules of paddy against brown planthopper is allowed. The rule is similar to IRRI standard rules, only the number of details is a little different because manual calculation is done per plant in the seedboxes whereas computation uses area approach.

TABLE II. RESISTANCE SCORE RULES BASED ON RATIO DAMAGE AREA

Resistance Score	Ratio Damage Area (D) %
0	D ≤ 45
1	45 < D ≤ 49
3	49 < D ≤ 55
5	55 < D ≤ 65
7	65 < D ≤ 80
9	D > 80

E. Classification Result

All images was scored using the ratio interval damage area in Table 2 and the error rate was calculated as 17.02%. Based on matrix confusion table, it may be seen that classification error happen mainly on the neighborhood classes.The illustration this process may be seen in Table 3 and 4.

TABLE III. CONFUSION MATRIX FOR IMAGE BASED RESISTANCE SCORING

		Predicted Score					
		0	1	3	5	7	9
Actual Score	0	15	28	57			
	1	13	7				
	3	18	32	40			
	5	17	16	22	103	72	
	7				57	32	11
	9					17	13

TABLE IV. ERROR RATE CALCULATION, ERROR PROPORTION BASED ON ERROR RATE AND THE FREQUENCY

Error Rate (ER)	Frequency (F)	ER x F
0	210	0
1/5	252	50.4
2/5	91	36.4
3/5	17	10.2
4/5	0	0
5/5	0	0
Total	570	97

Proportion Error (%) = (∑ER x F) / ∑ F = 97/ 570 = 17.02 %

In this experiment, the error proportion is not too high. Classification error happened on class with close resistance level, for instance class 0 (Highly Resistant), 1 (Resistant) and 3 (Moderate Resistant) also for class 5 (Moderate Susceptible), 7(Susceptible), and 9 (Highly Susceptible).

The following shows image sample for segmentation of healthy, sick and dead areas for each class.

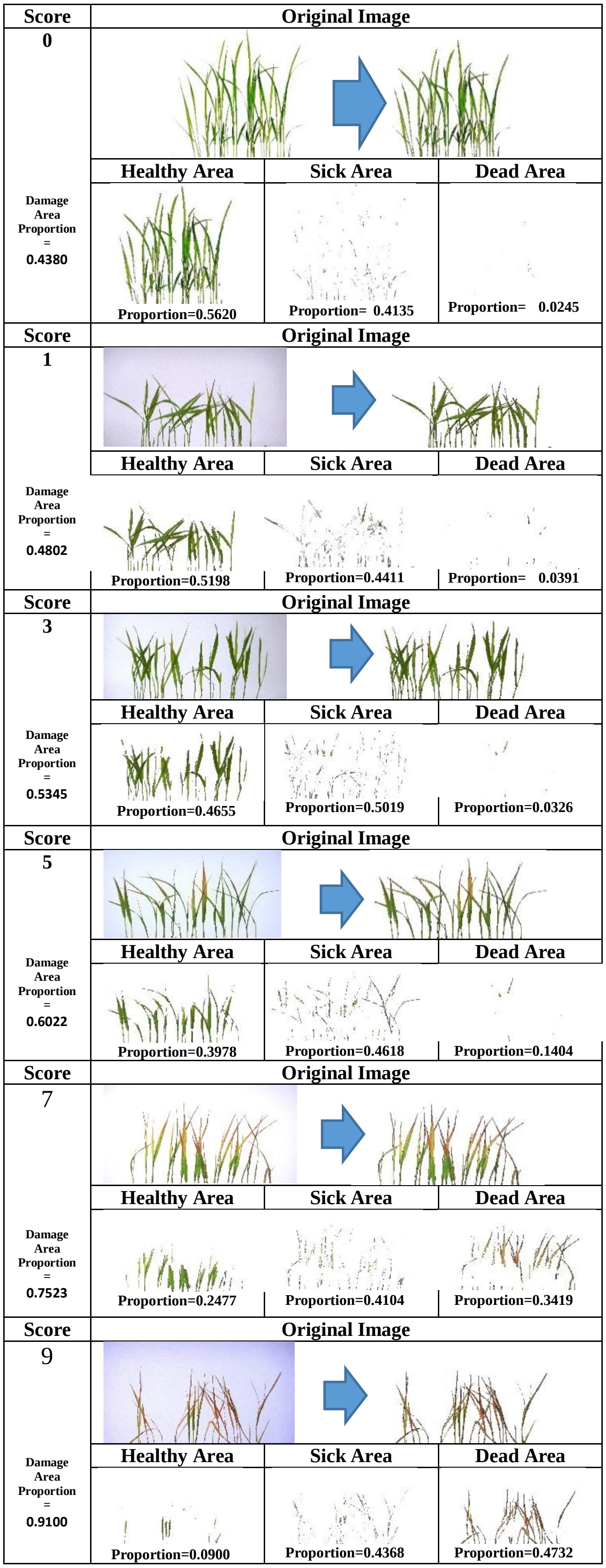


Fig. 10. Sample of Correct Resistance Level Classification

Figure 10 shows that the higher the score, the larger the infected areas (sick and dead leaves areas). Common mistake usually occur on the neighboring classes. The following is the sample of classification error to the neighboring classes. Figure 11 shows a class 3 was incorrectly classified into class

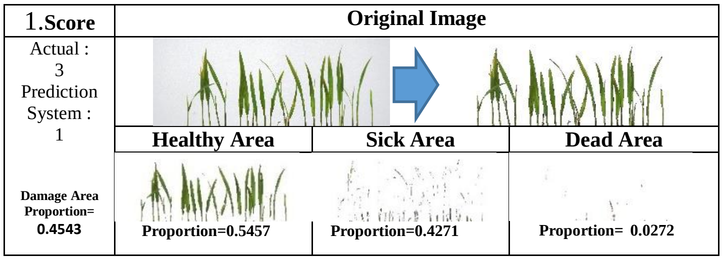


Fig. 11. Sample of incorrect resistance level classification

Quite fatal misclassification in Figure 12 occurred during identification process when the resistance score 5 was recognized as 0.

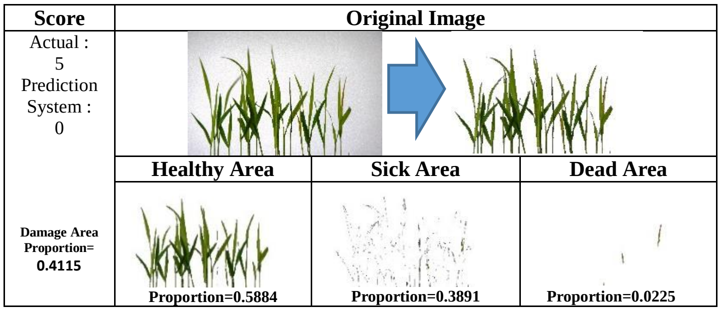


Fig. 12. Sample of incorrect resistance level classification

By only looking for the color feature, it tended to be recognized as score 1 or 3 because the color of the leaf was uniformly green, but the expert gave score 5. This was because the expert saw some spun leafs. Spun leaf image cannot be detected by the color feature because there were some green colored spun leaves. Classification error occurs because of overlapping in the classification limits of resistance level between adjacent classes. This may also be caused by different lighting level and contrast during image capturing or incomplete pre-processing. Classification error may also happen to manual scoring because quantitatively clear limitation is not yet available to differentiate between resistance level scores.

IV. Conclusion and Future Direction

In this paper assessing resistance Level of rice varieties using digital image processing are studied. We applied Multilevel Otsu to classify the resistance level by the damaged area ratio. Experimental result shows that all of images were assessed based on the rule and then the model was developed with an error rate of 17.02%. This result show that our proposed method is promising to measure resistance level of rice varieties automatically. Further to this, conditioning such as room lighting is necessary to obtain relatively uniform results of picture capturing to minimize segmentation error. Additionally, it also applies to other features such as shape, height, etc.

Acknowledgment

The authors would like to express their gratitude to the Directorate General of Higher Educationfor providing Bantuan Operasional Perguruan Tinggi Negeri (BOPTN) to fund this research. We also thankful to the Indonesian Center for Rice Research for giving permission to the authors to perform the research in the Field Biotype Brown Planthopper greenhouse.

References

[1] Du B, Zhang W, Liu BF, Hu J, Wei Z, Shi ZY, He RF, Zhu LL, Chen RZ, Han B, He GC.Identification and characterization of Bph14, a gene conferring resistance to brown planthopper in rice. Proc Natl Acad Sci USA.2009;106:22163–8.

[2] Sumiati, Ani. 2011. Pengendalian hama wereng batang coklat pada tanaman padi. Jambi (ID): Balai Pengkajian Teknologi Pertanian.

[3] Soemawinata, A. T. dan Soemartono S. 1986. Hama wereng cokelat dan masalah pengendalianya di Indonesia. Di dalam:*Prosiding Diskusi Ilmiah Wereng Cokelat dan Pengendalianya*. Bogor (ID): Fakultas Pertanian IPB.

[4] Madhogaria S, Schikora M, Koch W, and Cremers D.2011. Pixel-based classification method for detecting unhealthy regions in leaf images.In: *6th IEEE ISIF Workshop on Sensor Data Fusion: Trends, Solutions, Applications (SDF).*

[5] Chaudhary P, Chaudhari AK, Cheeran AN, Godara S. 2012. Color transform based approach for disease spot detection on plant leaf. *International Journal of Computer Science and Telecommunication.* 3(6):65-70.

[6] Asfarian A, Herdiyeni Y, Rauf A, Mutaqin KH. Paddy Diseases Identification with Texture Analysis using Fractal Descriptors Based on Fourier Spectrum.In : *International Conference on Computer, Control, Informatics and its Applications* (IC3INA) 2013. Jakarta Indonesia

[7] Heinrichs EA, Medrano FG, Rapusas HR.1985. Genetic evaluation for insect resistance in rice. Los Banos (PH): IRRI.

[8] IRRI. 2012. Standard Evaluation System for Rice. Los Banos (PH): IRRI.

[9] Huang DY, Ta Wei Lin, and Wu Chih Hu. 2011. Automatic Multilevel Thresholding Based On Two-Stage Otsu's Method With Cluster Determination By Valley Estimation. In : *International Journal of Innovative Computing, Information and Control* Volume 7, Number 10, October 2011.

Authors Profile

Elvira Nurfadhilah received her undergraduate degree from Bogor Agricultural University in 2011. She currently works in the Agency for the Assessment and Application of Technology as an Engineering Staff in the Intelligent Computing Lab. She is currently pursuing her Master Degree in Computer Science from Bogor Agricultural University with a research topic on Image Processing.

Yeni Herdiyeni learned a PhD in Computer Science with the Dissertation on Semantic Image Similarity using Tree from University of Indonesia (2010). Then she conducted Post Doctoral research at Department of Information Science, Graduate School of Science and Engineering, Saga University, Japan, for 5 months (September - January 2012).She had a Master of Computer Science from University of Indonesia with the thesis on 3D Face Recognition (2005). She obtained her first degree in Computer Science from Bogor Agricultural University (IPB), Indonesia (1999). Currently she is conducting research on digital image processing, computer vision and computational intelligence and biodiversity Informatics.

Aunu Rauf is a professor of agricultural entomology at the Bogor Agricultural University-Indonesia. He earned his M.Sc and PhD degrees from University of Wisconsin-Madison, USA in 1980 and 1983, respectively.

Rahmini received PhD from IPB (Bogor Agricultural University) in 2012 with subject Entomology. She works for Plant Protection Division in Indonesian Center for Rice Research (ICRR) under Indonesian Agency for Agricultural Research and Development (IAARD), Ministry of Agriculture, since 1995.

Probabilistic Algorithm based on Fuzzy Clustering for Indoor Location in Fingerprinting Positioning Method

Bo Dong
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

Jian Xing
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

Fei Wu*
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

Yan Zou
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

Abstract—**Recently, the location of the fingerprint positioning technology is obviously superior to the signal transmission loss model based on the positioning technology, and is widely concerned by scholars. In the online phase, due to the efficiency of the probabilistic distribution matching computation is low and when clustering the position fingerprint database, hard clustering lead to degrading the positioning accuracy, a probabilistic algorithm based on fuzzy clustering is proposed and applied to the indoor location fingerprinting positioning. Compared with hard clustering fusion algorithm, the proposed method has realized the fuzzy partition of the database, makes online positioning phase can effectively search the desired fingerprint data, and improve the positioning accuracy. Experiments show that the algorithm can effectively deal with the problem of the positioning accuracy of hard clustering.**

Keywords—Fuzzy Clustering; Fingerprinting Positioning; Indoor Location; RSSI; Probabilistic Algorithm

I. Introduction

With the development of wireless network technology, the demand for location-based services is increasing [1]. In the outdoor environment, GPS can satisfy people's demand for location service, but in the indoor environment, the satellite signal is shielded, so the location accuracy is not high. In daily life, handheld device, mobile terminal, domestic robot and other indoor intelligent equipment require the indoor location, and this demand has been spurred research on indoor positioning technology. Thanks to the development of WIFI technology, WIFI positioning technology has been widely concerned because of its low cost, easy to implement, and small influence [2].

Indoor positioning algorithm consists of Arrival of Angle (AOA) positioning method and Time of Arrival (TOA) positioning method, Time Difference of Arrival (TDOA) positioning method and positioning method based on signal strength (RSSI) [3]. With the location technology based on AOA, TOA, and TDOA, the positioning result is reliable and relatively accurate only in the visibility (Line of sight) signal under the condition of dominant [4-6]. In addition, TOA and TDOA positioning method require very accurate clock synchronization between the receiver and transmitter, and the positioning cost is high. In contrast, localization method based on signal strength has the advantages of low cost, simple implementation method, and so on [7]. With a higher precision, the location fingerprint positioning method based on the signal strength becomes the main research direction in the locating method based on signal strength [8].The main process of the location fingerprinting positioning method can be divided into two phases: the offline phase and the online phase [9].

In the offline phase, the signal intensity of the sample is collected at the reference point, and the location fingerprint database is constructed by using the collected data. In the online phase, the signal is acquired in real time, and the algorithm matches with the fingerprint data constructed in the offline stage to get the target position of the coordinate information.

In the actual situation, due to the efficiency of the probabilistic distribution matching computation is low, and when clustering the position fingerprint database, hard clustering lead to degrading the positioning accuracy, therefore, a probabilistic algorithm based on fuzzy clustering is proposed and applied to the indoor location fingerprinting positioning. Compared with hard clustering fusion algorithm, the proposed method has realized the fuzzy partition of the database, makes online positioning stage can search to the fingerprint data effectively, and improve the precision of position determination.

II. Introduction of Position Fingerprint Positioning Method

The location of the fingerprinting position method based on the signal strength is divided into two phases: offline and online phases [9].

The mission of the offline phase includes the following aspects:

1) Complete the deployment of Access Point (AP) and the determination of reference point.

2) Ensure that the wireless signal emitted by the AP can reach each reference point position.

3) Signal receiving device placed on the reference point location collect the wireless signal from the AP.

4) The server extracts feature parameters (signal strength, variance, etc.), then along with the location information, store it in the position fingerprint database (Radio Map) as a fingerprint.

In the online positioning phase, the mobile terminal collects the AP signal strength information, and the fingerprint matching algorithm is used to find the best match with the fingerprint data constructed in the offline phase, and finally obtains user location information of the mobile terminal. Its principle is shown in figure 1.

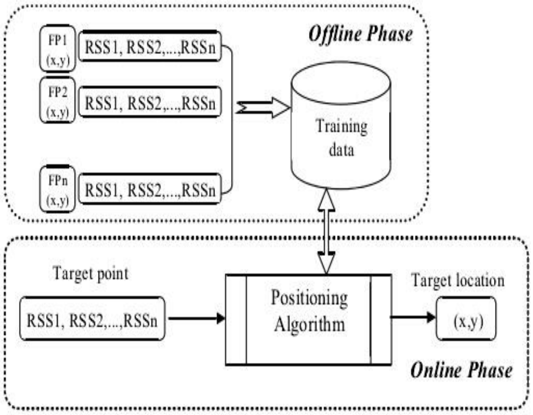


Fig. 1. The principle diagram of the fingerprint location method

III. Theoretical Analysis of the Algorithm

The probabilistic algorithm based on fuzzy clustering is presented in this paper [10]. Firstly, the fuzzy clustering fusion algorithm is used to get the coarse positioning. Then the algorithm selects the appropriate class family based on signal strength acquisition, and uses probabilistic method to obtain the coordinates of the target point, in order to achieve precise positioning location. The basic principle of the algorithm is shown in Figure 2.

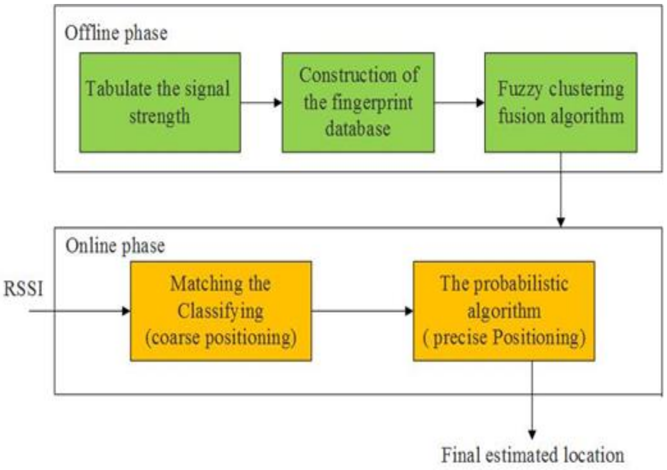

Offline phase
Tabulate the signal strength
Construction of the fingerprint database
Fuzzy clustering fusion algorithm
Online phase
RSSI
Matching the Classifying (coarse positioning)
The probabilistic algorithm (precise Positioning)
Final estimated location

Fig. 2. Basic flow chart of the algorithm

A. Fuzzy clustering fusion algorithm

Fuzzy clustering fusion algorithm is a relatively perfect algorithm in the theory development with the clustering algorithm based on the objective function. Fuzzy clustering algorithm can be used to divide n data $\{x_1, x_2, x_3, \cdots x_n\}$ into C data groups $X_1, X_2, X_3, \cdots X_c (X_1 \cup X_2 \cup X_3 \cup \cdots X_c = X)$, then, get the clustering centers of each group $p_i = (p_{i1}, \mathrm{p}_{i2}, \cdots, \mathrm{p}_{is}) \in R^c$ to make the minimum sum of squared errors between the samples in each class and the clustering center. Its formula is as follows:

$$J_{FCM}^{m}(U, \mathbf{P}) = \sum_{i=1}^{C} \sum_{j=1}^{n} \mu_{ij}^{m} d_{ij}^{2} \tag{1}$$

There, $\mu_{ij} \in [0,1]$indicates the degree of x_j is X_i, namely the membership, and must meet the conditions: $\sum_{i=1}^{c} \mu_{ij} = 1$, $\forall j = 1, 2, \cdots, \mathrm{n}$; $U = [\mu_{ij}]_{c \times n}$ is the fuzzy classified matrix; *m* is weighted index, if *m* is too large, the clustering effect will be very poor, if *m* is too small, the algorithm will be close to c-means clustering algorithm, so *m* usually assigned to 2 [11]; d_{ij} indicates the distance between the sample x_j and the class center (p_i), also, it can be written as the following form:

$$d_{ij}^2 = \| x_j - p_i \|_A = (x_j - p_i)^T A(\mathrm{x}_j - p_i) \tag{2}$$

There, when A is $I_{s \times s}$, d_{ij} is the Euclidean distance.

The minimum of $J_m(U, \mathrm{P})$ is the optimal solution of fuzzy clustering objective function:

$$\begin{aligned} \min\{J_m(U,\mathrm{P})\} &= \min\left\{\sum_{i=1}^{C}\sum_{j=1}^{n}\mu_{ij}^m d_{ij}^2\right\} \\ &= \sum_{j=1}^{n}\min\left\{\sum_{i=1}^{C}\mu_{ij}^m d_{ij}^2\right\} \end{aligned} \tag{3}$$

As the extreme constraint is $\sum_{i=1}^{n} \mu_{ij} = 1$, the Lagrange multiplier method is used to solve:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{m-1}}} \tag{4}$$

Make $J_m(U, \mathrm{P})$ is the smallest value, the value of μ_{ij} is:

$$\begin{cases} \mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{m-1}}} & I_j = \varnothing \\ \mu_{ij} = 0, \forall i \in \bar{I}_j, \sum_{i=1}^{n} \mu_{ij} = 1 & I_j \neq \varnothing \end{cases} \tag{5}$$

Also, $\mathbf{p}_\mathrm{i}$ can be obtained in a similar way:

$$p_i = \frac{\sum_{j=1}^{n} (\mu_{ij})^m x_j}{\sum_{j=1}^{n} (\mu_{ij})^m} \tag{6}$$

If the data set X, the number of cluster categories *m* and the weights *c* are known, optimal fuzzy classification matrix and cluster center can be obtained by the formula.

The pseudo-code of the algorithm is as follows:

Initialization: the number of data samples is n, and the class number is c $(2 \le c \le n)$,the fuzzy weighting exponent m is set to 2, the initial clustering center value $p^{(0)}$,ε indicates the iterative stopping threshold, iteration number $l = 0$;

TABLE I. ALGORITHM PSEUDOCODE

```
Repeat for $l$ = 1 , 2……
Step 1：compute the cluster prototypes(means)：
    $p_i^{(l)} = \frac{\sum_{j=1}^{n} (\mu_{ij}^{(l-1)})^m x_j}{\sum_{j=1}^{n} (\mu_{ij}^{(l-1)})^m}, 1 \le i \le c$
Step 2：compete the distance：
    $d_{ij}^2 = \| x_j - p_i^{(l)} \|_A$
    $= (x_j - p_i^{(l)})^T A(x_j - p_i^{(l)}),$
    $1 \le i \le c, 1 \le j \le n$
Step 3：Update the partition matrix：
For $1 \le j \le n$
If $(d_{ij})^2 > 0$ for all $i$=1，2，…，c
    $u_{ij}^{(l)} = \frac{1}{\sum_{k=1}^{c} (d_{ij}^{(l)} / d_{kj}^{(l)})^{2/(m-1)}}$
Otherwise
    $u_{ik}^{(l)}$ =0 if $d_{ik_A}$ >0， and $u_{ik}^{(l)} \in [0，1]$ with $\sum_{i=1}^{c} u_{ik}^{(l)} = 1$
Until $\| U^{(l)} - U^{(l-1)} \| < \varepsilon$
```

B. Probabilistic algorithm (PM)

The probabilistic distribution location algorithm is used to calculate the matching probabilistic of each point in the wireless signal intensity and location fingerprint, and the maximum of the probabilistic of the reference point is the estimated position of the target [12]. By introducing the probabilistic function of Gauss signal, the signal intensity distribution at any position in the indoor positioning environment is characterized by the expectation and variance of the signal. For the actual collection of wireless signal intensity, it is usually considered to be the general distribution of normal distribution $N(\mu, \sigma^2)$, the statistical parameters are μ and σ^2 [12]. The likelihood function is:

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{7}$$

Calculate:

$$\begin{cases} \mu^* = \bar{X} \\ \sigma^{*2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \end{cases} \tag{8}$$

According to the maximum likelihood estimation, μ^* and σ^{*2} are considered as the real value, and the signal intensity of the reference point is considered as the reference point. In the online phase, according to the signal intensity of real-time acquisition, the location of the indoor location fingerprint database is matched in order to achieve precise positioning. By using normal distribution probabilistic formula, the probabilistic of the point (x, y) is obtained:

$$P_i(x, y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(RSSI_i-\mu)^2}{2\sigma^2}} \quad (\sigma > 0) \tag{9}$$

After obtained the wireless signal strength of all APs in the database of probabilistic distribution, take the collected wireless signal strength value into the substitution probabilistic formula, and calculate the probabilistic of each reference point $P(x, y)$ in the class family, as shown in formula (10). There, *M* is the number of samples for the current class. The maximum probabilistic product is obtained and the coordinate (x, y) is used as the current position estimate coordinates.

$$P(x, y) = \prod_{i=1}^{M} p_i(x, y) = \prod_{i=1}^{M} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(RSSI_i-\mu)^2}{2\sigma^2}} \right) \tag{10}$$

IV. DESIGN OF EXPERIMENT AND RESULT ANALYSIS

In order to verify the effect of this algorithm, the experiment is carried out on the fourth floor of Building 1, local plan is shown in Figure 2. The location area includes the corridor area and the laboratory area. Among them, the laboratory is 6 × 9m, the corridor is 2 × 27m, the location fingerprint interval is 1.5 m; the wall thickness is 33cm, and a total of 138 reference points. The smart router with 5 models named JHR-N845R is used as AP, and 3C Honor is used to collect the RSSI signal of AP, and the sampling frequency is 30samples/min.

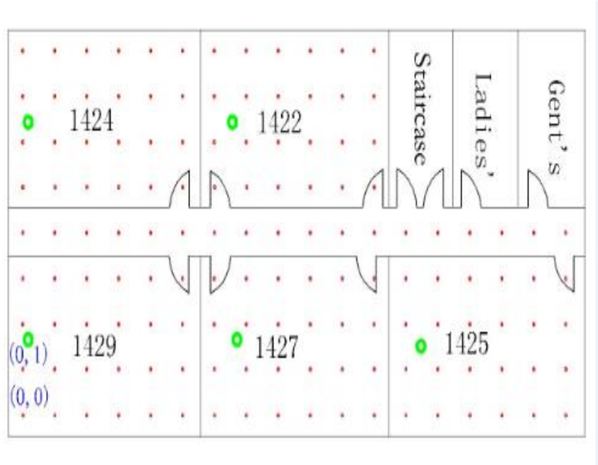


Fig. 3. Experimental environment plan

Figure 3 in the lower left corner is the origin, the vertical direction as y axis and horizontal direction for the X axis to establish coordinate system, and green rings represent the AP location, red dots represent the fingerprint minutiae. In the offline phase, the RSSI values of the reference points are sampled at 50 times, and <the location, the AP name, the RSSI mean, the variance > is recorded in the fingerprint database. The 138 samples are clustered into 6 classes, and the fuzzy weighted index *m* is set to be 2. In the online phase, N times sampling is taken at the target point, and respective calculation are performed in the test position.

In order to evaluate the effectiveness of the algorithm in this paper, the results in the online phase are evaluated, the evaluation parameters are:

$$e = \sqrt{(x_m - x)^2 + (y_m - y)^2} \tag{10}$$

$$e\_aver = \frac{1}{n}\sum_{j=1}^{n} e_j \tag{11}$$

There, (x, y) is actual coordinate, (x_m, y_m) is the calculated coordinates, *e* is the positioning error, e_aver is the multiple positioning error of the mean, *n* is the number of measurement at the current position. The sampling is taken *n* times (*n*=1, 2, 5, 10, 15, 20) at the point (8, 6), the formula (10) and (11) are used to deal with the data obtained, the average error is shown in Figure 4.

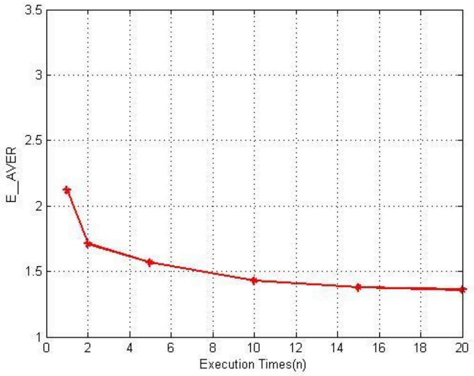


Fig. 4. The average error of N times sampling results

From Figure 4 we can see that in the online positioning phase, when *n*=1, the calculation error of the positioning is large, so the result is not ideal, and the positioning error is reduced after multiple sampling. In the actual situation, we can choose the mean value of the 10 sampling results as the final result.

Again, compared the algorithm proposed in this paper with hard clustering algorithm, the experiment select the hierarchical clustering algorithm (HCA) [13] and K-means clustering (k=6) [14], error results as shown in Figure 5.

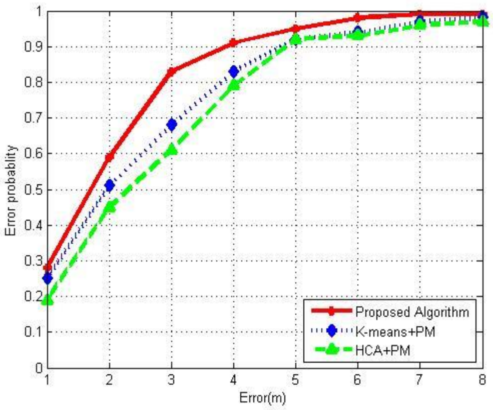


Fig. 5. The cumulative error distribution map

As is shown in figure 5, the results of the experiment show that compared with classic hard clustering algorithm, the proposed algorithm realizes the fuzzy partition of the database, so that the online positioning phase can effectively search to the desired fingerprint data. The probability of the positioning accuracy within 2 meters has reached 60%, and the probability of the positioning accuracy within 3 meters has reached 80%. Compared with the hard clustering algorithm, the positioning accuracy is improved.

V. Conclusion

In this paper, a probabilistic algorithm based on fuzzy clustering is proposed, which is used in the indoor location. This algorithm is an effective solution to the unsatisfactory results of the positioning accuracy with the hard clustering algorithm. Experiments show that the algorithm can improve the positioning accuracy by using the method of multiple sampling in the online positioning phase. And, compared with hard clustering fusion algorithm, the probability of the positioning accuracy within 2 meters has reached 60%, and the probability of the positioning accuracy within 3 meters has reached 80%. The proposed algorithm can satisfy the practical application of indoor positioning accuracy.

VI. Future Work

Aiming at the accuracy of the established database, for the future work, we plan to establish the database by the idea of feedback. Also, a better position of the AP is considered to improve the accuracy of the indoor positioning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.61272097, 61405045), the Shanghai Municipal Committee of Science and Technology Project (No.13510501400), the Shanghai Municipal Education Commission, A Key Innovation Project (No.12ZZ182).

References

[1] Zheng Yang, Chenshu Wu, Yunhao Liu.Location-based Computing: Localization and Localizability of Wireless Networks, Beijing:Tsinghua university press, 2014, pp.111-127.

[2] Dianjun Wang, Hongxing Wei, Fujun Ren. Autonomous mobile robot positioning technology[M]. Beijing:China Machine Press, 62-76, 2013.

[3] Zhang Wenjie, Dong Yuning, Wang Xinheng.Indoor positioning method using RFID and block clustering[J/OL]. Computer Engineering and Applications, 2015-02-16.

[4] Zhou J, Zhang H, Mo L. Two-dimension localization of passive RFID tags using AOA estimation[C]. Instrumentation and Measurement Technology Conference (I2MTC), 2011 IEEE. IEEE, 2011:1-5.

[5] Frank K, Julian L, Christ R. WLAN Mobile Robot Localization with Sensor Fusion[J]. IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications. Sep.2009, pp.649-654.

[6] ChengchengJi, Huamei Xin. Indoor fingerprinting localization algorithm based on weighted fuzzy C-means clustering algorithm[J]. Modern Electronics Technique, Vol.36, 2013, 45-47.

[7] Yong Zhang, Jie Huang, Keyu Xu. Indoor positioning algorithm for WLAN based on principal component analysis and least square support vector regression[J]. Chinese Journal of Science Instrument, 2015, 2(36):408-414.

[8] Xiaoxiao Xu, LinboXie, Peng Li. Design of indoor positioning system based on WiFisignal intensity characteristic[J]. Computer Engineering, 2015, 4(41):87-91.

[9] Shengbin Li. Application research on indoor location technology based on WiFi[D]. Nanjing Normal University. 2014.

[10] Li Tang. Indoor WLAN location based on fuzzy cluster KNN algorithm[C]. Harbin Institute of Technology. 2009.

[11] Yilin Du. A fuzzy clustering KNN location fingerprint positioning algorithm [J]. Micro machine and Application, 2012, 31(23):55-58.

[12] Dianjun Wang. Indoor mobile-robot localization system based on WRV[J]. Journal of Tsinghua university. 2011, 12:1849-1854.

[13] Tandong Bai. Classification and Clustering Theory in Data MiningBased Optimization on Fingerprint Indoor Localization[D]. Beijing Jiaotong University. 2014.

[14] Wang Chen, ZhenhongJia, Xizhong Qin, Chuanling Cao. Indoor positioning performance of WLAN based improved K-means algorithm [J]. LASERNAL, 2014, 35(7):11-14.

Implementation of Location Base Service on Tourism Places in West Nusa Tenggara by using Smartphone

Karya Gunawan
Academy of Information Management and Computer
AMIKOM Mataram
Mataram, Indonesia

Bambang Eka Purnama
School of Information Management and Computer
“Nusa Mandiri”, STMIK Nusa Mandiri
Jakarta, Indonesia

***Abstract*—The study is aimed to create an application that can assist users in finding information about tourism places in West Nusa Tenggara, Indonesia. West Nusa Tenggara is one of the provinces in Indonesia and one of the second tourists' destinations after Bali. It is a small Sunda Island which consists of two large islands located in the west of Lombok and Sumbawa that located in the east to the capital of Mataram on the island of Lombok. The area of West Nusa Tenggara province is 19,708.79 km^2. The application provides information such as descriptions of sights, tourism spot address, photo galleries, and the available facilities and the closest track to where the tourism spot is by using Google maps. Google maps can display map locations and closest routes from the user's position to the tourism place. Determination of the position of the user and position of Tourism sites using GPS (Global Positioning System). This application is made by using the Java programming language for Android with Eclipse 4.2.1 IDE that can be used on the Smartphone device based on Android. The result of the research is the creation of location-based services applications to search a place of Tourism in West Nusa Tenggara province so as to help tourists visiting the West Nusa Tenggara. The benefit of this application is that to determine the path in finding the tourism places.**

Keywords—location base service; GPS; Smartphone; tourism information

I. Introduction

West Nusa Tenggara Province is one of the provinces in Indonesia. It comprises the western portion of the Lesser Sunda Islands. West Nusa Tenggara Province consists of two large islands situated in the west of Lombok and Sumbawa are located in the east. Mataram city on the island of Lombok is the provincial capital and largest city in the province. The area of West Nusa Tenggara province is 19,708.79 km2[1]. West Nusa Tenggara has a Strategic Area Regional Tourism (Indonesian: KSPD) both located on the island of Lombok and Sumbawa islands. KSPD is the area that has the primary function of tourism or have the potential for the development of national tourism that have significant influence in one or more aspects, such as economic growth, social and cultural rights, the empowerment of natural resources, environmental support, as well as defense and security[2]. Lombok as the second tourist destination after Bali has many uniqueness, especially in tourism. In an effort to increase local tourism, the government of West Nusa Tenggara through the Department of Tourism to develop two programs: Visit Lombok-Sumbawa 2015 and Tambora Greet the World. Both of these programs are used as tourist information media of West Nusa Tenggara.

Information is a major requirement for most people. By using mobile devices, information can be obtained wherever the location is within a short time. Mobile devices are the most widely used today is the Smartphone with Android operating system. Besides, it is used to search for information, a Smartphone device can also be used to display the map as well as navigation[3].

The tourists who visit West Nusa Tenggara used Smartphone to search for information of tourist sites. They also used GPS (Global Positioning System) that exist in Smartphone to find the track to the tourist sites[4]. The information of tourism site has not been listed on Google Maps. It caused the difficulty in obtaining information and the correct location. This resulted in the lack of interest of tourists to visit tourist sites in Lombok and Sumbawa. To facilitate the tourists to get valid information about West Nusa Tenggara tourism with a closest track to the tourist sites, it is necessary to have a searching application and the location track of the tourist spots.

The application of tourist spot provides information such as descriptions of sights, address tourist attractions, photo gallery and the facilities available and the closest track to where Google Maps in the form of tourism[5]. Google Maps can display map locations and routes closest to the user's position sought after tourist sites. Determination of the position of the user and position of Tourism sites using GPS (Global Positioning System)[6]. The result of the study is the creation of location-based services applications to search a place of tourism in West Nusa Tenggara province so as to help tourists visiting the West Nusa Tenggara. Benefit from the application of this Smartphone is the closest in terms of determining the path articulated in this study. Not only that, this application can also display a map that can assist travelers in finding the shortest route to the location that will be addressed through the Google Maps APIs.

II. Review of Related Literature

A. Android

Android is a mobile operating system that grew in the middle of other operating systems are evolving today. Other operating systems such as Windows Mobile, i-Phone OS, and Symbian also offer a wealth of content and an optimal to walk on Smartphone devices[7]. However, the existing operating system is running with core apps built prioritize their own without seeing the great potential of third-party applications[5].

Android is a new generation mobile platform, the platform provides an opportunity for developers to develop applications according hoped. Android was developed by Google along with the Open handset Alliance (OHA)[8], which is an alliance of mobile device that consists of 34 companies hardware, software and telecommunications to support the development of open source on mobile devices[9].

Some of the main features of Android, among others WiFi hotspots, Multi-touch, multitasking, GPS, support java, supporting multiple networks (GSM / EDGE, IDEN, CDMA, EV-DO, UMTS, Bluetooth, Wi-Fi, LTE, and WiMAX) and also the basic capabilities of mobile phones in general[10].

B. Location Base Service

Location Based Service is the information service that can be accessed by mobile devices through the network and can display the geographical position where the mobile device [11]. Location Based Service technology works on the lines of Geographic Information System (GIS) and takes aid of Global Position System (GPS) to derive the position of receiver[12]. This technology consists of a device that collect, store, analyze, and distribute data based on the user requirements to the earth coordinate system.

Location based services are divided into two main elements [6]:

1) Location Manager (API Maps)

Application Programming Interface (API) Maps provide facilities to display, manipulate the map and other features.

2) Location Providers (API Location)

Location providers provide location search technology used by the device. API location associated with GPS data and location data in real time. By Location providers we can determine the user's location at this time, displacement and proximity to certain locations[13].

C. Google Maps

Google Maps is a free online map service provided by Google. In addition to providing free online map service, Google also provides APIs (Application Programming Interface) that enables application developers can combine Google maps into developed applications[4]. Google Maps API is a library in the form of JavaScript. Google maps display can be selected by with the wishes of both based on an original photo or an image map of the course[14][4].

D. Global Positioning System (GPS)

Global Positioning System (GPS) is a navigation system using satellite technology signals. Satellite sent to earth micro wave then received by a receiver on surface. The receiver will collect information such as time, location in the form of latitude, longitude and the elevation and speed [15].

At the Global Positioning System devices have GPS Tracker also called GPS Tracking. GPS Tracking is a technology AVL (Automated Vehicle Locater) which allows users to track the position of the vehicle, or fleet car in a state of Real-Time[16]. GPS Tracking utilizing a combination of GSM and GPS technology to determine the coordinates of an object, and then translate it in the form of digital maps [17].

III. Research Methods

A. Location and Place Research

The research location is housed in the Laboratory of Computer Engineering Academy of Information Management and Computer Mataram for three months, then two months later in the form of test applications from various access points to test application.

B. Data Collection Methods and Data Analysis Methods

Collecting data in this study used two types of data collection techniques are techniques of collecting primary data and secondary data collection techniques

1) Primary data collection techniques

Primary data were collected through interviews, questionnaires and observation of direct observation of the object data. Object data in question is data of Global Positioning System (GPS), Google Maps and information about the various tourist attractions in the Province of West Nusa Tenggara.

2) Secondary data collection techniques

In this study, secondary data obtained by studying, steeped in reference sources or literature relevant to this study are related about Android, Java programming and about the LBS application.

Data analysis method used in this research is descriptive qualitative method.

C. Research Procedure

The procedure describes the flow of research studies ranging from the preparation phase, analysis phase, design phase, implementation phase, test phase to the reporting phase. Research procedure can be described as in figure 1:

IV. Design and Implementation

A. Design Structure Navigation

Navigation structure illustrated the flow of display applications that can be comprehended by the users. Navigation structure that used in this application is navigation structure combination that consists of a hierarchical navigation structure, linear and non-linear. Navigation structure can be explained as in figure 2:

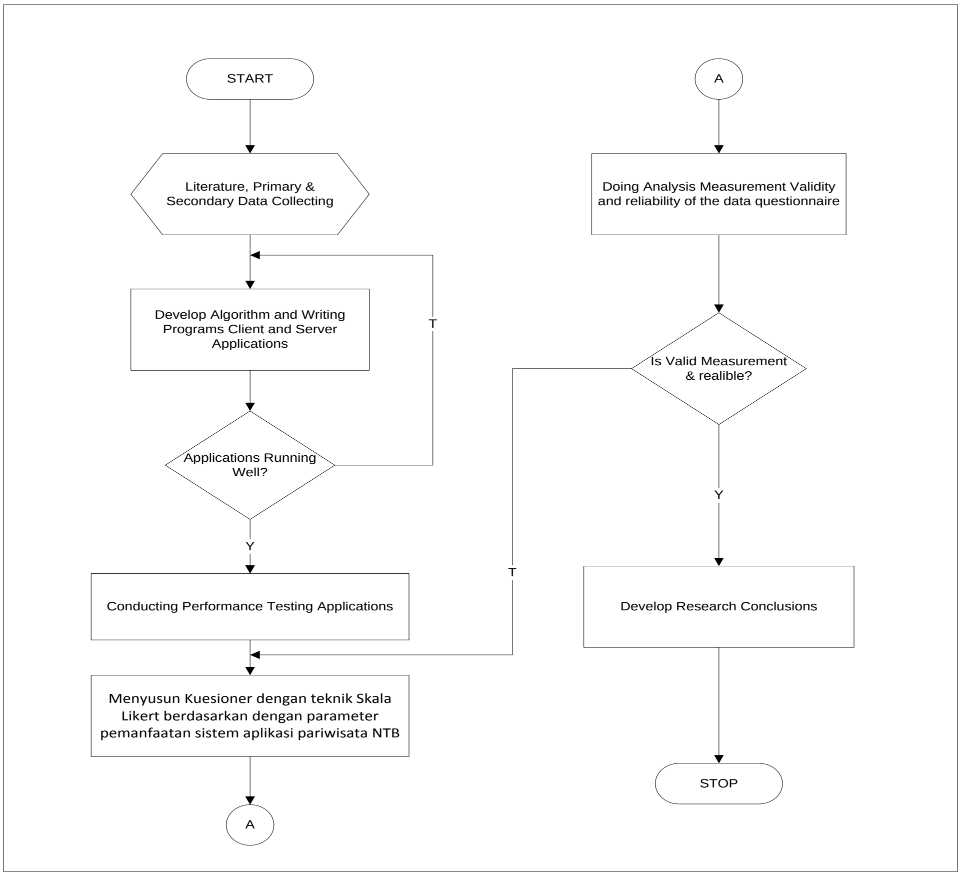

START
Literature, Primary & Secondary Data Collecting
Develop Algorithm and Writing Programs Client and Server Applications
T
Applications Running Well?
Y
Conducting Performance Testing Applications
Menyusun Kuesioner dengan teknik Skala Likert berdasarkan dengan parameter pemanfaatan sistem aplikasi pariwisata NTB
A
A
Doing Analysis Measurement Validity and reliability of the data questionnaire
Is Valid Measurement & realible?
T
Y
Develop Research Conclusions
STOP

Fig. 1. Research procedure

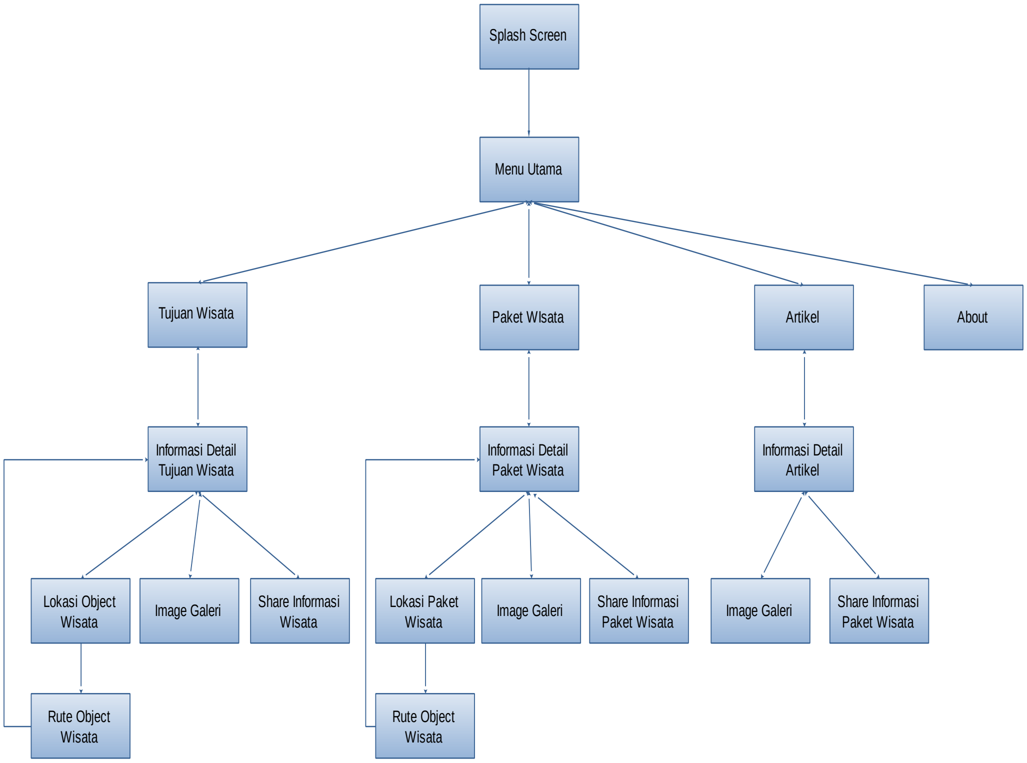


Fig. 2. Structure navigation

Application starts by displaying a splash screen for 5 seconds to give the application time data collection to the server, after which it will go to the main menu. Further on the main page there are 4 menus that is a tourist destination, travel packages, articles and about. When the user selects the menu of tourist destinations, then the application will display detailed information about tourist destinations in the form of drawings and a detailed explanation of the tourist attractions. On the detail page of the information contained three menus that tourism object location, photo galleries and share travel information. Location object only displays the coordinates tourist attractions on Google Maps, if the coordinates we choose the application will display the path to the navigation or in the form of tourist attractions ranging from the Google Maps to the user's position coordinates tourist attractions.

B. The Need of Hardware and Software

The application does not require high hardware specs, but the better the hardware specifications used, the better performance of the application will be.

1) Server Specifications

The making of tour android-based application uses the following hardware:

- Processor Intel Core i3-2367M @ 1.40 GHz
- Graphics Cards nVidia Geforce GT 620M 1 GB
- RAM 4 GB DDR 3
- Display 14.1" WXGA Acer Crystal Brite LCD
- Hard disk 500GB HDD

2) Android Specifications

To run the travel application android device requires the following specifications:

- Handheld Samsung Galaxy A3
- Operating System Android version 4.4 Kit Kat
- CPU Quad-Core 1.2 GHz Cortex-A53
- Memory RAM 1 GB
- External Memory 16 GB

3) Software Specification

The software used to create applications NTB tour are as follows:

- Java Development Kit (JDK) 8u45-nb-8_0_2 Windows-x64
- Android Software Development Kits (SDK)
- Eclipse Kepler and ADT
- Kies Samsung version 3

C. Design System

The steps of Application System development can be arranged as follows:

1) Literature Study

Literature study conducted to explore the theory and development of existing research and simultaneously understand the problems and solutions.

2) Collecting primary and secondary data

3) Developing algorithms and determine the features provided by the applications that run on both the client side and on the server side as well designing database.

4) Writing program at the Client and Server

5) Performing system testing applications in both the client side and server side.

D. Design Interface

1) Main Menu Page

This page is a page that will appear after the splash screen. This page contains a list of tourist attractions, travel packages, and Tourism news of West Nusa Tenggara. Each item can be selected to display detailed information, photo galleries and tourist sites such as Google maps of the item.

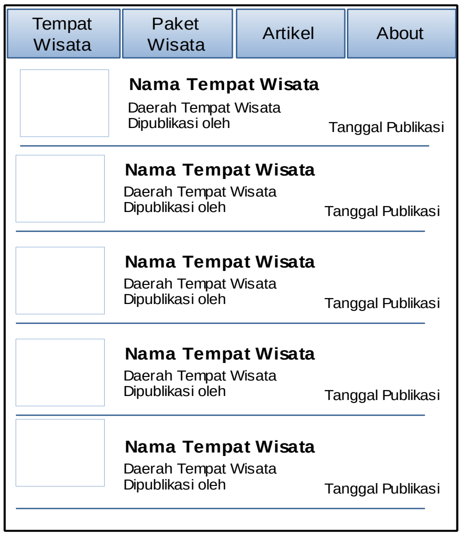


Fig. 3. Design main menu page

2) Detail Information Page

This page displays detailed information about tourist attractions. On this page, there is a menu photo gallery, tourist sites and share information through social media.

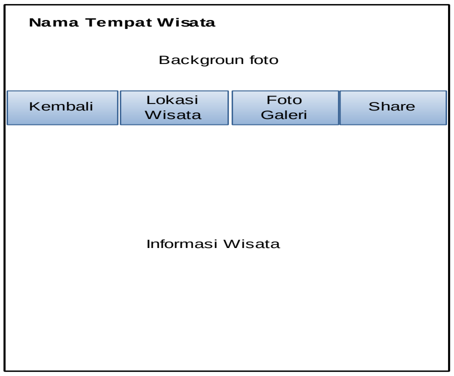


Fig. 4. Design detail information page

3) *Tourism Route Page*

Tourism route page showed travel route or track route to the destination from tourists' location to the place where the position is.

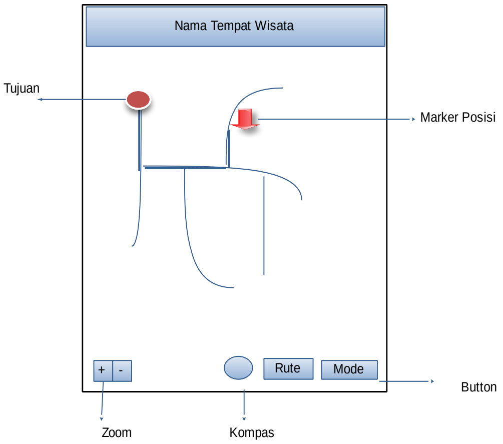


Fig. 5. Page tourism route design

V. Results and Discussion

A. *Test Result*

The results from testing of each component will be showed in the following table:

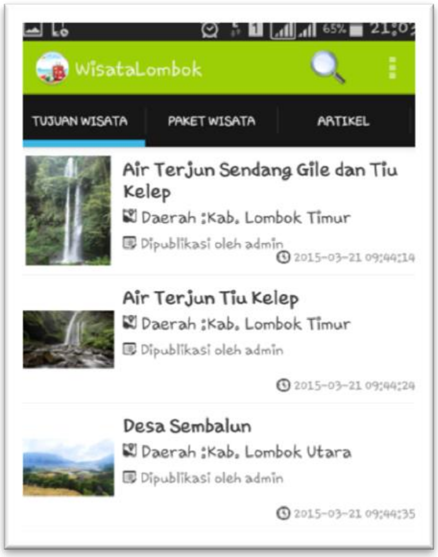
TABLE I. RESULTS FROM THE TESTING

No	Testing	Information	Result
1	Splash screen display	Splash screen displays for 5 seconds	Valid
2	Display a notification if there is no internet connection	This page appears when there is no internet connection on the Smartphone	Valid
3	Main menu display	The main menu displays a list of tourist attractions in a full page. If the page scrolls down next tourist spot, the data will be displayed.	Valid
4	Detail Tourism information display	This page showed detailed information about tourist attractions, including pictures and detailed description.	Valid
5	Google maps display	This page shows Google maps to show the location of the selected tourist spots and marked with a marker Google maps.	Valid
6	Google maps Tourism route display	This page appears when selecting marker on the page Google maps tourist sites. This page displays the route to the tourist attractions along the travel time.	Valid
7	Image gallery display	This page displays photos of the beauty of the tourist sites are selected. The photos displayed in the form of a slide show that slides left to the right.	Valid

From the table above, it can be concluded that the application is running by the good design without any error when run on the emulator on a Smartphone device. This application has also been tested by 50 different users on different Smartphone as well as on different screen sizes. From the results of these experiments, there were no problems or errors on run time.

B. *Display Results*

The result of the design that has been gained found that the application has been implemented in the form of Android applications. This application will run if the device is connected to the internet and GPS to determine the tourist position or user applications. The application will display a page splash screen for 5 seconds to allow time for data collection on the server and then it will display the main menu page as shown below:



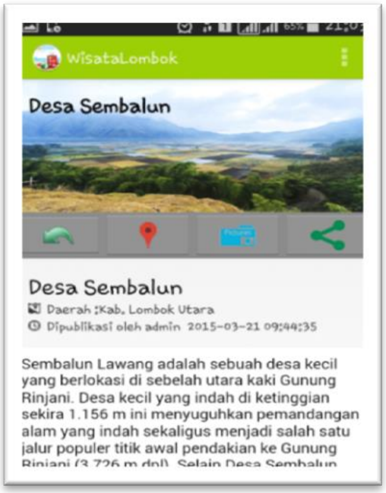
Air Terjun Sendang Gile dan Tiu Kelep
Daerah :Kab. Lombok Timur
Dipublikasi oleh admin
2015-03-21 09:44:14

Air Terjun Tiu Kelep
Daerah :Kab. Lombok Timur
Dipublikasi oleh admin
2015-03-21 09:44:24

Desa Sembalun
Daerah :Kab. Lombok Utara
Dipublikasi oleh admin
2015-03-21 09:44:35

Fig. 6. Main menu page

To view detailed information, select the type of Tourism information that will be displayed. On this page simply displays a list of tourist attractions, tour packages offered by the agent as well as news about Tourism activities. To view detailed information about each of the attractions, tour packages and select the articles or news items you want to display detail information pages that appear attractions such as the following picture:



Desa Sembalun

Desa Sembalun
Daerah :Kab. Lombok Utara
Dipublikasi oleh admin 2015-03-21 09:44:35

Sembalun Lawang adalah sebuah desa kecil yang berlokasi di sebelah utara kaki Gunung Rinjani. Desa kecil yang indah di ketinggian sekira 1.156 m ini menyuguhkan pemandangan alam yang indah sekaligus menjadi salah satu jalur populer titik awal pendakian ke Gunung Rinjani (3.726 m dpl). Selain Desa Sembalun

Fig. 7. Detail information Tourism page

This page displays detailed information about tourist attractions. To view the position or the select key tourist sites Google marker so that it will display the form Google Maps travel position such as the picture below:

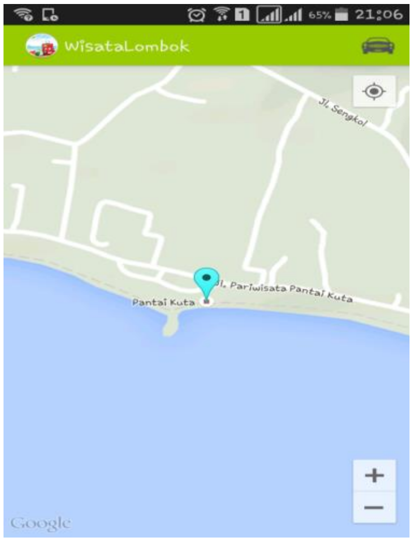


Fig. 8. Google maps page

This page simply displays the marker Google maps of the location of the tourist attractions. To view the route of the location, click the marker that appears. There will be a deal, if she or he wants to use Google maps or not and the choice of using the route or pathway that will be used. Choose the shortest path or paths with the shortest travel time on Google maps so it will display the route as shown below:

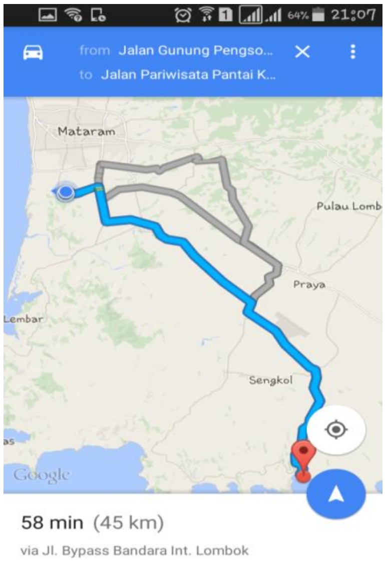

from Jalan Gunung Pengso...
to Jalan Pariwisata Pantai K...
Mataram
Pulau Lomb
Praya
Lembar
Sengkol
Google
58 min (45 km)
via Jl. Bypass Bandara Int. Lombok

Fig. 9. Google maps route tourism page

Also there is a button to display the position as well as the tourist attractions, there is also a menu to display the photo gallery that displayed the location of the tourist attractions in the form of a grid. The displayed image can be displayed in full-screen size by scrolling left or to right to display the next image.

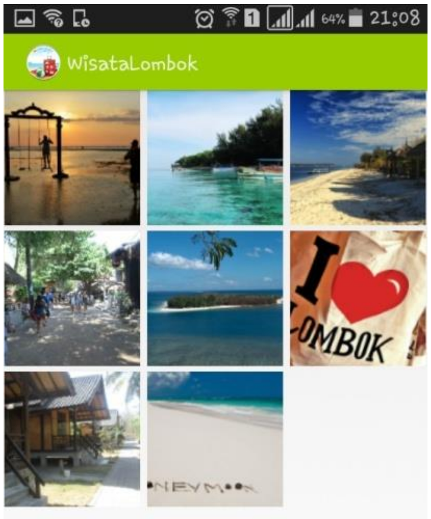


Fig. 10. Image gallery page

VI. Conclusion

From the results of the study, based on the correspondence amongst the input, process, and output, it can be concluded that the Tourism applications can run well on android or Smartphone. The application can be run if the Smartphone has a good access to the Internet and have an active GPS. The travel application provides convenience for the tourists when visiting West Nusa Tenggara in finding the Tourism sites.

Bibliography

[1] A. Subhani, "Potensi obyek wisata pantai di kabupaten lombok timur," 2010.

[2] Z. Amrulloh, "Pemberdataan Masyarakat Berbasis Pariwisata pada Dusun Tradisional Sasak Sade Lombok NTB," 2014.

[3] A. Sasongko, J. S. Informasi, F. Ilmu, and T. Informasi, "Aplikasi pemesanan makanan dan minuman pada rumah makan."

[4] D. G. Parrangan, "Pengembangan Indoor Location Based Service Menggunakan Wireless Positioning Pada Android," 2013.

[5] B. Anwar, H. Jaya, P. I. Kusuma, P. Studi, and S. Komputer, "Implementasi Location Based Service Berbasis Android untuk Mengetahui Posisi User," pp. 121–133, 2013.

[6] B. R. Rompas, a a E. Sinsuw, S. R. U. a Sompie, and a S. M. Lumenta, "Aplikasi Location-Based Service Pencarian Tempat Di Kota Manado Berbasis Android," no. 1, pp. 1–11, 2009.

[7] M. G. Aribowo, "Perancangan Aplikasi Pencarian Lokasi Bank Di Yogyakarta Dengan Location Base Service Untuk Android," 2013.

[8] N. T. Z. A, "Membangun Aplikasi Layanan Pencarian Lokasi Kuliner Terdekat Di yogyakarta Berbasis Android," 2012.

[9] B. D. Sarode and P. P. P. Karde, "Personal Service Areas for Location-Based Wireless Web Applications on Android Platform," vol. 4, no. 1, pp. 205–210, 2015.

[10] A. Sinsuw and X. Najoan, "Prototipe Aplikasi Sistem Informasi Akademik Pada Perangkat Android," pp. 1–10, 2013.

[11] W. Kusuma, A. K. Yapie, and E. S. Mulyani, "Aplikasi Location Based Service (LBS) Taman Mini Indonesia Indah (TMII) Berbasis Android," Semin. Nas. Apl. Teknol. Inf. 2013, pp. 13–18, 2013.

[12] A. Kushwaha and V. Kushwaha, "Location Based Services using Android mobile Operating System.pdf," Int. J. Adv. Eng. Technol., vol. 1, no. 1, pp. 14–20, 2011.

[13] L. Calderoni, D. Maio, and P. Palmieri, "Location-aware mobile services for a smart city: Design, implementation and deployment," J. Theor. Appl. Electron. Commer. Res., vol. 7, no. 3, pp. 74–87, 2012.

[14] A. D. Laksito, “Analisis Model Kematangan Tata Kelola teknologi Informasi di STMIK AMIKOM Yogyakarta menggunakan Framework COBIT,” 2012.

[15] M. Singhal and A. Shukla, “Implementation of Location based Services in Android using GPS and Web Services,” Int. J. Comput. Sci. Issues, vol. 9, no. 1, pp. 237–242, 2012.

[16] Wahyu Widayanto, “Perancangan Aplikasi Pengingat berdasarkan Location Base Service Berbasis Android,” 2013.

[17] S. Kumar, M. Qadeer, and a Gupta, “Location based services using android (LBSOID),” Internet Multimed. Serv. …, pp. 1–5, 2009.

Robust Fuzzy-Second Order Sliding Mode based Direct Power Control for Voltage Source Converter

D. Kairous and B. Belmadani
Laboratoire de Génie électrique et Energie Renouvelable –LGEER-. Departement of Electrical Engeneering.
Faculty of Technology. University of Hassiba Ben-Bouali at Chlef -UHBC-.
Chlef. Algeria

***Abstract*—This paper focuses on a second order sliding mode based direct power controller (SOSM-DPC) of a three-phase grid-connected voltage source converter (VSC). The proposed control scheme combined with fuzzy logic aims at regulating the DC-link voltage of the converter and precisely tracking arbitrary power references, in order to easily control the system's power factor. Therefore measures are proposed to reduce the chattering effects inherent to sliding-mode control (SMC). Simulations performed under Matlab/Simulink validate the feasibility of the designed Fuzzy-SOSM. Simulation results on a 1kVA grid-connected VSC under normal and faulted grid voltage conditions demonstrate good performance of the proposed control law in terms of robustness, stability and precision.**

Keywords—AC-DC power converters; Bidirectional power flow; Fuzzy logic; Sliding mode control; Direct Power control

I. Introduction

Voltage source converter (VSC) is a very useful device. Among many desirable characteristics [1], the VSC permit the independent control of active and reactive power, what makes it very attractive for power conditioning and transmission. As such, the VSC is a key element in power electronics-based equipment for flexible AC transmission systems (FACTS), high-voltage direct current (HVDC) systems and active power filters (APFs). Furthermore, with the development of smart grids involving distributed generation and the interconnection of renewable power generation systems [2], such as wind farms, photovoltaic, solar thermal plants, etc., the need for power flow steering increases. In this context, efficient control strategies for grid connected VSCs are relevant more than ever.

PI control is a mature and proven method which is very popular for industrial applications. Effective regulation of the active and reactive power flows between the grid and the VSC can be achieved with PI-based vector control. However, in order to apply PI regulation techniques at vector control, feed-forward compensation must be employed to eliminate cross-coupling and linearize the system's model. Owing to this linearization, the control scheme and the performance of vector control become respectively very sensitive to the system's parameters and dependent on their accuracy [3].

Sliding-mode control (SMC) is a variable structure control strategy that uses a special version of on-off control, or high frequency switching, to achieve robust control of non-linear systems [4-5]. A good mathematical background for SMC is presented in [6], along with a demonstration of its applicability to electric drives.

Sliding mode control is based on the theory of the variable-structure systems (VSS). In a closed loop system, one way to change the structure is to use different controllers depending on the state of the system. The main idea is to switch rapidly between strong control actions when the system deviates from the desired response. To do this, the closed-loop system behavior must be described by a switching surface in the state space. Then, the control actions can be chosen so that the net effect of the switching (*chattering*) will be to move the system towards the switching surface [4]. When this is achieved, the system will "slide" along the switching surface, giving what one calls the "sliding-mode". A regain of interest in SMC occurred in the 1980s with the increasing availability of powerful microprocessors. The robustness of the method then began to be recognized along with its relative immunity toward external disturbances and its low sensitivity to system parameters variations [5]. In this work, a second order sliding mode (SOSM) control scheme is proposed, to ensure tracking of the DC bus voltage and rotor power factor in a wind power systems. The control technique named super-twisting is based on a bounded continuous control with discontinuities in the control derivative [5].

In objective to attenuating the chattering effect the proposed control will be associated with fuzzy logic [7]. In fact, during the last decade, the fuzzy logic control (FLC) has been selected as suitable control solution in the field of power electronics and drives [8]. Among the advantages provided by this control approach over the conventional controllers in other hand it does not require accurate mathematical model. It can thus work with inaccurate inputs, handle nonlinear model systems and easily reach performances of ideal digital PI controllers. On the other hand, The SMC appears as a simple way to design robust controllers for electrical drives, a powerful technique to eliminate sensors in electrical machine drives. Furthermore, the SMC does not require many computational operations and remains insensitive to plant parameters variations. Accordingly, this paper aims at combining the advantages of FLC and SMC for robust control electronic power converter. This approach has been successful applied in wide area.

The rest of this paper is organized as follows. In section II, the global wind energy conversion system (WECS) is described. Section III then presents the model of the VSC to be controlled. In section IV and V, the details of the sliding mode control law and fuzzy logic are presented respectively.

Then, the section VI described the used space vector SV-PWM. Numerical simulations results and analysis for both transient and steady state are presented in section V and conclusions are drawn in section VI.

II. Exchange Powers by Grid Side Converter

This work addresses the usual scheme adopted for the doubly fed induction generator (DFIG) in a variable speed, constant frequency wind power generation system. Imposing slip frequency, amplitude and phase on the rotor voltage permits to achieve constant frequency, constant voltage output at the stator.

The structure of this system is depicted on Fig. 1. The rotor is connected to grid via two voltage-source PWM converters, connected back-to-back: the rotor-side converter (RSC) and the grid-side converter (GSC) [9].

As the Fig.1 suggests, the two converters do have the same circuit structure. They simply alternate between rectifier and inverter function, depending on the flowing direction of rotor energy.

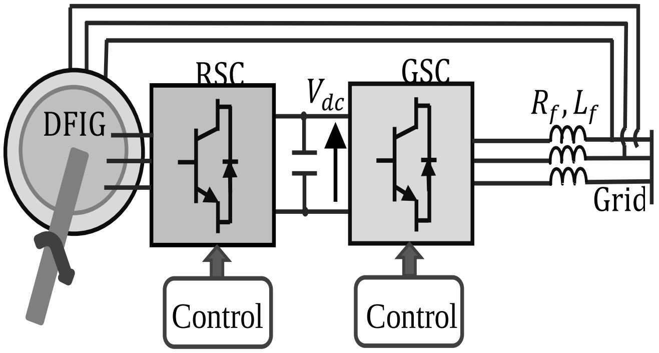


Fig. 1. Bloc diagram of WECS connected to the grid

The RSC feeds excitation current into rotor winding and achieves flux orientation, to catch maximum wind energy and adjust reactive power output. Depending on the rotor speed relative to the generator synchronous speed, the RSC works as either an inverter (low speed) or a rectifier (high speed). When the generator is working at synchronous speed, the RSC feeds direct current excitation into rotor, working as a Chopper. The GSC works in dual cooperation with the RSC, permitting seamless energy flow in both directions. It also controls DC bus voltage and adjust grid-side power factor, which makes the entire wind power system to have a flexible reactive power regulation.

III. System Model

The GSC system can be modeled as an ideal VSC. The equivalent circuit of Fig. 2 is a simplified representation of the grid-connected VSC in the stationary *αβ* reference frame.

KVL applied to the equivalent circuit of Fig. 2 gives the relation between line current and the supply voltage as [1]

$$U_{\alpha\beta} = I_{\alpha\beta}R_f + L_f\frac{dI_{\alpha\beta}}{dt} + V_{\alpha\beta} \tag{1}$$

From (1), the derivative of the current can be written as

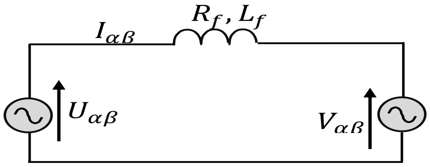


Fig. 2. Equivalent circuit of VSC in stationary frame

$$\frac{dI_{\alpha\beta}}{dt} = \left(1/L_f\right)\left(U_{\alpha\beta} - R_f I_{\alpha\beta} - V_{\alpha\beta}\right) \tag{2}$$

The power exchange with the electric grid, as seen from the network side (in stationary frame), is defined by (3)

$$\begin{aligned} P &= \left(-3/2\right)\left(U_\alpha I_\alpha + U_\beta I_\beta\right) \\ Q &= \left(-3/2\right)\left(U_\beta I_\alpha - U_\alpha I_\beta\right) \end{aligned} \tag{3}$$

From (3), the derivative of the current may be written as

$$\begin{bmatrix} dI_\alpha/dt \\ dI_\beta/dt \end{bmatrix} = \left(\frac{-2}{3U^2}\right)\frac{d}{dt}\left\{\begin{bmatrix} -U_\alpha & -U_\beta \\ -U_\beta & U_\alpha \end{bmatrix}\begin{bmatrix} P \\ Q \end{bmatrix}\right\} \tag{4}$$

Knowing that the derivative of the grid voltage, expressed in the stationary reference frame, is

$$\begin{aligned} dU_\alpha/dt &= wU\cos(wt) = -wU_\beta \\ dU_\beta/dt &= wU\sin(wt) = wU_\alpha \end{aligned} \tag{5}$$

Equating relations (2) and (4) gives an expression for the derivatives of the powers as

$$\begin{aligned} dP/dt &= \left(-3/2L_f\right)\left[\left(U_\alpha^2 + U_\beta^2\right) - \left(U_\alpha V_\alpha + U_\beta V_\beta\right)\right] \\ &\qquad - \left(R_f/L_f\right)P - wQ \\ dQ/dt &= \left(-3/2L_f\right)\left[-\left(U_\beta V_\alpha + U_\alpha V_\beta\right)\right] - \left(R_f/L_f\right)Q \\ &\qquad + wP \end{aligned} \tag{6}$$

This last expression relating the time derivative of the powers to the voltage is essential to the design of the SMC below.

IV. Controller Design

The control problem can be stated as: “find the sequence and the duration of the on-off states for converter switches so that the given dynamical specifications of the closed-loop system are satisfied”.

This formulation is a good starting point for introducing a sliding-mode control since it intimately links the selection of the PWM switching pattern to the dynamics of the system and will result in a PWM pattern that directly forces the close-loop behavior of the converter to slide along the desired switching surface.

The first design step consists in describing a suitable switching surface along which the system will be allowed to slide. Then one must define the reaching law that will force the system to move towards this very specific surface. Finally, an adequate technique should be introduced to alleviate the side effects of the chattering of the command on the switching surface [6].

A. Switching surface

The switching surface can be designed to ensure that the power P and Q track their references P^* and Q^*. The description of such a surface can be obtained by taking the errors on the actual values of the powers as [10].

$$\begin{aligned} e_P(t) &= P^* - P \\ e_Q(t) &= Q^* - Q \end{aligned} \tag{7}$$

Since the desired motion of the system will be obtained for $e(t) = 0$, the switching surface can be defined in terms of these errors as

$$\begin{aligned} S_P &= e_P(t) + K_P\int e_P(\tau)d\tau - e_P(0) \\ S_Q &= e_Q(t) + K_Q\int e_Q(\tau)d\tau - e_Q(0) \end{aligned} \tag{8}$$

Where K_P and K_Q are positive control gains.

Whenever the system reaches the switching surface and slides along it, we have:

$$S_P = S_Q = dS_P/dt = dS_Q/dt = 0 \tag{9}$$

According to (8), when the derivatives of S_P and S_Q are zero, we get the derivative for the errors as

$$\begin{aligned} de_P(t)/dt &= -K_P e_P(t) \\ de_Q(t)/dt &= -K_Q e_Q(t) \end{aligned} \tag{10}$$

The solution a=of (10) are decreasing exponentials, which ensure that the power errors converge asymptotically to zero with the time constants of $1/K_P$ and $1/K_Q$.

B. Sliding Mode based Direct Power Control (SM-DPC) Law

The second step focus on the synthesis of an appropriate control law to force the state trajectories to slide along the switching surface. Taking the time derivative of (8), we get

$$\begin{bmatrix} dS_P/dt \\ dS_Q/dt \end{bmatrix} = \begin{bmatrix} F_P \\ F_Q \end{bmatrix} + D\begin{bmatrix} V_\alpha \\ V_\beta \end{bmatrix} = 0 \tag{11}$$

This gives the equivalent control voltages

$$\begin{bmatrix} V_\alpha \\ V_\beta \end{bmatrix} = \begin{bmatrix} V_{Peq} \\ V_{Qeq} \end{bmatrix} = -D^{-1}\begin{bmatrix} F_P \\ F_Q \end{bmatrix} \tag{12}$$

Where, according to (6) and (7)

$$\begin{aligned} F_P &= \left(3/2L_f\right)\left(U_\alpha^2 + U_\beta^2\right) + \left(R_f/L_f\right)P + wQ \\ &\qquad + K_P(P^* - P) \\ F_Q &= \left(R_f/L_f\right)Q - wQ + K_Q(Q^* - Q) \end{aligned} \tag{13}$$

And

$$-D^{-1} = \left(2L_f/3U^2\right)\begin{bmatrix} U_\alpha & U_\beta \\ U_\beta & -U_\alpha \end{bmatrix} \tag{14}$$

Relation (12) above is the control law for the power controller. The control voltages are directly the converter's output voltages in the stationary reference frame, which are used to generate the PWM switching scheme that will keep the system on the switching surface.

In sliding-mode control design, one must derive the conditions under which the control law will stabilize the system by driving its state trajectory to an equilibrium surface with good robustness. We use the quadratic Lyapunov function

$$W = 0.5S^T S \geq 0 \tag{15}$$

The time derivate of which can be expressed in terms of the state trajectories (12)

$$dW/dt = d(0.5S^T S)/dt = S^T\, dS/dt = S^T(F + DV) \tag{16}$$

In order to guaranty stability, the control law must thus be modified so that the time derivative *dW/dt* is definite negative when $S \neq 0$. This condition can be achieved with the control law

$$\begin{bmatrix} V_\alpha \\ V_\beta \end{bmatrix} = -D^{-1}\left(\begin{bmatrix} F_P \\ F_Q \end{bmatrix} + \begin{bmatrix} K_{\mathrm{sgn}\,P} & 0 \\ 0 & K_{\mathrm{sgn}\,Q} \end{bmatrix}\begin{bmatrix} Sgn(S_P) \\ Sgn(S_Q) \end{bmatrix}\right) \tag{17}$$

where K_{sgnP} and K_{sgnQ} are positive control gains, while *Stgn*(S_P) and *Sgn*(S_Q) are sign functions. By setting appropriate constant values, stability can be achieved.

C. Power chattering attenuation

Because the SM involve fast switching of the command, unexpected chattering of the command may result in undesirable behavior of the system, e.g. excite some high frequency mode of the system generating instability.

One way to alleviate this problem is to modify the sign function in equation (17) to introduce a boundary layer that smooth the command around the sliding surface. Such a modification in the surface neighborhood can be written

$$Sgn(S_i) = \begin{cases} 1 & if\ S_i > K_{\mathrm{sgn}\,i} \\ S_i/\lambda_{\mathrm{sgn}\,i} & if\ |S_i| > K_{\mathrm{sgn}\,i} \\ -1 & if\ S_i < K_{\mathrm{sgn}\,i} \end{cases} \tag{18}$$

Where i stands for P or Q and $\lambda_{\mathrm{sgn}i}$ is the width of the boundary layer.

To ensure an even more robust tracking of the powers exchanged by the converter, we apply the super-twisting algorithm (STA), by adding a new term V_{iST} to the control law V_{ieq} given by (17).

The sliding along the surface (9) can be obtained by applying only the new law

$$V_i = V_{iST} + V_{ieq} \tag{19}$$

The new term V_{iST} is added to the equivalent control law, based on the SOSM (STA approach) [10] [11].

The idea of the SOSM control techniques is to zero a function of the system's states, the sliding variable S, and its first time derivative $\dot{s}$. The function is designed based on the desired control objectives, to guarantee their achievement when $S = 0$.

The condition of stabilizing $S = \dot{S} = 0$ determines the 2-sliding manifold in the state space.

The SOSM algorithm acts taking the trajectories in the state space to the 2-sliding manifold in finite time and keeping them operating robustly on it, i.e., makes the system operate in SOSM.

Discontinuous control input of the SOMC directly influences the sign and the magnitude of the second order time derivative of the sliding manifold. Since its control structure is relatively simple and no much information is needed, it has become the most widely used high order sliding mode control method [11].

The main problem with high-order sliding mode algorithm implementations is the increased required information. Indeed, the implementation of an nth-order controller requires the knowledge of $\dot{S}, \ddot{S}, \ldots, \overset{(n-1)}{S}$.

There are several SOSM algorithms, each of them with their own characteristics. In particular, the super-twisting algorithm has a quite simple law and allows synthesizing a continuous control action with discontinuous time derivative (in contradiction to those of the sub-optimal algorithm). and it only requires measurements of surface s .

The control law in equation (19) is composed of switching control terms and equivalent control terms. Where V_{iST} are the switching control terms. The switching control terms make the system in any initial state reach the sliding manifold in finite time, which are calculated through application of the super-twisting algorithm in this paper.

V_{ir} are the equivalent control terms. The equivalent control terms make the system move along the sliding manifold under ideal conditions, and these terms can speed up the response of the system and reduce the steady-state errors.

The equivalent control terms are derived by letting $\dot{S}_P = \dot{S}_Q = 0$. The power errors can be defined as below

$$\begin{cases} \dot{e}_p = \dfrac{-3}{2L_f}\left[\left(U_\alpha^2 + U_\beta^2\right) - \left(U_\alpha V_\beta + U_\beta V_\alpha\right)\right] \\ -\left(\dfrac{R_f}{L_f}\right)P - wQ - \dot{P}^* \\ \dot{e}_Q = \dfrac{3}{2L_f}\left[\left(U_\beta V_\alpha + U_\alpha V_\beta\right)\right] - \left(\dfrac{R_f}{L_f}\right)Q - wP - \dot{Q}^* \end{cases} \tag{20}$$

If we define the functions H_p and G_Q as follows:

$$\begin{cases} H_p = \dfrac{-3}{2L_f}\left[\left(U_\alpha^2 + U_\beta^2\right) - \left(U_\alpha V_\beta\right)\right] - wQ - \dot{P}^* \\ H_Q = \dfrac{3}{2L_f}\left[\left(U_\beta V_\alpha\right)\right] - wP - \dot{Q}^* \end{cases} \tag{21}$$

Then we have, the second derivate of the errors as

$$\begin{cases} \ddot{e}_{pg} = \dfrac{3}{2L_f}U_\beta \dot{V}_\alpha - H_{pg} - \left(\dfrac{R_f}{L_f}\right)\dot{P} \\ \ddot{e}_{Qg} = \dfrac{3}{2L_f}U_\alpha \dot{V}_\beta + H_{Qg} - \left(\dfrac{R_f}{L_f}\right)\dot{Q} \end{cases} \tag{22}$$

The indices P and Q being the active and reactive powers indices respectively

In finite time, based on the STA approach, the terms V_{iST} are calculated according to [11]

$$\begin{cases} V_{PST} = -\lambda_{pg}\left|e_{Pg}\right|^{0.5} Sgn(e_{Pg}) - \chi_{pg}\int Sng(e_{Pg})\,dt \\ V_{QST} = -\lambda_{Qg}\left|e_{Qg}\right|^{0.5} Sgn(e_{Qg}) - \chi_{Qg}\int Sng(e_{Qg})\,dt \end{cases} \tag{23}$$

Of which only the bounds Γ_m, Γ_M and Φ are known:

$$\begin{cases} 0 < \Gamma_{mpg} < \dfrac{3}{2L_f}U_\beta < \Gamma_{Mpg}, \left|\dot{H}_{Pg} - \left(\dfrac{R_f}{L_f}\right)\dot{P}\right| < \Phi_{pg} \\ 0 < \Gamma_{mQg} < \dfrac{3}{2L_f}U_\alpha < \Gamma_{MQg}, \left|\dot{H}_{Qg} - \left(\dfrac{R_f}{L_f}\right)\dot{Q}\right| < \Phi_{Qg} \end{cases} \tag{24}$$

The sufficient conditions for finite-time convergence are,

$$\begin{cases} \lambda_{pg} > \dfrac{\Phi_{pg}}{\Gamma_{MPg}}, \quad \chi_{pg}{}^2 \geq \dfrac{4\Phi_{pg}(\lambda_{pg} + \Phi_{pg})}{\Gamma_{mPg}{}^2(\lambda_{pg} - \Phi_{pg})} \\ \lambda_{Qg} > \dfrac{\Phi_{Qg}}{\Gamma_{MQg}}, \quad \chi_{Qg}{}^2 \geq \dfrac{4\Phi_Q(\lambda_{Qg} + \Phi_{Qg})}{\Gamma_{mQg}{}^2(\lambda_{Qg} - \Phi_{Qg})} \end{cases} \tag{25}$$

This ensures that uncertainty on L_f and R_f will not be a threat to the robustness of the algorithm. However, accurate value of the V_{ieq} control terms in (19) will greatly contribute to lower the control effort done by the STA.

In practice, the parameters are never assigned according to inequalities. Usually, the real system is not exactly known, the model itself is not really adequate, and the parameters estimations are much larger than the actual values. The larger the controller parameters, the more sensitive the controller to any switching measurement noises. The right way is to adjust the controller parameters during computer simulations.

The value of $k_{sngi}sign(S_i)$ is determined by the fuzzy structure detailed in the section follow.

V. Fuzzy Controller

Fuzzy-logic control has the capability to control nonlinear, uncertain and adaptive systems with parameter variation. Fuzzy control does not strictly need any mathematical model of the plant.

Its control rule can be qualitatively expressed on the basis of logic-language variation and the fuzzy model of a plant is very easy to apply. In fact, fuzzy control is good adaptive control among the techniques discussed so far. In this paper, fuzzy-logic control is associated with sliding-mode control to generate the value of the disconnect component gain which ensures the precision and robustness of the control [12].

The general structure of a fuzzy-control system is shown in Fig. 3. There are two input signals to the fuzzy controller, the error E and the change in error CE, which is related to the derivative DE/dt of error. The closed-loop error E and change in error CE signals are converted to the respective scale factors, e=E/GE and ce=CE/GC. The output plant control signal DU is derived by multiplying the per unit by the scale factor GU, that is DU=du*GU, and then integrated to generate the U signal [13].

The scale factors can change the sensitivity of the controller without changing its structure. The fuzzy controller is composed of three blocks: fuzzification, rule bases, and defuzzification. The function of membership of each input signal (E, dE) is illustrated in Fig. 4. The fuzzy subsets are as follows: NB (Negative Big), Nm (Negative Medium), NS (Negative Small), Z (Zero), PS (Positive Small), PM (Positive Medium), PB (Positive Big). There are seven fuzzy subsets for each variable, which gives 7 * 7 = 49 possible rules, where typical rule is: "If E(pu) is PS and dE(pu) is PM, then dU (pu) PB." (Table.1). Defuzzification is done by the centroid method based on the Takagi-Sugeno-Kang inference method.

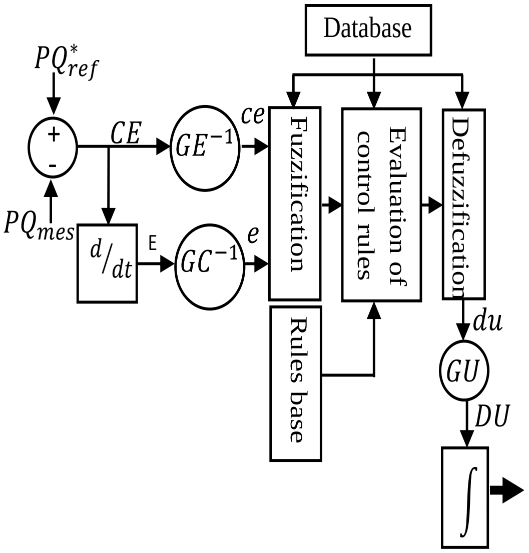


Fig. 3. Structure of the Fuzzy Controller

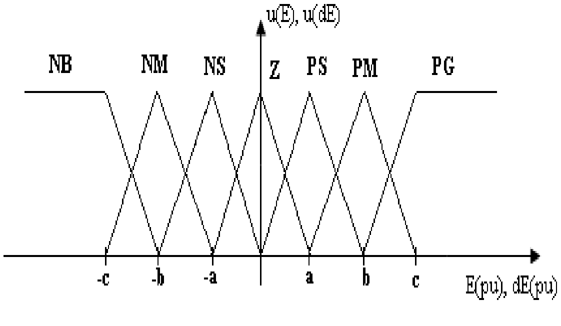


Fig. 4. a. Member functions

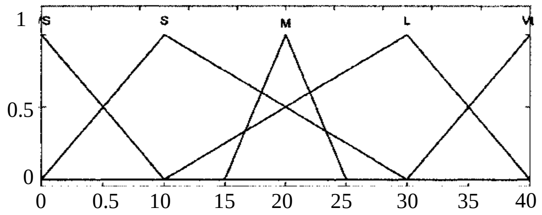


Fig. 4. b. Membership functions

TABLE I. RULES BASE

	NB	**NM**	**NS**	**Z**	**PS**	**PM**	**PB**
PB	Z	PS	PM	PB	PB	PB	PB
PM	NS	Z	PS	PM	PB	PB	PB
PS	NM	NS	Z	PS	PM	PB	PB
Z	NB	NM	NS	Z	PS	PM	PB
NS	NB	NB	NM	NS	Z	PS	PM
NM	NB	NB	NB	NM	NS	Z	PS
NB	NB	NB	NB	NB	NM	NS	Z

VI. SPACE VECTOR PWM

We use a Model of three-phase, tow-level source SV-PWM inverter with center-taped grounded DC bus. The relationship between the switching variable vector and voltage vector can be expressed below:

$$V_{\alpha\beta}=\begin{bmatrix}V_\alpha\\V_\beta\end{bmatrix}=C32\begin{bmatrix}V_{an}\\V_{bn}\\V_{cn}\end{bmatrix}=\frac{1}{3}\begin{bmatrix}2&-1&-1\\-1&2&-1\\-1&-1&2\end{bmatrix}\begin{bmatrix}C_a\\C_b\\C_c\end{bmatrix}\frac{V_{dc}}{2}\quad(26)$$

Where C_{32} is the Park transformation Matrix, C_a,C_b and C_c are the states of the interrupters.

Nabae and All in [14] divide the space vector plane into six sectors as shown in Fig.5. in the space vector approach, the desired reference vector is generated by time averaging the suitable discrete vltage vectors in sampling period T_s. For a given reference voltage *Uref* and angle in sector n , the volt-time balance is maintained by applying the active vectors 1, 2 and zero states together for durations T_1, T_2 and T_0 respectively, as given in following expressions :

$$\int_0^{T_z} U_{ref} = \int_0^{T_1} V_1 + \int_{T_1}^{T_1+T_2} V_2 + \int_{T_1+T_2}^{T_0} V_0 \quad(27)$$

$$T_1 = \frac{\sqrt{3}\,T_z \left|U_{ref}\right|}{V_{dc}}\left(\sin(\frac{n}{3}\pi)\cos(\alpha)-\cos(\frac{n}{3})\sin(\alpha)\right)\quad(28)$$

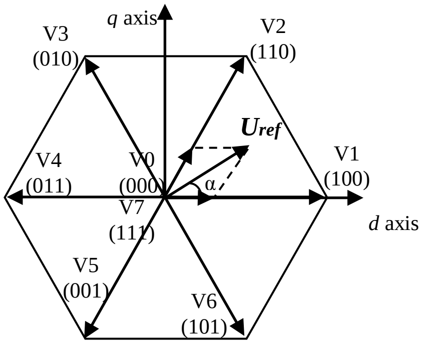


Fig. 5. Basic switching vectors and sectors

$$T_2 = \frac{\sqrt{3}T_z\left|U_{ref}\right|}{V_{dc}}(\sin(\alpha)\cos(\frac{n-1}{3}\pi) - \cos(\alpha)\sin(\frac{n-1}{3}\pi)) \quad (29)$$

$$T_0 = T_z - (T_1 + T_2) \quad (30)$$

With $(T_z = \frac{1}{fz})$

Where: n=1 through 6 (that is, Sector 1 to 6) and 0<α<60°

VII. Simulation Results

The power generated by the wind system is transmitted to the grid via the stator of the DFIG (Fig.1). On the other hand, only the slip power is exchanged between the grid and the rotor of the DFIG. When the DC link voltage is well regulated, the rotor current should be ensuring the smooth exchange of slip power.

Simulations have been carried out to demonstrate the effectiveness of proposed control strategy applied to the grid connected converters of a wind energy system. The proposed design is implemented in Matlab/Simulink software packages using parameters given in Table 2.

Simulink discrete models with step time of 1µs were used. A programmable 3-phase voltage source was used to simulate the grid with voltage fluctuations. The IGBTs of the GSC were driven by a SV-PWM generator.

The bloc diagram of Fig. 6 shows the proposed sliding mode power control for the GSC. An outer loop (not shown) contains PI controller to regulate the dc-link voltage and produce the active power reference. The control law developed in the previous section directly generates the converter voltage reference in the stationary αβ frame, according to the instantaneous values of the errors on the active and reactive powers.

The system is simulated for two different cases: first insteady state, in order to demonstrate the DC link voltage regulation and power factor control. In the second case, a transient state with system parameters variation was simulated. The objective of this second simulation is to prove the robustness of proposed strategy.

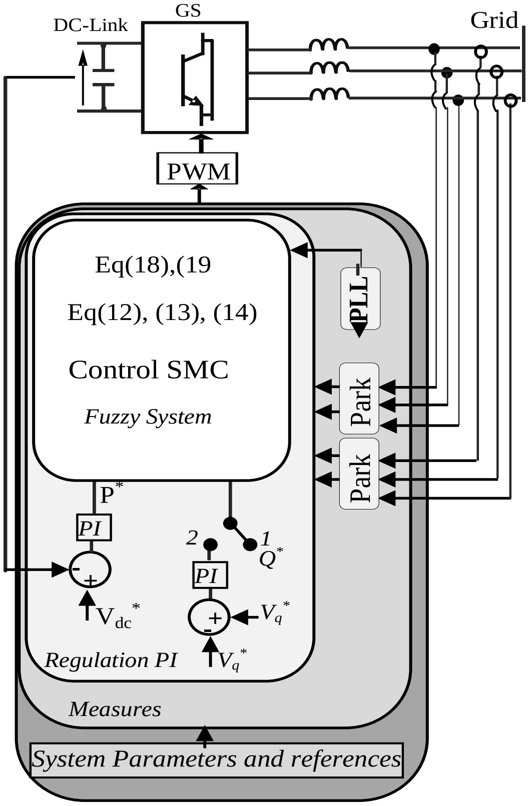


Fig. 6. Bloc diagram of WECS connected to the grid

Case1:

In this case, the voltage of the programmable source is fixed to constant frequency and magnitude.

The RSC and DFIG rotor are simulated by a constant current source and a well smoothed wind power is assumed. Imposing Q = 0, an initial current of Is =1.5A is supplied to the DC-link circuit in order to simulate a slip power extracted from the rotor. At t = 0.3s, a current step is induced, bringing its value to -1.5 A (Fig.7). This is done to demonstrate the system's ability to transmit and received power to/from the grid. At t = 0.6s the current is brought back to its initial value of 1.5 A. The results presented on Fig. 8 and 10 show that both the DC- link voltage and reactive power are unaffected by the dc current steps. Fig.9 shows the active power dynamics, starting at -600W and reaching the value of +600W on the first step at t=0.3s, and then going back to the initial value of -600W at 0.6s.This behavior corresponds to the transition from hypo- to hyper-synchronous operating modes of the DFIG. The power involved, which is transported by the

TABLE II. Parameters of the Tested GSC System

Rated power	1.2 kW	Line resistance	0.001 H
Grid rated voltage	208 V	Line inductance	0.001 Ω
Grid rated frequency	60 Hz	DC link voltage	400V
DC link capacitor	35000 µF		

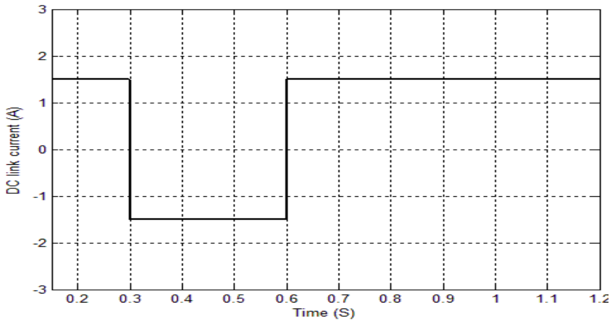


Fig. 7. DC link current

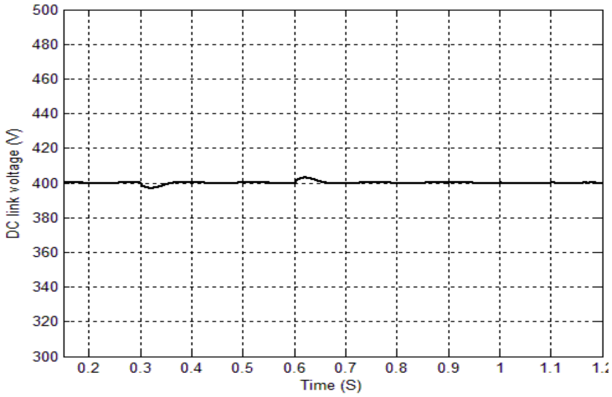


Fig. 8. DC link voltage

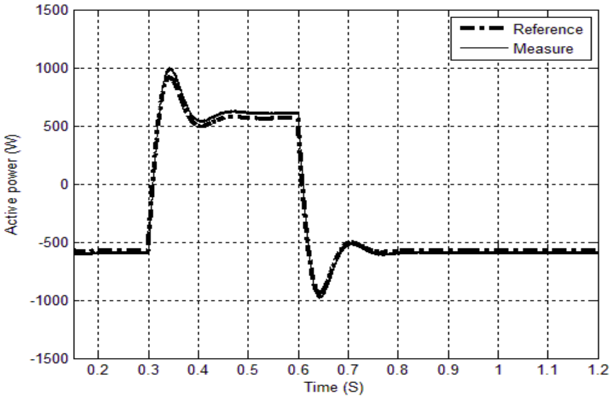


Fig. 9. Active power reference and response

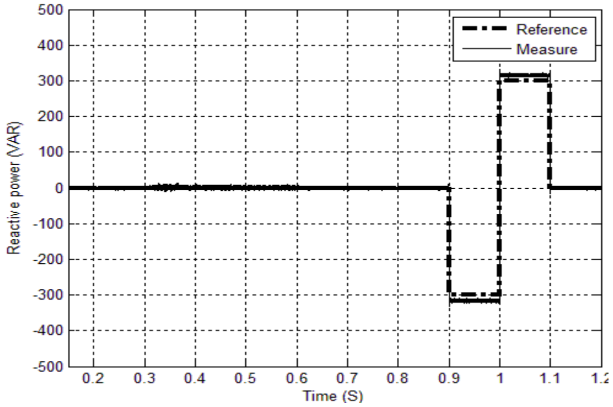


Fig. 10. Reactive power reference and response

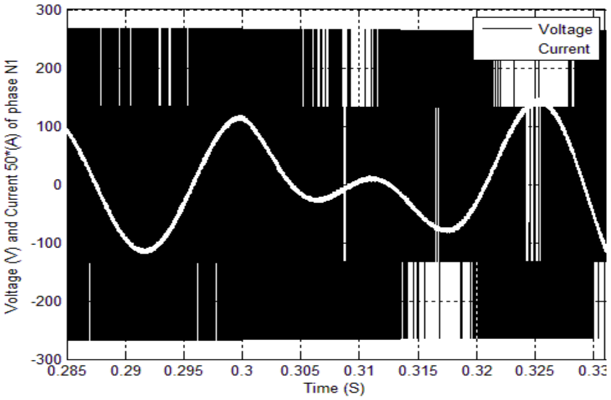


Fig. 11. Transition of current at first transition of the DC current

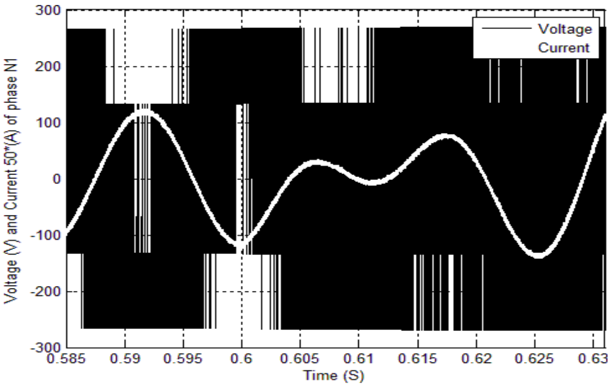


Fig. 12. Transition of current at second transition of the DC current

GSC, is 50% of the GSC nominal power. The passage between hyper- and hypo-synchronous modes is made without great transient current, which give attenuation at the power transient state. An overshoot in power peaking to twice the permanent value is considered reasonable.

In order to demonstrate the decoupling between active and reactive powers, the latter is varied by steps from 0.9 s to 1.1 s. At t = 0.9 s a step brings Q from 0 to -300 VAR then to +300 VAR at t = 1.0 s and finally back to 0 VAR at t = 1.1 s. (Fig.10)

As can be seen the step change of one control variable, i.e. active or reactive power, does not affect the other, and there is no high overshoot of the active and reactive powers. Fig.11 and Fig.12 show that the initial power factor of 1 is changed to -1 after the first step in the current Is, and then changed back from -1 to 1.

Case 2:

For this case, a perturbation is introduced under the form of a short circuit defect that causes a symmetrical voltage amplitude dip on the grid.

On the simulated model, the programmable voltage source is set to change its magnitude at t = 0.4s, 0.7s and 1s, in order to demonstrate the ability of the designed control scheme to maintain the system performances. The corresponding voltage magnitude passes from nominal value Vn to 0.7Vn, then to 1.3Vn, and finally back to the nominal value (Fig.13).

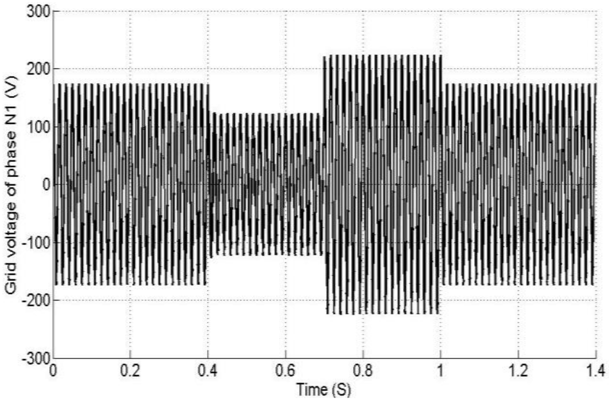


Fig. 13. Grid voltage at transient state

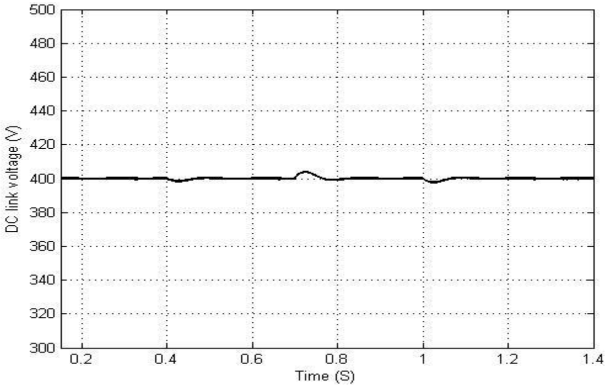


Fig. 14. DC link voltage at transient state

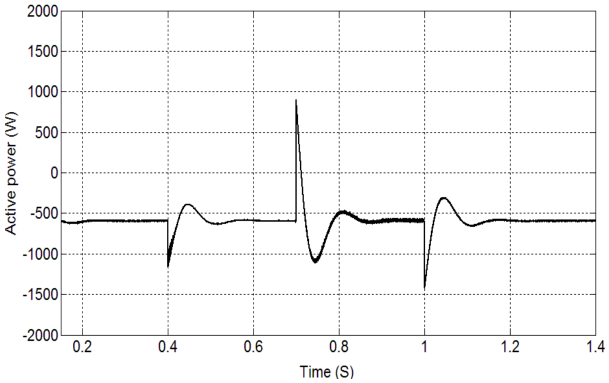


Fig. 15. Active power at transient state

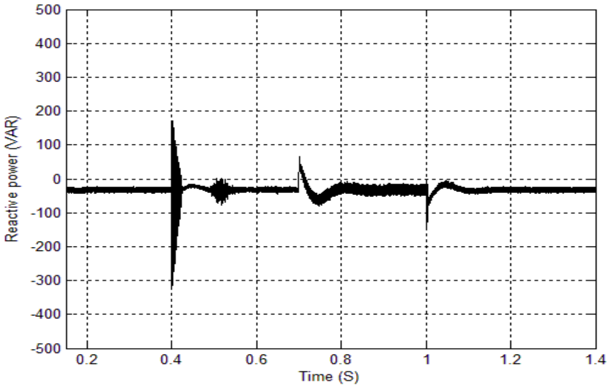


Fig. 16. Reactive power at transient state

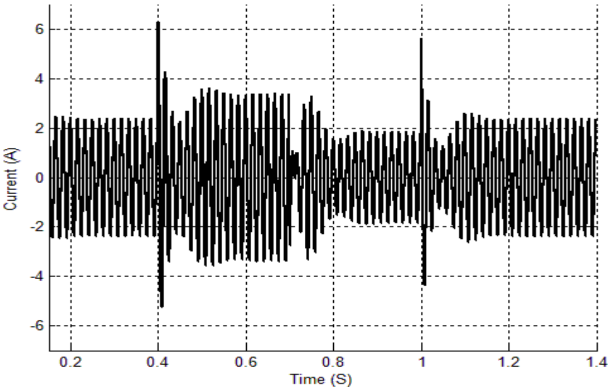


Fig. 17. Grid current at transient state

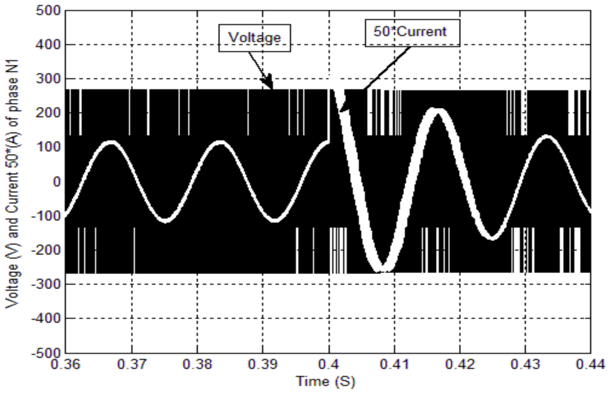


Fig. 18. Grid current and voltage at first transition of voltage

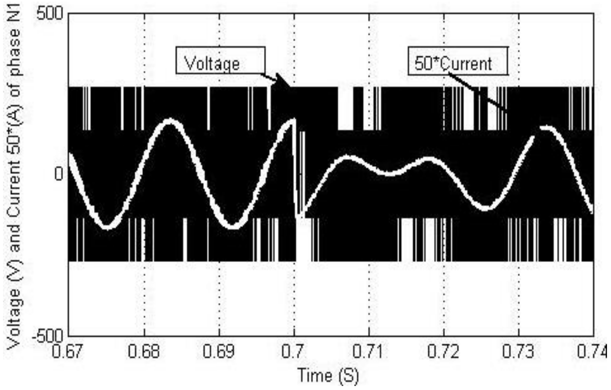


Fig. 19. Grid current and voltage at second transition of voltage

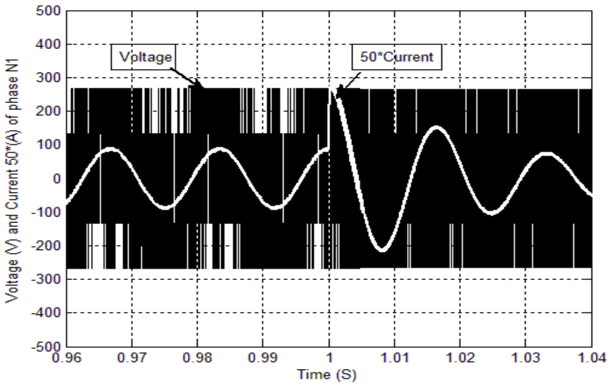


Fig. 20. Grid current and voltage at third transition of voltage

In addition to the voltage variations, the values of inductance and resistance were changed to 50% of their value. The objective in that is to demonstrate the control scheme robustness towards both perturbations: grid fault and parameters identification.

Fig.14 to Fig.20 presents the dynamic behavior of the system for this case. In Fig.14 the DC link voltage appears roughly constant for all perturbations.

Fig.15 shows a more violent transient on the active power compared to the Fig.9 in for the first case. At t=0.7 s, the peak reach 800W which represents a variation of nearly three times the permanent value (-600). Meanwhile, the reactive power shown on Fig.16 reaches -300 VAR at t=0.4s. This value represents 15 times the permanent value of -20 VAR. However, this peak takes just a few milliseconds, which makes it negligible.

The current also reaches peak values during transient states of few milliseconds (Fig.17). These peaks are predicable due to the voltage disturbance. In all transient states, the current gets back to a steady value after 200 ms approximately, which we think is acceptable considering there are not high transient peak values. Zooms on the current are helpful to see the behavior of the current and voltage during the first transient state at 0.4s (Fig.18), during second transient state at 0.7s (Fig.19) and finally during third transient state at 1s (Fig.20).

VIII. Conclusion

A SOSM control scheme for independent control of active and reactive power of a VSC (RSC) has been presented. The proposed SM-DPC scheme directly generate the SV-PWM switching pattern that forces the close loop system to slide along a desired state trajectory, designed to minimize the active and reactive power errors. The control law has also been formulated to ensure stability and alleviate the chattering effect.

To demonstrate the feasibility of the proposed fuzzy-SOSM strategy, two simulation cases were considered: constant and faulted (variable) grid voltage conditions. The simulations results obtained demonstrate the robustness and the good performances in terms of stability and tracking of the reference state trajectory. The pretty well features of the proposed fuzzy SOSM based DPC strategy are as follows

1) No rotating coordinate transformation and angular information of grid voltage are required.

2) Enhanced transient performance

3) The steady-state and transient responses are insensitive to the system parameter’s variations.

4) Chattering-free behavior, a finite reaching time, and robustness with respect to external disturbances (grid)

The contribution of the proposed control to enhance system performances and experimental validation will be the subjects of a coming paper.

References

[1] J. M. Carrasco, L. G. Franquelo, J. T. Bialasiewicz, E. Galvan, R. C. P. Guisado, M. A. M. Prats, J. I. Leon, and N. Moreno-Alfonso, “Power–electronic systems for the grid integration of renewable energy sources: a survey,” *IEEE Trans. Ind. Electron*., vol. 53, no. 4, pp. 1002-1016, Aug. 2006.

[2] S. Janardhanan,, and B. Bandyopadhyay, “Output Feedback Sliding-Control for Uncertain Systems Using Fast Output Sampling Technique,” *IEEE Transactions on Industrial Electronics*, Vol. 53, N°. 5, October 2006.

[3] T. C. Kuo, Y. J. Huang, C. Y. Chen, and C. H. Chang,“Adaptive Sliding Mode Control with PID Tuning for Uncertain Systems,” *Engineering Letters*, Vol.16 N°3, EL_16_3_06, August 2008.

[4] V. I. Utkin, J. G¨uldner, and J. X. Shi, “Sliding Mode Control in Electromechanical Systems,” *FL: CRC Press*, Boca Raton, 1999, pp. 115–130.

[5] Slotin, J.J. and Li,W., Applied non-linear control, Prentice Hall 1991.

[6] V. I. Utkin, “Sliding mode control design principles and applications to electric drives,” *IEEE Trans. Ind. Electron*., vol. 40, no. 1, pp. 23–36, Feb. 1993.

[7] O.A. Morfin, A.G. Loukianov, R. Ruiz, E.N. Sanchez, F.Valenzuela, M. I. Castellanos, “Grid side converter controller applied in wind systems via second order sliding modes,” *8th International Conference on Electrical Engineering Computing Science and Automatic Control* (CCE), Mexico, 2011, pp.1-6.

[8] Yao.08 Yao, Chuanbao Yi, Deng ying, Jiasi Guo and Lina Yang,“The Grid-side PWM Converter of the Wind Power Generation System Based on Fuzzy Sliding Mode Control,” International Conference on Advanced Intelligent Mechatronics, Xiang, China,2008, pp.973–978.

[9] D. Kairous, R. Wamkeue, “Sliding-mode control approach for direct power control of WECS based DFIG,”EEEIC, Italy, 8-11 May 2011. Pp. 1-4, 2011.

[10] B. J. Parvat, B. M. Patre, “Second Order Sliding Mode Controller for Second Order Process with Delay Time,” *International Conference on Industrial Instrumentation and Control (ICIC)*, India. May 28-30, 2015. Pp 280-284.

[11] L. Fridman, and A Levant, “Higher order sliding modes as a natural phenomenon in control theory,” *Ser. Lectures Notes in Control and Information Science, F. Garafalo, and L. Glielmo, Eds*. New York, Springer-Verlag, 1995, Vol. 217, pp. 107-133.

[12] Y. Ren, H. Li, J. Zhou, Z. An, J. Liu, H. Hu, and H. Liu, "Dynamic Performance Analysis of Grid-Connected DFIG Based on Fuzzy Logic Control," *ICMA*, August 9-12, China 2009.

[13] L. A. Zadeh, "Fuzzy Setes,"*Information and Control*, vol. 8, pp.338-353, 1965.

[14] Nabae A., Ogasawara S., Akagi H., A novel control scheme for current controlled PWM inverters, *IEEE Trans. Ind. Applicat.*, 22 (1986), No. 4, 312-323.

Research on Energy Saving Method for IDC CRAC System based on Prediction of Temperature

Zou Yan
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

Xing Jian
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

WU Fei
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

Dong Bo
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai, China

***Abstract*—Amid the information era, energy consumption of IDC Computer Room Air Conditioning (CRAC) system is becoming increasingly serious. Thus there is growing concern over energy saving and consumption reduction. Based on the analysis of the energy saving application of the air conditioning system in the present computer room, a new energy saving method of the IDC CRAC system, which presents energy saving decision based on the prediction of temperature, is proposed. Its principle is the collection of CPU utilization reflected the change of equipment working load, the temperature in hot spots and cold area. Then, to build a BP Neural Network model, taking the working load and the temperature in hot spots for the actual input, taking the temperature in a cold area for actual output. The BP Neural Network model can predict the temperature in hot spots of the next, when a set of real-time data into the model. Choosing a reasonable and effective decision-making scheme of the air conditioning system can realize energy saving control. Preliminary simulation results show, through the establishment of BP network model obtain approximation error of training samples and the prediction error of testing samples, both to highlight the advantages of the model. Finally, the distribution of temperature change about CRAC system whole day obtained by simulation shows that the proposed energy-saving method can reduce the energy consumption of IDC, fully embodies the effect of energy saving.**

Keywords—IDC; CRAC system; BP Neural Network Model; Forecast of temperature; Energy saving

I. Introduction

With the rapid development of information technology and global business, the building and demand of the Internet data center (IDC) computer room are growing and the roll out of the high power density computer and server have raised new challenges on the air conditioning and refrigeration technology. The proportion of power and cooling costs is increasing in the investment of the IT industry, and IDC business boasts high energy consumption. Therefore, how to solve the reasonable running of the IDC CRAC system has become a problem that should be considered in the process of building and using IDC computer room.

A. Features of the Air Conditioning System

The IDC CRAC system mainly serves all types of computers, servers and other equipment which enjoys big Calorific value and high requirement of the humiture in the run, with the following characteristics: 1) with the big cooling load and the small wet load, the indoor air needs to keep humid while removed after heat in the room; 2) all-weather operation, equipment of the room run 24 hours a day. Even in Winters, heat release from the computer room to the outdoor is less than calorific value of equipment, so cool supply is still required [1]; 3) air supply modes of the air conditioning vary, which are mainly under-floor air supplying system and up-supply down-return mode;

In order to ensure normal reliable operation of computers, servers and other high-precision equipment, and the use of IDC room building process must strictly ensure the engine room temperature, humidity and other parameters to obtain precise control. The following provides domestic norms and standards.

TABLE I. DOMESTIC STANDARD "ELECTRONIC INFORMATION SYSTEM ROOM DESIGN SPECIFICATIONS" (GB50174-2008)

Project Level	**Level A**		**Level B**
	Summer	***Winter***	***Annual***
Temperature	23±2℃	20±2℃	18-28℃
Humidity	45-65%		40-70%
Temperature variation rate	<5℃/h And no condensation		<10℃/h And no condensation

Note: The environmental conditions refer to the parameters measured at a cold area or the inlet of the equipment, rather than the average parameter of the room or parameters of air conditioner return air.

Extensive use of air-conditioning equipment guarantees effectively the normal operation of equipment in the computer room, but it is the main equipment of power consumption. Therefore, striking a balance between environment-guarantee and energy consumption in the computer room should be taken as an important issue for energy saving work, especially focusing on air conditioning energy saving [2]. Meanwhile, relevant statistics signify that: while running phases of 24 hours a day for the current IT equipment in the IDC computer room, it is running at a low load of work status some time and at a high load the other. Therefore, it is needed to know the specific working loads of computer room equipment in designing and the effective running time under different conditions.

B. The common energy saving technology of CRAC

Now commonly used energy-saving technologies in the computer room include natural cooling source in the computer room , air conditioning condensers and water cooling system, inverter technology, air-conditioning energy saving additives, CRAC units adaptive control technology, scientific air distribution in CRAC [3] . The natural cooling source uses the outdoor natural cold air to reduce the running duration of the CRAC cooling. Air conditioning condenser and the water-cooling system uses atomized water to realize its cleaning, heat dissipation, cooling and energy saving. Inverter technology adjusts the air conditioning unit operating parameters to save energy, based on the outdoor temperature changes [4]. Air conditioning energy saving additives may work well to prevent oil bubble, oil slick, oil corrosion and other problems caused by long running of the air conditioning. CRAC units adaptive control technology refers to air conditioning operating parameters are set from the manual setting to automatic setting by the computer supervisory control system [5]. Scientific air distribution in CRAC is mainly reflected in the air supply mode of the computer room and rationalizing of air distribution inside the cabinet.

II. Neural Network

A. Artificial Neuron model

Biological neurons are the basic units of the nervous system. It is made up of cell body, dendrite and axon. From the viewpoint of Biological Cybernetics, biological neurons can be used as the basic unit of control and information processing [6]. By abstracting process of biological neurons can be formed an artificial neuron model, shown in Fig.1.

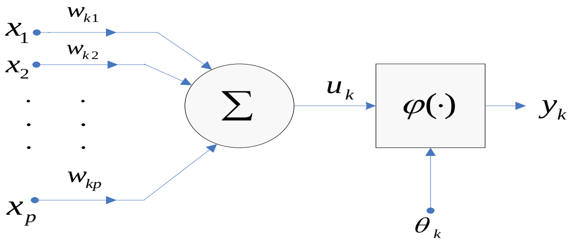


Fig. 1. Artificial Neuron Model

As shown in Figure 1 presents that the artificial neuron model consists of three basic elements, namely, the connection weights, summing unit and activation function. Among them, the connection weights $w_{k1}, w_{k2} \cdots w_{kp}$ corresponding to a biological neuron synapses, the connection strength between each neuron is represented by the weight of the connection weights. Weight values indicate activation of positive representation, inhibition of negative representation. Summing unit means for obtaining a weighted sum of input signals $x_1, x_2 \cdots x_p$. Activation function $\varphi(\cdot)$ plays the role of non-linear mapping, and limits the artificial neural output amplitude to a certain range, generally limited to between (0, 1) or (-1, 1). Activation function generally have several forms, such as a step function, piecewise linear function, Sigmoid type function. In addition, there is a threshold value. At the same time, the role of Figure 1 can be expressed mathematically, as in:

$$u_k = \sum_{j=1}^{p} w_{kj} x_j \, , y_k = \varphi(u_k - \theta_k) \tag{1}$$

B. Artificial Neural Networks ANN

Artificial neural networks ANN is composed of vast artificial neurons connected broad, which can be used to simulate the structure and function of the brain neural system. Artificial neural networks can be viewed as a directed graph, which use artificial neurons as node, connected by directed weighted arcs. In this directed graph, artificial neuron is a simulation of biological neurons, and the weighted arc is simulation of the axon - synapses - dendrites. The weight of directed arc can indicate the strength of interplay in two artificial neurons.

Artificial neural network is an information processing system consists of a large number of interconnected processing units, which have nonlinear, adaptive characteristics [7]. Artificial neural networks have four basic characteristics: 1) Non-linearity, nonlinearity relationship is a common feature of nature; 2) Non-limiting, a neural network is usually composed of a plurality of neurons connected together widely. Overall behavior of a system depends not only on features of a single neuron, but also on the interaction between the cells; 3) Non-qualitative, artificial neural network have adaptive, self-organizing, self-learning ability. The information processed by neural networks have a variety of changes, at the same time, nonlinear dynamical system are changing; 4) Non-convexity, the evolutionary direction of system under certain conditions will depend on a particular state function. In general, artificial neural network model need to consider the topological structure of the networks, the characteristic of neurons, and the learning rules. Depending on the connection method, artificial neural network can be divided into feed forward networks and feedback networks [8].

In the past ten years, the study of artificial neural networks is deeply, also has made considerable progress. Meanwhile, it has successfully resolved many practical problems which difficult to solve of the modern computer in pattern recognition, intelligent robot, automatic control, predictive estimate, and biological, medical, economic and other fields [9].

C. BP Neural Network

BP (Back Propagation) network has been proposed by D.E.Rumelhart and J.L.McClell in 1986, which use the error back-propagation training algorithm [10]. BP network is a multi-layer feed forward network with hidden layer, solves the learning problems of connection weights in the hidden units of multi-network. The BP network structure has shown in Fig.2.

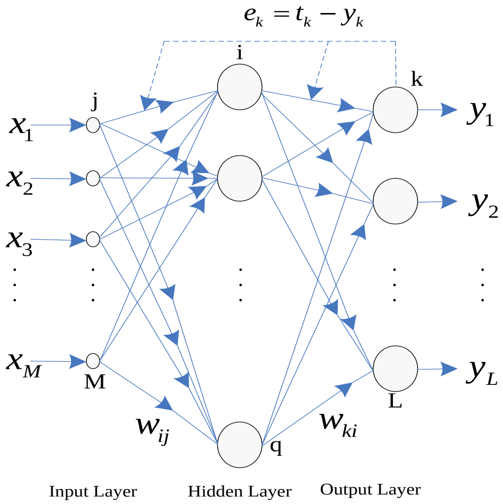


Fig. 2. BP Network Structure

An important feature of neural networks is the ability to acquire knowledge through to learn the environment and improve their performance. So, it can adjust its parameters by using different learning algorithm (such as weight). The basic principle of BP learning algorithm is the gradient steepest descent method; its central idea is that adjusting the weights to make the total error of network becomes a minimum. Also using a gradient search technique, the mean square error between actual output value and expected output value of network is minimized. Network learning is a process that error is spreading backwards while correcting weights.

Figure 2 shows that the network has M input nodes, L output nodes, there are q neurons in hidden layer. Among, x_1, x_2, x_3 ... x_M are the actual inputs, y_1, y_2 ... y_L are the actual outputs, $t_k(k=1,2,...,L)$ are the target outputs, and $e_k(k=1,2,...,L)$ are the output errors.

III. Energy Conservation Design Idea

A. *Overall Implementation Process*

Based on the analysis of the energy saving application of the air conditioning system in the present computer room, a new energy saving method of the IDC CRAC system, which presents energy saving decision based on the prediction of temperature, is proposed. Design details: First, establishing an experimental test environment, under certain circumstances, is able to better reflect the effectiveness of the design. Second, to confirm the distribution division of the equipment working loads. As the design takes into account the temperature changes caused by different equipment working loads, so it is necessary to classify the device specific working conditions. Third, collecting multiple sets of data is analyzed by BP neural network, achieving the prediction of temperature in hot spots. Through the effective control of the CRAC parameters to achieve energy savings effect. Specific energy-efficient room air conditioning system control process shown in Fig.3.

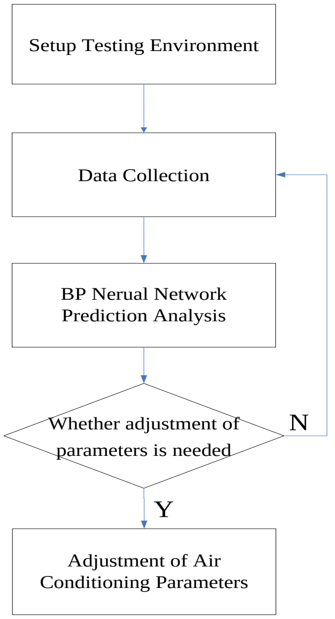

Setup Testing Environment
Data Collection
BP Nerual Network Prediction Analysis
Whether adjustment of parameters is needed
N
Y
Adjustment of Air Conditioning Parameters

Fig. 3. The Control Process of energy saving in the CRAC system

B. *Setup Testing Environment*

There are all types of computers, servers and other equipment within the IDC room. Because of the reasonable use of resources, servers and other equipment are placed in a rack / cabinet [11]. Therefore, there are three type of cooling, namely, 1) the cold air enter into the server from the front of the rack / cabinet, the server waste heat discharged from the rear of the rack / cabinet; 2) the cold air enter into the server from the front of the rack / cabinet, the server exhaust heat discharged from the top of the rack / cabinet; 3) the cold air enter into the server from the front of the rack / cabinet, the server exhaust heat discharged from the back and top of rack / cabinet. In this paper, choose the first mode as shown in Fig.4.

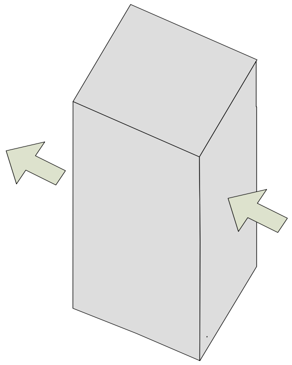


Fig. 4. Type of Cooling

According to the type of cooling, it can put the front area of the regional rack / cabinet known as cold area, the rear area of the regional rack / cabinet known as hot spots. IDC room by the use of the process required to meet the temperature, humidity and other parameters. This paper chooses the domestic standard "electronic information system room design specifications" (GB50174-2008). Control the temperature of cold area within the scope of the class B from 18 ℃ to 28 ℃, and the temperature of the hot spots were measured for judgment basis to adjust the air conditioning outlet temperature. At the same time, combined with changes in operating conditions of the devices within the IDC room, analyzing the temperature changes caused by different working conditions.

C. *The Experimental Data*

1) *Demarcation of Load*

This design need to consider the different temperature variation in the hotspot arise from different conditions of equipment. So the interval division of working condition of equipment is necessary. Generally, heat dissipation of the computer system is mainly because of the electric energy loss, i.e., the electrical power consumption. Combined with the relationship between thermal energy and power distribution, the calculation formula of power consumption used is presented as in:

$$P=K\times C\times F\times VDD^2 \qquad (2)$$

Wherein, P is the dynamic power consumption; K a coefficient; C the load capacitance; F the operating frequency and VDD the operating voltage. Thus it can be seen that power consumption of the circuit is mainly determined by such two variables that the operating frequency and the operating voltage [12]. At the same time, the CPU utilization can fully reflect the operational procedure of relevant equipment at some point in time. The paper takes the network video server for the study which is the special equipment for compression, storage and processing of video and audio data. For this type of server, its load is the task request of different video and audio data. When the load of equipment increase or decrease, its CPU utilization will increase or decrease correspondingly. Thus, the CPU utilization can represent working conditions of equipment with different loads. Taking the operating frequency F as the reference amount and the CPU utilization as the collected volume, the congruent relationship between the two can be analyzed so as to represent the different environmental impacts caused by different working conditions of equipment. The data collected are used to draw the corresponding curve through the SPSS statistical analysis software, taking F for normalization processing. The curve is smoothed to remove random fluctuations, which leads to the one shown in Fig.5.

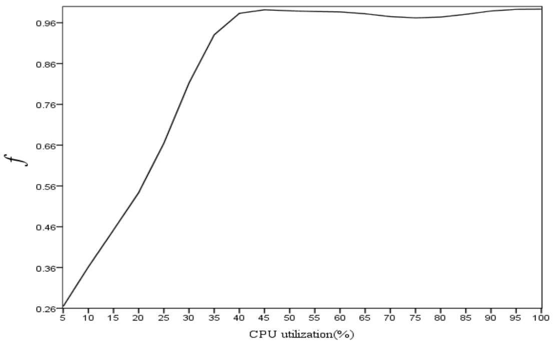


Fig. 5. Corresponding relationship curve between the CPU utilization and the operating frequency

It can be seen from the above curve that when the CPU utilization is in the range of less than 40%, a linear relationship between it and the operating frequency F will be presented, but equipment will work with the highest frequency when the utilization is over 40%. On this basis, working loads can be divided in intervals as shown in Table.2.

TABLE II. DEMARCATION INTERVALS OF LOAD

Working condition NO.	**Percentage distribution of loads**
1	0-10%
2	10-20%
3	20-30%
4	30-40%
5	40-60%
6	60-80%
7	80-100%

2) *Data Collection and Analysis*

In the IDC machine room test environment, the paper needs to collect multiple sets of experimental data. The basic constituent element of each set of data contains the hot zone multi-point temperature, cold area multi-point cold area, and equipment working loads. Among them, the data about cold zone multi-point temperature and equipment working loads collected for the current point in time. The corresponding hot zone multi-point temperature value indicates that the temperature collected in the next moment. Each set of test data combined in this form, therefore, the data elements could be used for the actual input in the BP neural network, which include the cold zone multi-point temperature and the equipment working loads. Also, the hot zone multi-point temperature could be used for the actual output in the BP neural network. Analyze the mapping relation between the input and the output, establishing a forecasting model.

In practice, the cold area temperature and the working load enter this model through real-time data collection, so it can get the predicted value of the hotspot temperatures. In order to adjust the temperature of CRAC, observing the predicted value to meet the requirements of the room temperature that completing the energy saving control of the air conditioning system. In summary, the specific data collection and analysis process shown in Fig.6.

IV. PRELIMINARY PRACTICES

A. *Prediction of Temperature*

1) *Data Preparation*

The simulation experiment can conduct preliminary practices to feasibility of the temperature forecasting and energy saving decisions. Taking the network video server as an example in practice, the sampling frequency is 30 minutes. To find the position of temperature measurement point in the hot and cold zones by determining the initial test time. Collecting multiple sets of data determine the definite value for actual input and actual output in BP neural network. The data form shown in Table.3. In the test data table, the former seven sets of data are used as training samples, and the later two sets of data are used as testing samples of the neural network.

2) *Establish BP Neural Network*

During the design of BP network, there are several aspects need to be considered ,such as the layer number of network, the number of neurons in each layer of network , the activation function and learning rate, etc [13].

a) The Layer and The Number of Neurons of Network

In 1989, Robert Hecht-Nielson had proved that the mapping relation of any continuous function can be approximated by a BP network with a hidden layer. A BP network with s-type hidden layer and a linear output layer approach any continuous function with any closed interval.

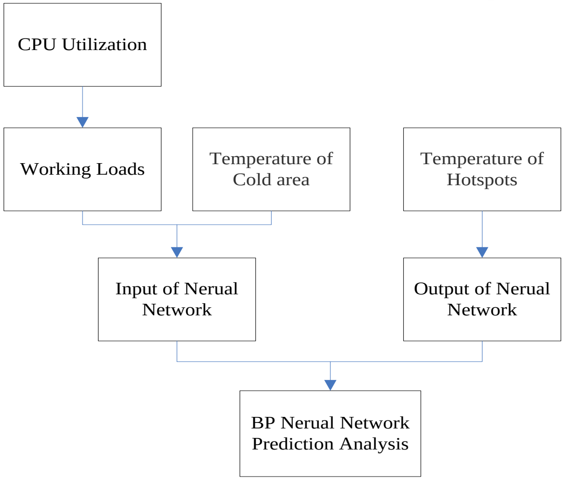


Fig. 6. Data Collection and Analysis

TABLE III. TEST DATA TABLE

Actual Input								**Actual Output**						
Working Load	**The Temperature of Cold Area(℃)**							**The Temperature of Hot Spots (℃)**						
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	y_1	y_2	y_3	y_4	y_5	y_6	y_7
1	21.3	21.5	20.3	20.4	20.9	21.0	20.7	22.0	22.4	21.8	21.4	21.0	21.1	21.9
2	22.0	21.8	20.7	20.9	21.3	21.5	21.1	23.1	22.8	22.5	22.1	22.0	21.7	22.6
3	24.1	23.9	22.1	21.8	23.3	23.0	22.7	25.0	25.2	24.6	24.7	24.0	23.8	24.8
4	24.4	24.7	23.3	23.4	23.7	24.0	23.6	25.8	26.1	25.4	25.1	24.7	25.0	25.3
5	25.6	25.7	24.0	24.4	24.7	25.3	25.0	27.2	27.6	26.8	26.9	26.4	26.7	27.0
6	26.0	25.9	24.8	25.0	25.3	25.6	25.4	30.4	30.1	29.4	29.1	28.8	28.4	29.7
7	27.8	27.5	25.7	25.4	26.3	26.0	25.8	32.5	32.4	31.8	31.4	31.3	30.8	32.0
4	24.6	25.0	23.7	23.2	24.0	24.4	23.9	25.6	26.0	25.3	25.4	25.0	24.8	25.5
6	26.1	26.3	25.0	24.7	25.4	25.6	25.2	30.2	30.5	29.2	29.6	28.7	28.4	29.8

So, a BP network with three-layer accomplishes the map from any n-dimensional space to m-dimensional space [14]. Therefore, increasing the number of hidden layer neurons improve the accuracy of the network.

We should know the number of neurons in input layer and output layer, according the complexity of specific issues.

In order to improve the accuracy of network training, it can use a hidden layer, increasing the number of neurons in the hidden layer, to achieve such design. In the specific design, the number of neurons in the hidden layer is ensured by empirical formula. General empirical formula used to determine the number of neurons in the hidden layer, as in:

$$n_1 = \sqrt{n+m} + a \tag{3}$$

$$n_1 = \log 2^n \tag{4}$$

Wherein, n_1 is the number of neurons in hidden layer, n is the number of neurons in input layer, m is the number of neurons in output layer, a is a constant between 1 and 10. By Table.3, during the design process, n have set for 8, m have set for 7. By using the equation 3, the paper have selected 5 for the initial number of neurons in the hidden layer, and then training the network to find the right number of neurons in the final.

b) The Learning of BP network

After determining the structure of BP network, it needs to learn and correct threshold value and weight value of the network, in order to realize the mapping relationship between the input and the output [15]. The main contents of this article are that to establish and implement the neural network forecasting model about temperature of the hot zone. Mainly, in the simulation the paper have used the neural network toolbox offered by MATLAB software to program the model.

For the content of this paper, select the improved elastic gradient descent algorithm. In the elastic gradient descent algorithm, when occurs the oscillation of the training, the variation of the weight will reduce. After several iterations, the variation of the weight will increase when the weight change in one direction. A large number of practical applications proved that the elastic gradient descent algorithm is very effective. In MATLAB neural network toolbox, the training function of elastic gradient descent algorithm is the Function trainrp.

3) Analysis

During the network training, in order to ensure the data for the same order of magnitude, firstly, it have pretreated the data of input and output in the neural network. So it can speed up the training of the network. The transfer function established in the neural network model is based on the S-type function, its input range as [0, 1] is best [16]. Therefore, the algorithms used in the paper should take the data normalized to [0, 1].Suppose $X_{\max}$ is the maximum value of the original data elements in the same column of the table, $X_{\min}$ is the minimum value, X and X_i are the normalized data before and after, and the normalization function , as in:

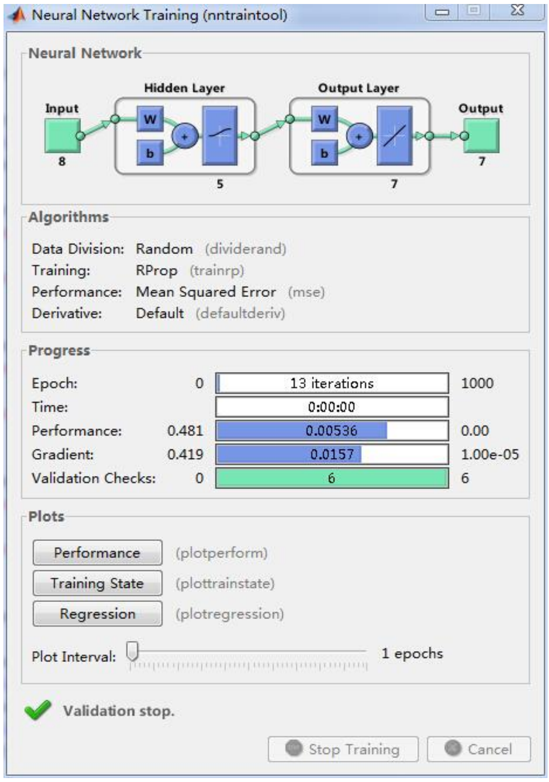
$$X_i = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \qquad (5)$$

In summary, there is a BP neural network model that could predict the temperature of hot zone in IDC room. Using the MATLAB software conducted the simulation experiments of the design. BP network training effect can be obtained by simulation, as shown in Fig.7.

Through several simulation experiments, to compare multiple sets of training data, this study found that for the purpose of BP network needs to have a S-type hidden layer and a linear output layer, the hidden layer contains five neurons, the linear output layer contains seven neurons. Meanwhile, the training function of network is trainrp. After the completion of training for the network, we need to model the training samples approximation error, as shown in Fig.8. In order to test the effect of the model, using the test samples in the table to test the model that calculate the prediction error of the output, as shown in Fig.9.

As shown in Figure 8 and Figure 9, seven groups approximation error value of training samples can be kept at 0.2 or less, and even 5-set error values below 0.1, indicating that BP network model better reflects the direct relationship between input and output . Meanwhile, two groups prediction error value of test sample can be kept between 0.1235 and 0.1240.

Through the establishment, training and testing of the above-mentioned BP network model, when there are sets of data of equipment working load and sets of data of cold zone temperature as input data into the model that can predict the temperature of hot spots in next time. The next time improve energy conservation control of CRAC systems.



Neural Network Training (nntraintool)

Neural Network

Algorithms

Data Division: Random (dividerand)
Training: RProp (trainrp)
Performance: Mean Squared Error (mse)
Derivative: Default (defaultderiv)

Progress

Epoch:	0	13 iterations	1000
Time:		0:00:00	
Performance:	0.481	0.00536	0.00
Gradient:	0.419	0.0157	1.00e-05
Validation Checks:	0	6	6

Plots

(plotperform)
(plottrainstate)
(plotregression)

Plot Interval: 1 epochs

✓ Validation stop.

Fig. 7. Training effect of BP network

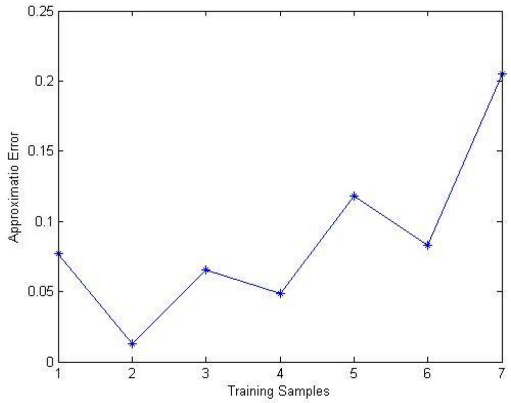


Fig. 8. Approximation Error of Training Samples

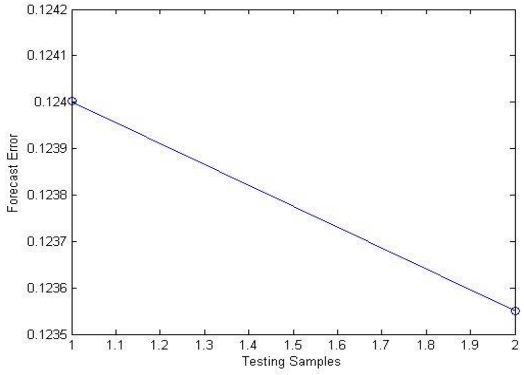


Fig. 9. Forecast Error of Testing Samples

B. *Energy Saving*

In order to reflect the energy saving of paper design, this article assumed two cases: 1 .Throughout the day, to keep the air temperature of CRAC unchanged at 20 ℃; 2. Combining the BP network model, the output data of the temperature of hot spots predicted by the model adjust the air temperature of CRAC.

According to the domestic selection criteria GB50174-2008, firstly, the air temperature of CRAC was adjusted to 28 ℃, and then collected the data of working load and cold zone temperature to predict hot spots temperature in next time, finally achieve energy efficiency goals. By means of software, it can map out the changes in air temperature throughout the day, as shown in Fig.10.

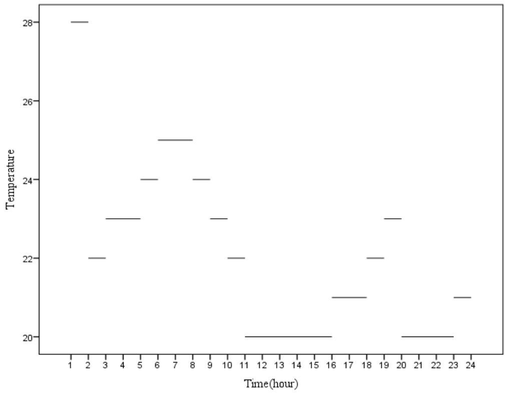


Fig. 10. Temperature of CRAC System

From Figure 10, the air temperature is constantly changing. Combining different energy consumption caused by different air conditioning temperature, to some extent, the design scheme proposed in this paper can reduce the energy consumption of the cost of IDC.

V. CONCLUSION

The energy consumption of the IDC CARAC system accounts for a large proportion in that of the computer room, so it is needed to reconstruct energy saving of the CRAC system. From the perspective of service objects of the air conditioning system, temperature changes are primarily caused by the all-day changing working conditions of computer room equipment. Understanding and grasp to the running status of equipment can produce a more effective energy saving method to achieve the goal by adjusting air conditioning operating parameters.

In this paper, CPU utilization for the reference data to represent the working conditions of the equipment, while collecting multiple sets of real-time temperature of hot and cold zones. The equipment working load and the cold zone temperature act as the actual input and the hot zone temperature act as the actual output establish a BP network model. Through a set of real-time input data, it predicted the temperature of the hot zone the next time that as a data basis to adjust air temperature of CRAC systems. Feasibility of the research method proposed is analyzed by simulated experiment and results show that the BP network model can well determine the mapping relationship between input and output. Also, the value of approximation error and prediction error are able to show that the design proposed method has achieved the purpose of the experiment better.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No.61272097, 61405045), the Shanghai Municipal Committee of Science and Technology Project (No.13510501400), the Shanghai Municipal Education Commission, A Key Innovation Project (No.12ZZ182).

REFERENCES

[1] FanQiang, "Air Conditioning System Design for Large Data Rooms," Heating Ventilating & Air Conditioning, Vol.12, No.2, 2013, pp.33-36.

[2] Qian Xiaodong, Lizheng, "Research on Energy Saving of Air Conditioning System in Data Center," Heating Ventilating & Air Conditioning, Vol.9, No.3,2012, pp.91-96.

[3] GaoRui, "Introduction to the Air Conditioning System Energy Saving from the Engineering Example," Cryogenics and Superconductivity, Vol.4, No.6, 2014, pp.48-53.

[4] Zhang Guangli, Chen Liping, "Operational Energy Optimization of Refrigerating Station," Fluid Machinery, Vol.12, No.8, 2012, pp.75-80.

[5] Lil, "Energy Saving and Temperature Control Method of Adaptive Control for Air Conditioning System of Communication Room, " Logistics Technology, Vol.5, No.5,2013, pp.323-326.

[6] GuDi, ZhouJun, "The research is based on the MATLAB about neural network forecast model," Logistics technology, Vol.125, No.2, 2006, pp. 125-128.

[7] Ping Sun Leung, L.T.T . "Predicting shrimp disease occurrence: artificial neural networks vs. logistic regression ," Vol.12, No.5, 1999,pp.15-24.

[8] G..Thimm, P. Moerland, E. Fiesler, "The interchangeability of learning rate and gain in back propagation neural networks,"

[9] Neural Computation, Vol.8,No.2, 1996,pp.451-460.

[10] Adachi M, Kotani M. "Identification of Chaotic Dynamical Systems with Back-Propagation Neural Networks," IEICE trans, Vol.12, No.6,1994,pp.324-334.

[11] Qiquan Li, Changquan Wang, " The spatial distribution simulation method of China surface soil organic matter based on neural network model ." Earth Science Progress, Vol.8, No.2, 2012, pp.20-24.

[12] Shuzhen Wei, Feifeng Wang, "Natural grassland classification based on SOFM network," Journal of grass industry, Vol.4, No.1, 2011, pp.52-57.

[13] Wan Suhai, Shaokun, Liu Zongtian, " Dynamic Power Management Scheme Based on Linux Dynamic Frequency Scaling," Computer Engineering, Vol.5, No.10, 2011, pp.238-239.

[14] Burcue, Tulay.Y, "Improving classification performance of so-Nar targets by applying general regression neural network with PCA," Expert Systems with Applications, Vol.2, No.7, 2008, pp.28-33.

[15] Parojcic. J, Ibric .S, Djuric. Z, "An investigation into the usefulness of generalized regression neural network analysis in the development of level A in vitro-in vivo correlation," European Journal of Pharmaceutical Sciences . Vol.3, No.11, 2007, pp. 38-42.

[16] Qingle Pang, "A Rough Set-Based Neural Network Load Forecasting Algorithm and Its Application in Short-Term Load Forecasting," The grid technology,Vol.3,No.4,2010,pp.26-30.

[17] Moody.J, C.J.Darken, "Fast Leaning in Networks of Locally-turned Processing Units," Neural Computation, Vol.25, No.13, 2012, pp.281-294.

Bio-Inspired Clustering of Complex Products Structure based on DSM

Fan Yang
[1]Institute of Systems Science and Engineering
Wuhan University of Technology
[2]College of Computer
Hubei University of Education
Wuhan, China

Pan Wang*, Sihai Guo*, Qibing Lu
Institute of Systems Science and Engineering
Wuhan University of Technology
Wuhan, China

Xingxing Liu
School of Management
Wuhan University of Technology
Wuhan, China

***Abstract*—Clustering plays an important role in the decomposition of complex products structure. Different clustering algorithms may achieve different effects of the decomposition. This paper aims to proposes a bio-inspired genetic algorithm that is implemented based on its reliable fitness function and design structure matrix (DSM) for clustering analysis of complex products. This new bio-inspired genetic algorithm captures the features of DSM, which is base on the biological evolution theory. Examples of these products include motorcycle engines that are presented for clustering. The five cluster alternatives are obtained from the regular clustering algorithm and the bio-inspired genetic algorithm, while the best cluster alternative comes from the bio-inspired genetic algorithm. The results show that this algorithm is well adaptable, especially when the product elements have complicated and asymmetric connections.**

***Keywords*—Bio-inspired computation; genetic algorithm (GA); reliable fitness function; DSM; complex products; clustering**

I. INTRODUCTION

In this case, complex products are those whose structure is more complicated and whose module partition is scattered. There are many methods for the designing the physical structure of complex products, including the DSM method [1], the Hatley/Pirbha method [2], the stochastic block model [3], and biclustering in machine learning. With fierce competition in the manufacturing industry, however, the issue of physical structure is a problem worthy of concern. The clustering methods for the physical structure of complex products are thus a key consideration for core competition in business.

In recent years, studies on the function module partition have been frequently discussed. The goal is to construct a symmetric matrix, transformed from attributes of many products [4, 5]. Then, the scholars create new methods to deal with this matrix, in order to gather similar elements for optimized clustering. Erixon et al. [6] propose a module partition method based on impact factors like technological innovation, planning, parameters, and style. Gu et al. [7] propose a product modularization for life cycle engineering, a method of constructing correlation matrixes for the stage targets. At this time, the design professors then select the better module partition result by comparing the results from different design objectives. Salhieh et al. [8] use design for modularity by the Fuzzy Theory, a quantitative description and analysis algorithm. Stone et al. [9] use a heuristic method for identifying modules for product architectures based on energy flow, material flow, and signal flow in architecture charts. Gao et al. [10] identify functional modules using generalized directed graphs. Chen et al. [11] propose the dynamic cluster analysis of a fuzzy equivalence matrix to be applied to a dynamic partition function module.

The above studies offer better routes for structuring partitions for new products, but few of them consider the physical structure or the asymmetric connections between product components. DSM, first proposed by Dr. Steward, could solve these problems partly [1]. It consists of a matrix with homogenous rows and columns, and the value of each cell is not a random probability. Thus, it is different from the matrix in the Stochastic Block Model or the biclustering model. Many studies concentrate on how to cluster and evaluate the module partition based on DSM [12]. For the efficiency of clustering, some studies use GA for identifying modules. Tseng et al. [13] use a grouping genetic algorithm for clustering the components. Crossover mechanisms are modified according to the need of modular design by a reasonable evaluation. Zhou et al. [14] propose a function module partition method for innovative product design in order to meet customers'diverse requirements and to shorten design and assembly time. Liu et al. [15] establish a method of decomposition and clustering of the product architecture performed by GA. Their method is based on a building the product with DSM.

With the help of literatures [11, 15, 16], in this paper, a bio-inspired genetic algorithm is proposed for a components cluster based on DSM. This algorithm uses a more reasonable fitness function and an asymmetric matrix. Compared with a regular cluster method, the bio-inspired genetic algorithm is suitable for a cluster optimized for complex product architecture.

II. DSM IMPLEMENTATION AND EVALUATION

The aim of DSM is to decompose elements of a product first, and then divide the elements into different clusters. It is possible that all the elements may cluster into one, determined by the connection degree between elements. Finally, the best one can be selected by several different cluster alternatives through an evaluation function.

*Corresponding authors.

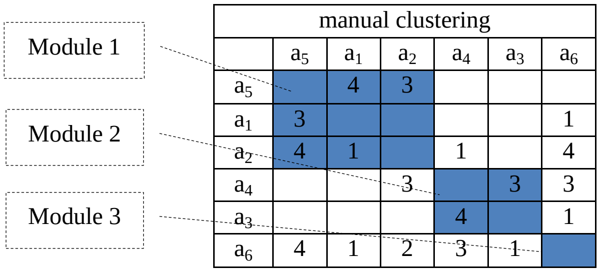
A. Physical DSM

DSM is a product engineering matrix tool, which has four applications: components clustering, team design, task assign and action design [16]. The first one is used here. Each element in DSM corresponds to each component of the product. Cells indicate the connection among elements, except those that are diagonal. Empty cells mean no connection between corresponding elements, while cells with high value indicate high connections intensity. As shown in figure 1, six components as (a_1, a_2, a_3, a_4, a_5, a_6) compose product A. The value of the connection degree between a_1 and a_3 is two. Product A can be divided into three modules using manual classification. Module 1 includes a_1, a_2, and a_5. Module 2 includes a_3 and a_5. Module 3 includes only a_6.

The relations of space, energy, information, or material could be considered to confirm the value of the connection degree with the weighted average. A larger connection degree shows that a closer relationship between the two elements.

initial model						
	a_1	a_2	a_3	a_4	a_5	a_6
a_1	■				3	1
a_2		■		1	4	4
a_3			■	4		1
a_4		3	3	■		3
a_5	4	3			■	
a_6	1	2	1	3	4	■

(a)



manual clustering						
	a_5	a_1	a_2	a_4	a_3	a_6
a_5	▨	4	3			
a_1	3	▨	▨			1
a_2	4	1	▨	1		4
a_4			3	▨	3	3
a_3				4	▨	1
a_6	4	1	2	3	1	▨

(b)

Fig. 1. DSM of product A

B. Clustering and evaluation function

The evaluation influences the clustering process. The common simple clustering algorithm is based on specific evaluation functions, which are performed using software tools like SPSS and Matlab. In these tools, the clustering methods are almost the same, meaning that it is necessary to generate a symmetric matrix. We also must complete clustering with evaluation methods using maximum or minimum distance as a parameter. This matrix is similar to the DSM mentioned above, with the exception of diagonal symmetry. The internal concentration and the external linkage degree are key parameters to the cluster alternatives. They can be calculated by asymmetrical DSM and the related formula after elements clustering. Then, the cluster alternative with the higher internal concentration and the lower external linkage degree is considered to be the better one. In order to obtain a reasonable evaluation model, the equation for the internal concentration and the external linkage degree are set with some assumptions.

The elements in DSM representing components of a product often convert into one or more modules after clustering. The assumptions are as follows:

1) Assume a product is comprised of m components. The number of components is the same as the elements of DSM. The DSM is m×m matrix. And d_{ij} is the value at the ith row, jth column (expressed cell (i,j)) in DSM where i, j ∈(0,m). It represents the value of connection degree.

2) The individual element means it is the only element in a module where the internal concentration is 0.

3) Overlapping elements are allowed, meaning that one element may belong to two or more modules.

4) Unless correlation occurs, clusters decompose. We maximize the total internal concentration value and minimize the total external linkages.

5) Assume g_d as the internal concentration of the dth module. For the purpose of system optimization, g_d is in direct proportion to the total value of cells in the dth module; while it is inversely proportional to the number of connections inside the dth module. A higher total value shows strong cohesion, while the more connections may disperse the concentration. The formula for the internal concentration of the dth module is as follows:

$$g_d = \begin{cases} 0, & k_d = 1 \\ \dfrac{1}{k_d(k_d-1)}\Big(\sum\limits_{d=1,i,j\in M_d} d_{ij}\Big), & k_d > 1 \end{cases} \tag{1}$$

Assume the total number of modules is D; k_d is the number of elements in dth module. $k_d = 1$ means the case of the only element in a module where the internal concentration is 0, namely the "individual element". The total value of internal concentration of system G is given by:

$$G = \sum_{d=1}^{D} g_d \tag{2}$$

6) Assume r_{db} as the external linkages between dth and bth modules. Also, for the purpose of optimizing the system, r_{db} is in direct proportion to the total value of cells between the dth and the bth modules. It means that the higher the total value, the tighter the connection between the two modules. The formula for the external linkages between the dth and the bth modules is as follows:

$$r_{db} = \sum_{i\in d,\, j\in b} d_{ij} \tag{3}$$

The total value of external linkages of system R is as follows:

$$R = \sum_{d=1}^{D-1} \sum_{b=d+1}^{D} r_{db} \tag{4}$$

7) According to the explanation of G and R, it is obvious that a higher G and a smaller R means a better clustering effect. Thus, the evaluation value can be set as follows:

$$F = G - R \tag{5}$$

III. The Bio-Inspired Genetic Algorithm

A. Fitness function

When a product is not complex, it can cluster the DSM with manual methods. The optimal alternative can then be selected with the above evaluation model from the set of alternatives. However, in the case that there are too many components, connections between components are made complicated, and handling it manually is no longer suitable. In this section, the GA with an optimized fitness function is presented to select the clustered alternatives.

Let's take product A as an example. One of its chromosomes and the clustered alternative is shown in Table 1. It can be described as a 01-matrix. When an element belongs to a cluster, the gene value is 1, others are 0. The elements in one column should not all be 0, meaning an element must belong to one or more cluster(s). The length of a chromosome is $D \times m$，D⩽m/2.

TABLE I. The Clustering of Modules and the Encoding of a Chromosome

Encoding of a chromosome						
	1	2	3	4	5	6
Module 1						
Module 2						
Module 3						

The chromosome 01-matrix can be presented as:

$$rst = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{6}$$

The corresponding clustering alternative is:

$$A = \{[1,2,5],[3,4],[6]\} \tag{7}$$

Based on formulas (1) to (5), the fitness function of every chromosome can be designed as:

$$Fit = \frac{F(h) - F_{\min}}{F_{\max} - F_{\min}} \tag{8}$$

Where $F(h)$ is the evaluation value of the hth chromosome, F_{max} and F_{min} is the maximum and the minimum respectively. Then, the selection probability is:

$$p(h) = \frac{F(h)}{\sum_{h=1}^{n} F(h)} \tag{9}$$

B. The flow of the bio-inspired algorithm

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

The parameters, specifically chromosome size *n*, crossover, and mutation probability, of the bio-inspired genetic algorithm are fixed a priori. Based on the fitness function and the selection probability above, the steps of the GA are as follows:

Step 1: Generate randomly *n* chromosomes, actually 01-matrixes by the requirement above;

Step 2: Generate corresponding clustered alternatives from chromosomes;

Step 3: Calculate the evaluation values of the alternations based on the evaluation function;

Step 4: Calculate the fitness function values and the selection probabilities based on the evaluation values;

Step 5: Select the top m chromosomes and eliminate the bad ones, and adds new chromosomes so as to the total number is *n*;

Step 6: Realize the crossover and mutation with the 2-dimension multiple-points crossover and 2-dimension basic mutation [16] (the suitable values may be obtained by trial-and-error.).

Step 7: Judge the terminated condition — if the optimal one in one generation keeps constantly evolves, stop; for others, return to step 2.

IV. Case Study

A. Construction of the DSM

For example, a motorcycle engine with seventeen components would create a DSM matrix like Table 2 by experts' evaluation and the AHP method.

TABLE II. DSM Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
2	2	0	0	0	0	2	2	0	0	0	0	0	4	0	4	4	4
3	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0
5	0	2	2	0	0	0	0	0	4	0	0	0	0	0	0	0	0
6	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0

8	0	0	0	0	0	0	1	0	4	0	0	0	0	0	0	0	0
9	0	0	0	0	4	0	2	4	0	0	0	0	0	0	0	0	0
10	0	2	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
12	0	2	0	0	0	0	0	0	0	4	2	0	0	0	0	0	0
13	0	4	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0
14	0	4	0	0	0	0	0	0	0	0	0	0	4	0	4	0	0
15	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	4	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
17	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

B. *Regular clustering*

Clustering it with MATLAB, firstly the asymmetric matrix is transferred into the symmetric one, taking the means of d_{ij} and d_{ji} as the new d_{ij} and d_{ji}. The following results are obtained by adopting the clustering algorithm of the MATLAB (for better effectiveness, the element 2 and 6 are eliminated, and regarded as independent elements).

Then several clustering alternatives can be obtained：

A_1={[1,17],[2],[3,4,5],[6],[5,7,8,9],[10,11,12],[13,14,15,16]}

A_2={[1,17],[2],[3,4,5,7,8,9],[6],[10,11,12],[13,14,15,16]}

A_3={[1,17],[2],[3,4,5],[6],[7,8,9],[10,11,12],[13,14,15,16]}

A_4={[1,17],[2],[3,4],[6],[5,7,8,9],[10,11,12],[13,14,15,16]}

C. *Genetic clustering*

Considering the data in Table 2, the genetic algorithm is adopted for clustering. When compared with the randomness of the original chromosomes, the results obtained by running the algorithm are quite different. The optimal one (shown as in Fig. 2) is obtained from the many iterations.

In Fig.2, the x-coordinate represents the number of iterations and the y-coordinate means the optimal value of each generation and the global optimum is gotten in the early evolution. From $A_{3,}$ in Table 3, elements 3 and 5 are shown as independent, while element 1 is overlapping.

A_5={[4,10,11,12],[8],[3],[5,9],[1,17],[1,2,6,7,13,14,15,16]}

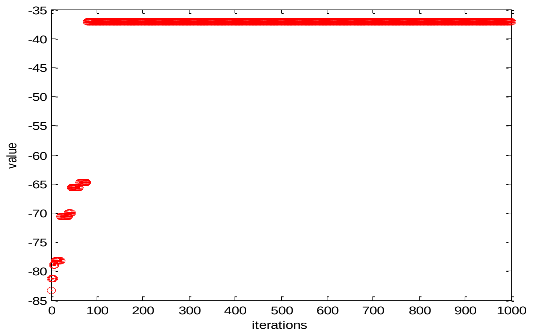


Fig. 2. Genetic clustering

TABLE III. COMPREHENSIVE EVALUATION ALTERNATIVES

	cluster alternatives	Value
A_1	[1,17],[2],[3,4,5],[6],[5,7,8,9],[10,11,12],[13,14,15,16]	-63.8333
A_2	[1,17],[2],[3,4,5,7,8,9],[6],[10,11,12],[13,14,15,16]	-49.9500
A_3	[1,17],[2],[3,4,5],[6],[7,8,9],[10,11,12],[13,14,15,16]	-55.4167
A_4	[1,17],[2],[3,4],[6],[5,7,8,9],[10,11,12],[13,14,15,16]	-54.1667
A_5	[4,10,11,12],[8],[3],[5,9],[1,17],[1,2,6,7,13,14,15,16]	-37.0536

Table 3 shows the values of the five comprehensive evaluation alternatives. Among them, the result of the genetic clustering is the best. The genetic clustering not only accelerates optimization, but also obtains a high quality. Moreover, alternatives 2, 3, 4 are better than alternative 1, showing that element 5 should not belong to two clusters.

V. CONCLUSIONS

In this paper, the bio-inspired algorithm with a reliable fitness function and DSM for complex products is proposed. This genetic algorithm can achieve the better effect for decomposing complex products. Comparing the results of the regular clustering and this genetic algorithm, the basic conclusions are as follows: (1) the regular clustering method has a limited range of applications; it cannot adapt to the physical structure clustering of complex products; especially, in the case of the connection matrix being asymmetric, the optimal cluster alternative cannot be obtained by the regular clustering; (2) it is significant that this genetic algorithm is designed with the fitness function and evaluation function for clustering; (3) Introducing of the overlapping elements shows the better effect. Meanwhile, the independent elements may appear in some situation. However this genetic algorithm is a lack of considering efficiency. There may be other methods to improve these clustering results, so the algorithm can be further improved.

ACKNOWLEDGMENT

This work was supported in part by the Chinese National Natural Science Foundation (No. 71371148); China postdoctoral Fund Projects (No. 2013 M542080); China postdoctoral Fund Special projects (No.2014T70750).

REFERENCES

[1] R. ZhiJun, L. Ming, D. Binbin, C. Kuisheng, “Analysis of clustering hierarchy for product design structure,” Journal of Wuhan University of Science and Technology, vol. 38, pp. 190-196, June 2015.

[2] A. Tarek, E. Hoda, “Optimum granularity level of modular product design architecture,” CIRP Annals - Manufacturing Technology. vol. 62, pp. 151-154, April 2013.

[3] J. Pandremenos, G. Chryssolouris, “A neural network approach for the development of modular product architectures,” International Journal of Computer Integrated Manufacturing, vol. 24, pp. 879-887, September 2011.

[4] B. Xiaoyong, Z. Tianxu, Z. Xiaolong, Y. LuXin, L. Bo, “Clustering-based extraction of near border data samples for remote sensing image classification,” Cognitive Computation, vol. 5, pp. 19-31, March 2013.

[5] M. K. Parag, S. J. Tim, H. L. Robin, H. M. John., “Clustering of gaze during dynamic scene viewing is Predicted by Motion,” Cognitive Computation, vol. 3, pp. 5-24, March 2011.

[6] G. Exixon, A. V.Yxkull, M. A. Arnstr, "Modularity-the basis for product and factory reengineering," Annals of the CIRP, vol. 45, pp. 1-6, July 1996.

[7] P.Gu P, S.Sosale, "Product modularization for life cycle engineering," Robotics and Computer Integrated Manufacturing, vol. 15, pp. 387-401, October 1999.

[8] S. M. Salhieh, A. K. Kamrani, "Macro level product development using design for modularity," Robotics and Computer Integrated Manufacturing, vol. 15, pp. 319-329, August 1999.

[9] R. B. Stone, K. L. Wood, R. H. Crawford, "A heuristic method for identifying modules for product architectures," Design Studies, vol. 21, pp. 5-31, January 2000.

[10] G. Fei, X. Gang, T. Simpson, "Identifying functional modules using generalized directed graphs: definition and application," Computer in Industry, vol. 61, pp. 260-269, April 2010.

[11] C. Jiwen, Z. Jinsheng, W. Zhi, H. Bo, W. Fakai, "Research on function module dynamic partition for product innovation Design," The Chinese mechanical engineering, vol. 24, pp. 251-256, January 2013.

[12] L. Jiangang, T. Dunbing, Y. Chun, H. Xiangdong, J. Taotao, W. Ningsheng, "Clustering and evaluating of product architecture based on physical DSM and row-column transform," Systems Engineering and Electronics, vol. 30, pp. 1904-1908, October 2008.

[13] T. Hwaien, C. Chienchen, L. Jiadiann, "Modular design to support green life-cycle engineering," Expert Systems with Application, vol. 34, pp. 2524-2537, May 2008.

[14] Z. Kaijun, L. Dongbo, T. Yifei, "Function module partition method for product innovative design," Journal of Computer Aided Design & Computer Graphics, vol. 19, pp. 73-83, January 2007.

[15] L. Jiangang, H. Xiangdong, W. Ningsheng, Q. Xiaoming, M. An, "Clustering and reorganization of product architectures based on genetic algorithm," Mechanical Science and Technology, vol. 25, pp. 1318-1337, November 2006.

[16] T. Guobing, Q. Xiaoming, L. Jiangang, Product design and development based on design structure matrix (DSM), Bei Jing: Science press, 2009.

Contemplation of Effective Security Measures in Access Management from Adoptability Perspective

Tehseen Mehraj
Dept. of Computer Science & Engineering
Islamic University of Science and Technology
Awantipora, Kashmir

Bisma Rasool
Dept. of Computer Science & Engineering
Islamic University of Science and Technology
Awantipora, Kashmir

Burhan Ul Islam Khan
Dept. of Computer Science & Engineering
Islamic University of Science and Technology
Awantipora, Kashmir

Asifa Baba
Dept. of Electronics & Communication Engineering
Islamic University of Science and Technology
Awantipora, Kashmir

Prof. A. G. Lone
Dept. of Computer Science & Engineering
Islamic University of Science and Technology
Awantipora, Kashmir

***Abstract*—With the extension in computer networks, there has been a drastic change in the disposition of network security. Security has always been a major concern of any organization as it involves mechanisms to ensure reliable access. In the present era of global electronic connectivity where hackers, eavesdroppers, electronic frauds, viruses are growing in number, security has proved to be indispensable. Although numerous solutions have been put forth in literature to guarantee security, they have botched with related traits like efficiency and scalability. Despite the range of security solutions that have been presented by experts, not a single approach has been wholly agreed upon to provide absolute security or standardized upon unanimously. Furthermore, these approaches lack adoptable user implementation. In this paper, various approaches and techniques that were introduced in the past for the purpose of enhancing the authentication and authorization of the user while performing sensitive and confidential transaction are discussed. Each of works performed by the researchers is taken single-handedly for debate followed by analysis. Finally, the open issues in the current domain of study are formulated from the review conducted.**

Keywords—Access Management; Authentication; One-Time-Password (OTP); OTP generation; User adoptability

I. Introduction

Access management includes diverse ways a user can gain access to an identity and the various technologies that offer protection to that identity. Moreover, the access to an identity by the right user at the right time is also taken care of in Access Management [1]. Access management in public and private networks is a serious concern in the field of Information Technology. The exponential growth in computer networks has led to the emergence of a range of security issues [2]. As information happens to be more substantial and a large figure of people joining the Internet, the security of information about various sectors of society has become vital [3]. Although, security is a serious concern in the present day scenario in sectors like educational institutions, governmental applications, financial institutions, military organization, etc., it needs to be ensured that the security measures should be user ergonomic and adoptable. Therefore, enhanced security measures should be taken for safeguarding computer networks while operating network applications [4]. Further, there is a need for performing secure commercial and business transactions that make access management a top security requirement that involves authentication and authorization [3]. Moreover, a security system relies on Access Management that is found to be the first checkpoint in network security [5].

A. Authentication and Authorization:

In IT security, Authentication is the essential building block as well as the crucial line of defense [6]. Authentication is used as a fundamental technology that is the process of validating the identity of someone or something. On the other hand, authorization is the process of determining whether the user has permission to access, read, insert, delete or modify certain data, or to execute certain programs [3][7]. In general, authentication necessitates the user to present the credentials to prove what he claims to be. Conventionally, the various authentication mechanisms have been categorized as:

1) Something the individual knows, e.g., a password, a Personal Identification Number (PIN), a passphrase

2) Something the individual possesses, e.g., electronic key cards, physical keys, smart cards, cryptographic keys

3) Something the individual is (static biometrics), e.g., recognition by Iris pattern face, fingerprint and other biometrics

4) Something the individual does (dynamic biometrics), e.g., recognition by handwriting characteristics, typing rhythm and voice pattern.

The last couple of years has brought news of a variety of distinguished security breaches focused on authentication, in some cases, severe consequences. A not uncommon pattern could be a revelation that some server is hacked, and an outsized variety of account passwords have been probably exposed. As a result of, whereas, we tend to know files containing things like countersign hashes were traced, there's typically no succeeding data on the actual fallacious use of the information or real harm done. An example of a security breach where the damage resulted is that the attack on the Associated Press Twitter account of Apr 2013 [8]. A counterfeit tweet about explosions at the White House caused a quick, but serious, disruption to the monetary markets. The business is slowly reacting to password attacks and is commencing to try and notice higher ways to stop them.

In the current era of security modernization, password-based authentication i.e. single factor authentication system still finds a use for the processes of authentication and authorization. As passwords form the utmost level of sensitive and confidential information, consequently unauthorized access of user-sensitive password in large scale networking system has been exceedingly studied in the past research works [9][10][11][12]. One of the momentous issues with traditional password-based security system is that users have a higher proclivity to select intrinsically unsafe passwords for the sake of easy memorization, etc. This phenomenon directly leads to surfacing of dictionary attacks [13], where the adversary over the network attempts various permutation and combination of strings leading them finally to arrive at the correct password by the genuine user. With the current availability of various password hacking tools [14][15][16] as well as keylogger software [17], the task of password retrieval has become much easier for the attacker. Not only public network but also private network is not secure in existing set of hacking methodologies and tools. Various events have been reported in the past, where it is found that number of reputed enterprises like eBay, ICICI Bank, World Bank, Walmart, etc. have been literally hacked costing massive loss of property and highly confidential data [18].

Above all, every publicized password attack is sometimes followed by a series of articles decrying the “end of the password” and business for the implementation of Multi-Factor Authentication (MFA) [19]. MFA proves to be a more secure remedy that piggybacks multiple authentication items from multiple factors [20]. A website using MFA is tougher to attack – to “break into” – than a website authenticating users with solely one issue like a password. The widespread adoption of MFA would improve on-line security and facilitate a reduction in fraud. Moreover, MFA is not a replacement plan rather it has been enforced in online systems for several years. Till recently, however, MFA has seldom been deployed with success in very large-scale websites meant for communities like consumers within the light-weight of the increasing word attacks, practices are setting out to an amendment.

Combining multiple factors for authentication enhances the accuracy and usability of the authentication process. It also reduces the False Rejection Rate (FRR) of genuine users [3].

One such technique is Two Factor Authentication (2FA) that combines any of the two above-mentioned authentication factors. For the reason of high cost and complex design, biometric-based security systems are not widely adoptable [3]. Currently, the usage of One-Time-Password in 2FA is extremely popular [21]. OTP authentication, also known as session authentication, utilizes a password that can be used only once unlike the traditional re-usable passwords which are being used over relatively longer periods of time. Moreover, OTPs are encrypted before being transmitted to reduce their interception [3].

B. OTP Generation and Distribution:

To ensure the security of the system, the generated OTP should not be traced, guessed or retrieved easily by the intruders. Thus, a secure authentication mechanism needs to be developed for OTP generation. OTP generation can be achieved by any of the three mechanisms:

1) Use of a mathematical algorithm for generating a new password on the basis of previous password,

2) Use of a time-synchronous scheme, where the designed algorithm executes in both the client and the authentication server for generation of password,

3) Use of a challenge-response schemes in that the server or any other authenticating system issues a challenge to the host seeking authentication and expects a response. In addition to this, a counter is employed instead of the previous password [22].

Though the generated One Time Passwords are immensely employed but at the same time they face with several limitations including synchronization problem and the inability of the algorithms to provide security with continued existence.

Distribution of OTP to the user is achieved in two ways: i) the client can request the service provider i.e. a bank or authentication server to provide the OTP. The server can deliver the OTP via GSM network or it can provide the client with hardcopy of OTPs i.e. TAN lists (Transaction Authentication Number) [23][24], ii) Alternately, the service provider provides client with tokens that can generate OTP on client side i.e. EMV Card and Reader, Hardware/Software OTP Tokens, etc. [25][21].

It needs to be ensured that the method used for OTP generation and distribution should be acceptable widely i.e. it should be user ergonomic. The methods used should be intuitive, simple, and economical. Also, the cost of devices used should be taken into consideration in the case of device failure i.e. they should be reliable and robust.

The prime objective of the paper is to study various security measures that were introduced in the past for effective access management. Section II is entirely devoted to the concise survey of numerous current authentication/authorization techniques. Section III highlights the prominence of open issues from the discussed literature and finally the paper is concluded with some clinching remarks in Section IV.

II. Related Work

A detailed study of the relevant literature that has been introduced in the past for the purpose of mitigating security issues about authentication and authorization are discussed in this section. The study involves authentication schemes that are characterized into two categories: OTP based and non-OTP based schemes.

A. Review of OTP Based Schemes

Many authentication schemes have been studied by researchers but the strongest of all have been found to be the ones based on the concept of One-Time-Password. There are numerous schemes of generating One-Time-Passwords.

In [22] a noisy password technique has been implemented that results in passwords that are robust against any shoulder surfing or eavesdropping. This technique enables users to use passwords suitable for them to remember. The password provided by the user together with the noisy part is used to generate complex and different password at each login. The noisy password imbeds the actual password within it. Noisy password, in particular, has four parts that make it extremely challenging computationally to deduce password from the noisy part. User is authenticated with a different password for each login by employing proposed technique. The system achieves user authentication by permitting the user to enter the username i.e. an ID and a noisy password. The technique utilizes two algorithms; first algorithm is used for choosing the password and its storage in a database. The second algorithm is employed for the extraction of the password and its recognition. The proposed work has also limited the number of incorrect login entries to five attempts. The system has made use of noisy passwords that exponentially increase the processing time.

In [26] authors have put forward a scheme known as 'Infinite Length Hash Chains' to increase the efficiency and extensibility of the conventional chaining idea with the help of public-key techniques. The ILHC scheme is the scheme where the length of the hash chain can be increased infinitely thereby facilitating its use without the need of system restart or bootstrapping. The proposed ILHC scheme employs a public-key algorithm to produce an infinite one-way-function that forms the one-time-password production core. Unlike the one-way function chains, the proposed system allows the owner of the chain to go in whatever direction any time and doesn't limit the length of the hash chain. Usually, certificates expire after certain interval of time; say a year or two but the usage of ILHC scheme enables the user to use server-supported signatures whereby the user is no longer restricted by the number of messages to be signed. In the proposed scheme, there is an increase in the computation cost as a consequence of operations involved in public-key algorithms. This can be owed to the exceedingly large number of cascaded exponentiations. As a result, this scheme is difficult to be employed in devices with restricted computational resources (say, mobile phones).

An algorithm proposed by authors in [27] implements a knock sequence that is secure and employs AES (Advanced Encryption Standards) encryption scheme. Thus, the proposed scheme cannot be sensed by sniffing or spoofing. This authentication scheme is an improvement over the port-knocking mechanism in place. The authors have proposed a novel framework that further enhances the security level in the prevailing port-knocking models by making it more complex for the attackers to reveal the correct port knocking sequence. In this system, the concept of One Time Password (OTP) is used that acts as the One Time Key for complete AES encryption technique in addition to Quadratic Residue Cipher (QRC) to spoof the source IP addresses a number of times thereby leading to raised complexity in discovering the sequence of IP addresses and packets. Also, the pseudo random number generator (PRNG) is being used to produce the random numbers in real time that shall be used both in QRC and as a key for the AES encryption. Using this scheme, various problems like out-of-order-delivery problem of packets are swept away since all the packets irrespective of their source being assembled at the server at first and then decrypted to obtain the correct knock sequence. The system makes use of SMS to send OTP and a random number to the client. However, sending an OTP over an SMS to the user has certain restrictions like cost, lateness, security, etc. In addition to this, the SMS-based OTP can be compromised.

A secure and quite compatible web/mobile-based authentication mechanism is presented in [28] with an attempt to enhance multifactor authentication. The authors put forward a method of generating one-time password keys in OTP clients with the help of PingPong128 stream cipher. Stream ciphers have been developed to approximate the behavior of one-time code. The proposed system makes use of an innovative approach for producing an authentication method that employs IC's (Identification code) to implement an additional security level in the conventional login system. IC's are the identification codes that are unique to each user and each transmission and are provided by companies, banks or other financial institutions to the user. In this scheme, the authentication server produces a secret code only once and then changes the value of the secret code in the next instance. The system makes use of Advanced Encryption Scheme (AES) for encrypting the generated OTP. The protocol employs multifactor authentication for verifying the user and the current transaction. The client side has been simulated using The Sun J2ME Wireless Toolkit, that comprises of build tools, utilities, and a device emulator.

The authentication server is based on J2EE technology with web server Glassfish and database Mysql. Module IC codes are pseudo-random codes that can be generated employing pseudo-random number generation algorithm. It is only an authorized person who distributes the IC's to the user's phone through the web browser or a Bluetooth device. The distribution process also includes the encryption of IC's. Once implemented, the proposed protocol does not add to the expenses of users considerably as it has an easy implementation and can be executed on the current costs that the servers incur from users. However, this model involves two communication channels viz. TCP/IP and GSM.

A novel OTP authentication scheme based on true random numbers has been put forward in [29]. In the proposed scheme, the true random numbers are generated using physical methods i.e. from digital physical noise sources and are then applied in the process of secure authentication. The generation of random numbers in this scheme is based on the characteristics of the common digital-analog hybrid circuit. The random numbers generated by this method are non-repetitive, stochastic and unpredictable that ensures that it is almost impossible to crack those random numbers. This scheme provides a defense against attacks of human sources and can be viably used in the areas where efficiency and high security is needed such as stock exchange system, finance system, etc. The authentication scheme uses the true random numbers as the password that is encrypted and decrypted by the client and server twice respectively. The proposed authentication mechanism is a two-way authentication scheme i.e. both the server and the client authenticate each other. The client uses challenge/response method to ask authentication each time and the encryption algorithm used is RSA public key cryptography. The cryptographic algorithm employed in this scheme is easy to implement as there is not an involvement of third party. A large amount of random numbers can be generated using this scheme at a fast rate that decreases the expenses paid by the server. This, in turn, enhances the efficiency of the server as well as the entire authentication process. However, the proposed system is found to be vulnerable to phishing attacks.

A secure authentication scheme has been proposed in [30] using smart cards that preserves the properties of previous authentication schemes but does not put any restriction on the number of login attempts. At the same time, the proposed scheme can be said to provide resistance against stolen-verifier attack, replay attack, password guessing attack. As a result, this scheme has been found to improve both the applicability of OTPs as well as the security level in the system using this concept. The proposed scheme is composed of three phases: the registration phase, the login phase, and the proof phase. It is the server that issues the smart card to the user that holds the initial secret seed. A one-way hash function was employed in this scheme that is collision resistant. The server also transmits a timestamp to the user along with the previous results of the computation. It is this timestamp that is used to check the legality of the server. The proposed scheme has been found to be resistant to several attacks, e.g., server spoofing attack, user impersonation, replay attack, password guessing attacks (offline as well as online), stolen-verifier attack. Moreover, this authentication system ensures a high level of security since both the server and the user authenticate each other. The proposed authentication system is reliable and complete in terms of the fact that the important data cannot be retrieved even after analyzing the data transmitted. However, the use of an extra device i.e. a smart card in the proposed authentication scheme can cause inconvenience to the user and at the same time may prove expensive to the service provider.

An efficient end-to-end one-time-password authentication scheme has been put forward in [31] that eradicates the drawback of the existing authentication protocols. The proposed scheme utilizes two cryptographically strong base elements viz. the Authentication Key Exchange (AKE) protocol and the keyed Hash Message Authentication Code (HMAC). Such an authentication protocol has been found to provide transparency in the mutual authentication between the two participating end-points. In this authentication system, the various operations involved like Key Setup, Key Scheduling, and Key Update are performed independently at the two end-points with no interaction between them. As a result, the proposed authentication scheme ensures a high level of security since there is no involvement of a trusted third party. Moreover, this scheme is highly secure cryptographically owing to its resistance against a range of attacks that the present OTP authentication schemes encounter. The proposed scheme makes use of HMAC-SHA512, that is based on the cryptographic hash function SHA3. HMAC is used for message authentication and to verify data integrity. This authentication protocol can be operated independently in various scenarios because it is relatively simple and has less computational overhead as compared to the previous protocols. Owing to the simple operation of the proposed scheme, it can be very suitably adapted by resource-constrained mobile devices. However, this authentication mechanism suffers from a limitation that it is not much strong cryptographically as it cannot be used in cases when the number of iterations exceeds the length of the mutually agreed upon Master Key. Furthermore, the security of the system relies on secure handling and storage of Master key i.e. the Master key should not be stolen or compromised.

In [32] authors have put forward a novel authentication mechanism based on Chebychev chaotic mapping. The chaotic mapping has various characteristics such as sensitive dependence on the initial condition and structural parameters; and unpredictability that are the key factors of authentication. The proposed system has designed an authentication mechanism between intelligent electronic devices (IEDs) in substation automation. This system has been proposed taking into consideration three protocols—GOOSE, SMV and GSE, that have rigorous performance requirements as a result of that encryption or various other security measures can drastically affect their rate of transmission. The proposed chaotic sequence based authentication scheme has been found to be efficient than the general OTP authentication schemes since there is no need to store the whole sequence thus utilizing less memory. The chaotic mapping employed in this scheme being an identity mark sequence ensures that the sequence generated cannot be imitated by others. The use of one-way function F ensures secure and easy authentication, since it is practically infeasible for anyone to generate the initial seed ‘s’ from the value y = F(s). The chaotic sequence that is generated in the proposed authentication scheme is found to be responsive to the initial condition such that a small change in y_0 produces large changes in the sequence y_n. Although, the chaotic sequence based authentication schemes have been found to be fast and secure but a large number of those systems have been effectively cryptanalyzed owing to the finiteness in its computing precision that is used as a means to represent the floating point output of the system, thus making the current system susceptible to attacks.

The paper [33] has proposed a novel authentication mechanism that exploits location and time information of mobile device as physical parameters to generate the One-Time-Password (OTP). The framework restricts the validness of OTP in a definite time-period along with the tolerant geometrical location, thus enhancing security. Present statistics of moving directions and movement of the mobile device are employed to improve the precision of location prediction. Transparent authentication of the user is presented, provided the user moves steadily thus avoiding manual typing of credentials every time while being transparently logged into the server. The authors have utilized both event based and time synchronized OTP mechanisms. In event-based, the system considers that the user should be in the tolerant region while in time-based, mobile device and server are clock synchronized. An assumption is made in the system that the users are already registered on the application server and are organized in Public Key Infrastructure (PKI). The system involves two phases. In the first phase, it calculates the user-tolerant range of expected future location, from the user's current location and time. For user to login the server, the location obtained from GPS receiver and the current time is used to generate OTP. The generated OTP and the International Mobile Subscriber Identity (IMSI) are concatenated and encrypted by a public key and then sent to the server. The server decrypts the received message using a secret key. OTP is extracted by the server and from it, the time and location coordinates are retrieved. If the coordinates are in the tolerant region of predicted destination, the server authenticates the user otherwise it rejects the login request. System provides secure user authentication while accessing crucial Internet services such as e-commerce and online banking transactions; immunity against several types of attacks like eavesdropping, replay, man in the middle, dictionary, brute force, and other user impersonation attacks. The proposed authentication mechanism works with GPS-enabled mobile phones. Further, clock synchronization between the server and the mobile device is required, that is difficult to achieve in the case of mobile phones as mobile phones are certain to move out of the network due to that synchronization fails.

A novel OTP authentication mechanism that resolves counter de-synchronization problem by employing symmetric encryption algorithm and one-way hash function was proposed in [34]. It has effectively minimized Denial of Service (DoS), guessing, and replay attacks. The approach makes use of symmetric encryption scheme i.e. AES, that offers resistance to differential and linear cryptanalysis in coordination with a one-way hash function i.e. MD5. The symmetric cryptographic algorithm provides one-way functionality. The symmetric cryptographic algorithm encrypts counter value to generate OTP. Further, the symmetric encryption algorithm generates an output that is capable of carrying the counter value to the server thus avoiding the counter de-synchronization problem. A single comparison is done by the server so as to solve counter desynchronization problem. Using only symmetric encryption algorithm will result in a successful attack. To avoid such attack, the system makes use of a digest that are the verification bits of counter need to be added i.e. MD5. In the scheme, OTP has two parts P = C | | D, where C signifies cipher text, D signifies verification bits. 48 bits one-time password using 24 bits of C and 24 bits of D are used in the system, MD5 calculates one part and AES calculates another part of the password. Illegal submission is recorded by the server if password authentication fails. The server restricts a user if the number of illegal submissions reaches a specific threshold. The scheme has minimized the security requirement of server and provides an easy integration in the present enterprise applications.

The paper [35] has proposed a scheme that makes use of one-time-passwords as dynamic passwords that change according to time, user ID and some other factors unique to user i.e. Media Access Code (MAC). The scheme has utilized time and space factor to provide secure authentication mechanisms. In this system, various problems associated with the current 2FA scheme have been eliminated by minimizing user's effort or/and the high overhead associated with the generation of one-time-password. It has employed software for OTP generation thus providing easy user adoption and an economical solution. Proposed dynamic authentication mechanism involves four steps: sign-up, OTP generation, transferring authentication information to the server and finally, the verification. In this technique, the OTP generation and user authentication are performed on the same machine. In the sign-up phase, the user enters the static password provided by the server that holds the master key. The hash function is used to generate Password Digest (PD) from the static password entered by the user and the digest is saved in server's database. The second phase comprises of small software that is used to generate the OTPs from the static passwords of the user. The generated PD is XORed with a master key that is present in the software, so as to generate Master-key and Password Digest (MPD). The generated digest together with MAC and time are concatenated and fed to hash function to generate an OTP. This phase may include compression of the generated OTP into a smaller length OTP. The generated OTP is considered user's signature at specific machine and time and helps to obtain time and space dynamism. The generated OTP at user side is denoted as U-OTP (user side OTP). The third phase involves transferring authentication information to the server. Since, the hash function generates output as binary strings therefore encoding technique Radix64 is utilized to present the user an OTP that can be typed and read easily. The generated U-OTP is manually entered or copied on the login screen of website. In addition to other authentication information, MAC also has to be transferred to server. MAC can be captured using JavaScript techniques by a web page. The final phase involves verification. The server will extract the MAC address of the user and generate OTP by the same mechanism and compare the details upon reception of authentication information. The server side retrieves MAC and PD easily retrieved but for obtaining time synchronization information it may use any of the two strategies: add a time factor to OTP and server guesses method. As the scheme has utilized space and time dynamism it enables the proposed authentication mechanism to resist Perfect-Man-in-the-Middle and replay attacks. In the proposed scheme, the hashed output of static password is stored in server's database that enables the proposed authentication mechanism to resist stolen-verifier attack. However, this system has been found to be resistant to basic phishing attack as the user is not allowed to enter the password on login form.

Traditional hash chains suffer from updation problem in that seed should be updated after generating finite hash values. In [36] authors have introduced an enhanced self-updating hash chain that eliminates the practicality and security issues associated with the conventional Infinite Length Hash Chains (ILHC). The authors have demonstrated a self-updating hash chain, which is based on Linear Partition Combination Algorithm (LPCA). LPCA, which is a data distribution scheme, splits the data or seed in 'm' parts and then departs those parts. No information about the seed can be obtained from these parts as they do not hold any information about the seed. This scheme eradicates the shortcomings of traditional infinite length hash chains used for authentication that are associated with increased computational cost in case of public key based infinite length hash chains and weak security due to leak of part of next hash seed value at every transmission in case of one bit information exchange algorithm. The model has made use of LPCA algorithm, the computations for data dispersal and data restoration functioned in Galois Function (2). Further, the technique requires additional storage than the conventional authentication mechanisms. Also, the security of the system is centered on LPCA scheme and one-way functionality of the hash function. OTP has solved the security issues associated with static passwords, but the manageability of same is a problem for consumers.

A novel authentication mechanism has been presented in [37] for home networks by using smart cards based on One-Time-Password (OTP). The proposed scheme provides strong authentication mechanism by using OTP and one-way hash function that reduces the computation load enabling the system to meet the security requirements of home networks. Mutual authentication is established using three-way challenge-response handshake. The scheme has accepted the current home networks based on one-time-passwords. Authors have designed this system between the user and home gateway. The proposed authentication mechanism involves three major phases: registration phase, login phase and authentication/service request phase. In this model, the secret key has to be shared between Home Gateway (HGW) and Internet Authentication Server (IAS). The system enables the server to verify multiple access requests from a home user in single verification. It makes the home networks resistant against passive attacks such as eavesdropping, replay, Man-in-the-middle and Denial of Service and Stolen-verifier attacks. Mutual authentication between the home user and the authentication server is done thus avoids phishing attacks. The system discards timestamp to eradicate serious time synchronization problem. Further, Session key agreement is used in every session to provide secure connection and minimizes the burden on systems involved in authentication mechanism as no verification tables need to be stored. The security of the system is based on the non-invertible property of hash function and a nonce that is used to prevent time synchronization problem. However, this system proves to be unsuccessful to offer protection against various active attacks.

OTP has solved the security issues associated with static passwords, but the manageability of OTP is a problem for consumers. In [38] authors have proposed a technique that has extended the password generator to increase the manageability of OTP. The system generates website specific passwords by applying a one-way cryptographic hash function over website domain name and password. The proposed scheme has proved to provide superior performance by taking into consideration transmission bandwidth and computational cost of password verification/generation. The scheme makes use of Manageable One Time Password (M-OTP) module that can be any firmware module or some software program on the consumer device. The user submits only one password to this module and obtains a website-specific OTP. The web browser transmits the generated OTP to the web server, which performs authentication. In this scheme, Advanced Encryption Standard (AES) algorithm is used as to provide one-way functionality as well as for encryption. A large number of iterations of one-way cryptographic hash function has been employed to generate the password, due to that the proposed system has been able to resist offline dictionary attacks.

The manageability of credentials and identities for different internet services has become difficult for users. A One-Time-Password (OTP) MIDlet working on a mobile phone for integrated authentication designed for various types of internet services is presented in [39]. The proposed system minimizes the burden on users by automating the solution. This technique results in reliable multi-channel authentication mechanism by combining internet connection and GSM to give-and-take authentication messages. In this technique, challenge-response mechanism is employed to generate OTP. The model comprises of Java MIDlet installed on java supported mobile phone, an applet on the user terminal to pass the OTP to Authentication Server (AS) i.e. a servlet. The core of the scheme is that there exists a closed loop among all components in the system. The user accesses the internet services through the browser having a java applet on the user computer, equipped with internet connection. The service providers are associated with authentication servers to handle authentication mechanism. The authentication server connects to the applet on HTTPS connection and with the mobile phone through SMS over GSM network. Finally, the mobile phone inputs the credentials via Bluetooth or if there is no Bluetooth, user enters the credentials manually.

The paper [40] aggregates the advantages of both software and hardware tokens by integrating them on mobile phones equipped with hardware mobile trusted module (MTM). The technique presented provides usability with strong security and scalable OTP solution using mobile phones as hardware tokens together with trusted computing technology. In the proposed scheme, the trust factor is established between the service provider and the Mobile Local Owner Trusted Module (MLTM) equipped mobile phone. The MLTM acts as a secure processor to create a series of OTPs. MLTM supports SHA-1 as OTP generator function. The proposed model considerably reduces the cost and need of having separate hardware tokens for different service providers and thus allows users to handle multiple OTP service providers on a single mobile phone hence, eliminating problem of economics and scalability. In the presented scheme, user authentication as well as data origin authentication is taken into consideration. Usage of two separate channels concurrently i.e. the internet and the mobile network lead to a complex system that will minimize the man in middle attack. The separation of the mobile phones from the user client terminal restricts the usability that necessitates the user to copy the OTP manually from mobile phone to the client terminal.

Top military commands, government agencies, etc. require absolute privacy and security that lasts endlessly. The intervention of “Top Secret” in a month or after 100 years can prove disastrous. In [41] authors have proposed a framework that delivers absolute security by making use of one-time-pad and supplying the random keys by using a high throughput binary random sequence generator. The framework has introduced an up-to-date usage of one time pads for achieving absolute security by introducing 100Mbit/s hardware binary random generator. The proposed scheme has solved the problem of availability of long one time keys (OTK) or one time pads (OTP). It presents an infinite source of one time keys by making use of the random generator. However, this system involves a high cost for secure physical distribution of keys thus, hampering adoptability.

The various problems about security and authentication of accessing private and highly privileged information studied by many researchers are tabulated below for benchmarking. The contributions and associated shortcomings of the various OTP based authentication approaches already discussed are formulated in Table I as follows:

TABLE I. REVIEW OF OTP-BASED SCHEMES

AUTHOR	CONTRIBUTION	RESULT OBTAINED	LIMITATIONS
(Bicakci and Baykal, 2002)[26]	Proposed Infinite Length Hash Chains that use a public-key algorithm to generate one-time-password	• Network overhead and system-restart complexity is evaded. • Owner of the hash chain is allowed to go in whatever direction anytime with no limit on the length of the chain. • Enables the user to use server-supported signatures whereby user is no longer restricted by number of messages to be signed.	• Increase in computation complexity as a consequence of public key operations makes it difficult to be employed in devices with restricted computational resources e.g., mobile phones.
(Chang et al., 2004)[30]	A secure authentication scheme using smart cards with no restriction on the number of login attempts	• One-way hash function and XOR operations employed ensure its efficiency. • Defensive against a number of attacks e.g., server spoofing, user impersonation, replay attack, password guessing attacks(offline and online), stolen-verifier attack. • Highly secure since both server and user authenticate each other. • Important data cannot be retrieved even after analyzing the data transmitted.	• Extra hardware token involvement.
(Long and Blumenthal, 2007)[38]	Designed Manageable-OTP by extending password generator for consumer applications.	• Increased consumer convenience by offering manageability for OTP based authentication systems. • Resistant against offline dictionary attacks.	• Employed symmetric encryption algorithm (AES-128) and one-way hash function (MD-5) that have already been compromised.
(Hallsteinsen and Jorstad, 2007)[39]	Presented a unified authentication scheme based on mobile phones	• Increases user adoptability and eliminates the weakness associated with many existing time synchronization based OTP schemes. • Reduces cost and burden of managing hardware token compared to other OTP schemes. • Offers resistance against eavesdropping, man-in-middle, replay, hacking, sniffing and guessing attacks.	• SMS has been used by the authentication server to perform key exchange and communication with user, that can be compromised. • Reduces user friendliness when client terminal does not possess Bluetooth facility.
(Jeong et al., 2008)[37]	Provided a novel authentication mechanism for home networks by using smart cards based on one time passwords.	• Minimized the computational overhead and communication cost. • Immune against various passive attacks viz. passive eavesdropping, replay, Man-In-The-Middle, Denial of Service and stolen verifier attacks. • Discards timestamp to eradicate serious time synchronization problem. • Offers user convenience by enabling home user to freely choose the password. • Avoided phishing attacks by providing mutual authentication between home user and authentication server. • Minimized burden on the systems involved in authentication mechanism as there is no need to store verification tables.	• Fails to preserve the privacy of transmitted data. • Protection against active attacks is not provided. • Smart cards are used for achieving authentication, that are not devoid of short comings.
(Li and Zhu, 2009)[32]	Designed a novel authentication mechanism between IEDs in substation automation based on Chebychev chaotic mapping	• Utilizes less memory as there is no need to store the whole sequence. • Chaotic sequence generated cannot be imitated by others. • One-way function employed ensures secure and easy authentication. • Chaotic sequence is sensitive to initial condition.	• Large number of chaotic sequence based authentication systems have already been cryptanalyzed effectively.
(Alghathbar and Mahmoud, 2009)[22]	Designed a novel one time password authentication	• Robust against shoulder surfing or eavesdropping.	• Usage of noisy passwords

	mechanism based on noisy password technique.		exponentially increases the processing time. • Less user-friendly.
(Liao et al., 2009)[34]	Presented a one-time-password authentication mechanism eliminating counter de-synchronization problem.	• Resolved counter de-synchronization problem. • Minimized security requirements of the server and provided easy integration in present enterprise applications. • Effectively minimized guessing and replay attacks.	• Employed symmetric encryption algorithm (AES-128) and one-way hash function (MD-5) that have already been compromised.
(Davaanaym et al., 2009)[28]	Proposed a secure and market-compatible mobile/web based authentication mechanism that generates OTP using PingPong128 stream cipher	• Resistant to attacks based on basic key-stream properties like period and linear complexity. • Overcomes time-memory tradeoff. • Easy implementation and can be executed on current costs incurred by servers from users.	• AES has been employed for encryption of generated OTP.
(Tao et al., 2009)[29]	Designed a novel two-way authentication scheme based on true random numbers generated by physical methods	• Password generation is random, fast and dynamic • Provides defense against many attacks e.g., interception, forgery, server-forged attacks, etc. • Improved efficiency of the server as well as the entire authentication process	• Cannot thwart guided phishing attack.
(Min-Qing et al., 2009)[36]	Introduced an enhanced self-updating hash chain based on LPCA.	• Eliminates practicality and security issues associated with the conventional infinite length hash chains (ILHC).	• Additional storage than the conventional authentication mechanisms is required since a part of next root seed is stored at each process
(Alzomai and Josang, 2010)[40]	Presented mobile phone as scalable OTP device based on trusted computing.	• Provided solution for achieving scalability and usability. • Minimized man in the middle attacks.	• SHA-1 has been employed as OTP generator on that theoretical attacks have been reported. • Restricts the user to generate valid OTPs when attacker masquerades the service provider. • Separation of mobile phones from user client terminal restricts usability. • Wide technical adoptability of the proposed system is not supported.
(Srivastava et al., 2011) [27]	Proposed an algorithm that implements a knock sequence employing AES capable of withstanding spoofing or sniffing attacks	• Almost impossible to detect and interpret the successive knock sequences • Implementation of multi-packet authentication mechanism prevents data to be divulged • Eliminates out-of-order delivery problem of packets • A range of attacks viz. man-in-the-middle attack, denial of service attack can be avoided	• OTP generated is sent over a GSM network
(Hsieh and Leu, 2011)[33]	Proposed an authentication mechanism based on time and location of mobile phone.	• Provided secure user authentication for accessing crucial internet services • Immune against various attacks such as eavesdropping, replay, brute force, and user impersonation attacks • Transparent user authentication • Improved precision of location prediction.	• GPS enabled mobile phones are required • Clock synchronization is required between mobile device and server
(Ren and Wu, 2012)[35]	Provides a secure authentication mechanism utilizing time and space factors	• Minimizes user's effort or/and the high overhead associated with generating OTP • Easy user adaption • Effectively resists Perfect-Man-in-the-Middle, stolen-verifier, replay and basic phishing attacks.	• Susceptible to phishing and other attacks that are highly sophisticated and on rise.
(Borowski and Lesniewicz, 2012)[41]	Presented an up-to-date usage of old one time keys or pads by introducing 100Mbits/sec binary generator.	• Provides absolute security by making use of 100Mbits/sec hardware binary random generator • Provides infinite source of one time keys	• Involves high cost for secure physical distribution of keys thus hampering adoptability
(Castiglione et al., 2014)[31]	An efficient end-to-end OTP authentication scheme involving AKE protocol and the keyed HMAC	• Simple and less computational overhead thus can be operated independently. • Provides transparency in addition to efficiency. • Resistant against wide range of attacks e.g., password guessing attack, offline dictionary attack, brute-force attack, replay attack, eavesdropping, stolen-verifier attack and denial-of-service attack • Suitably adoptable.	• Cannot be used when the number of iterations exceeds the length of the mutually agreed upon Master Key. • Security of the scheme relies on secure handling and storage of the Master Key.

B. Review of Non-OTP Based Schemes

Other than OTP based authentication schemes, researches have been conducted on security solutions that are based on biometrics, fuzzy vault schemes, chaotic mapping, etc.

The paper [5] demonstrates the usage of speech for identity authentication i.e. they make use of speech features obtained from speech recognition. Usage of speech features is highly beneficial as the same are having stability and uniqueness characteristics. Moreover, it is difficult to be forged and can be easily carried by the user, resulting in better user ergonomics. Speech authentication is associated with speech recognition that incorporates two fundamental phases, viz., feature extraction and matching. The security of speech authentication system can be easily compromised if the intruder succeeds in recording the voice of the authenticated user and uses this recorded voice to break the system.

In [42] authors have proposed a federation Single Sign-On (SSO) authentication scheme based on network identity. The one-pass authentication technique presented is very fast and secure since it ties together two authentication schemes viz. Network Attachment Control Function (NACF) and IP Multimedia Subsystem (IMS). The bundled authentication scheme is useful for the mobile users in the Next Generation Network (NGN). In NGN, the Federation SSO is a method used for authenticating IMS service and web application service, i.e., even if, only network operator is authenticated, there is no need to authenticate application service. It is thus evident that this method reduces the complexity as compared to previous approaches. To realize the federated SSO scheme, the prerequisites are Service Control Function (SCF), Network Access Control Access Function (NACF), Web Application Service Control Function (WASCF) and NGN Terminal Function (NTF). In this scheme, authors have introduced the Authentication and Key Agreement (AKA) vector in 3GPP that comprises of an Integrity Key (IK), a Cipher Key (CK), and a credential for authentication. The access network operator is there to provide the unified access authentication to both wired and wireless networks. The service network operator is there to provide IMS service authentication for the user equipment making use of SIP REGISTER. Here, MD5-Digest and MD5-AKA are employed for authentication purposes. The proposed scheme proves to have higher security and reliability as against the previous SSO authentication mechanisms owing to the use of a reliable network operator—NACF with Identity Provider. Thus, unlike the earlier systems where authentication is provided between application services, this federated single sign-on considers authentication between web application service and NACF. This scheme prompts the user to select or subscribe the bundled authentication process; hence the user is given privilege over the federation operator. It can prove beneficial when the user has to access the network about multiple service network operators. But it has been found that the proposed scheme fails to provide security against spoofing attack as it only involves information about the location of the user equipment and does not consider security operations. Moreover, some access identifiers need to be added to the profile of user equipment to identify the user because it is devoid of fixed line information.

An authentication scheme has been put forward in [43] to ensure secure user authentication in a cloud computing environment. Enormous volume of data has to be handled in real time in cloud computing, therefore it is imperative to devise an authentication system that is lightweight, cost-effective, fast and most importantly secure i.e. robust against attacks. Thus, the authors have presented a lightweight and efficient multi-user authentication mechanism that is based on Cellular Automata (CA) in cloud computing environments. The proposed scheme works in almost the same way as the One-Time-Password (OTP) authentication, the only difference being use of non-linear Cellular Automata (CA) for the purpose of random key generation.

The process of authentication has been illustrated between the user and the authentication server and is accomplished in two distinct phases: a setup phase and an authentication phase. The proposed system makes use of a CA-based Pseudo-Random Number Generator (PRNG) that gives the system several properties like vast area complexity, uniform structure, fast operation, prompt hardware implementation, uniform structure, etc. The security of the proposed scheme has been experimentally proven with the help of a DIEHARD test suite. The DIEHARD test has generated the p-value pass rate ≥ 85% that is measured as “good”. As a result, this authentication system proves to be secure. However, the security of this system can be further enhanced by improving the randomness of the pseudorandom generator. Furthermore, this authentication system is based on non-linear CA whereas linear CA could provide a much higher degree of randomness.

The paper [44] introduces a system that makes use of a fuzzy vault scheme for the protection of biometric information. The user authentication with the help of biometric data can prove to be a stronger security measure. The proposed scheme makes use of biometric information for authentication purpose that makes it more reliable since a biometric data cannot be lost, changed, copied and guessed. The fuzzy fingerprint vault has been found to be the most accepted solution to safeguard the fingerprint features. With the help of fuzzy vault scheme, the data can be made secure by combining it with a biometric template in such a way that only the user who is authorized can be granted access to the secret data after providing the genuine biometric. In previous systems, it was assumed that the fingerprints were already aligned but this was not a rational assumption regarding authentication systems based on fingerprints.

As a result, three solutions have been put forward by the authors for fuzzy fingerprint vault that make the biometric-based data authentication more secure and efficient i.e. automatic alignment of fingerprints based on a geometric hash table; a better and secure fuzzy fingerprint vault that can provide resistance against correlation attack; fuzzy fingerprint vault employing One-Time-Template (OTT) producing a diverse biometric template every time, just like One-Time-Password (OTP). These solutions have been proposed to improve the security of biometric data.

The proposed system has shown a performance of 91.17% Genuine Accept Rate (GAR) without affecting False Accept Rate (FAR) (0.6%) with an 8-degree polynomial in FVC2002 DBI. Also, the performance of 92.1% GAR and FAR (0%) has been reported with a 7-degree polynomial. In this scheme, there is no need to store additional information, e.g., geometry hash table, helper data, etc. that may otherwise degrade the security level of the system. The proposed scheme makes sure that the access to secret data is granted only to the authorized user, but there remains liability that some sophisticated attacks may compromise the secret data or biometric information. Moreover, this scheme is suitable for restricted applications only as the system cannot scale well to large service pool.

In [45] authors have presented a secure and efficient authentication system based on a smart card to protect against the vulnerabilities as well as to improve the security in the existing systems. The proposed authentication system allows the user to choose a password very conveniently and even modify it offline. This system blocks stolen user smart card attack because the smart card does not hold any important information. Also, the server attack has been eliminated by shifting the user authentication from the server to registration center that ensures that each server possesses a unique private key. Thus, this system provides high-level security and is more practical. There are three participating entities in this system i.e. user, server and registration center. The proposed system works on four protocols viz. the registration protocol, the login protocol, the authentication protocol and the password change protocol. The proposed scheme is flexible for the users in a way that it is the concealed identity that is transmitted and not the actual identity of the user. It provides an efficient and secure mechanism for changing the password because the user can very conveniently change his password without relying on the registration center. The proposed system was found to be more cost-efficient after performing a comparative analysis with other systems.

A new identity authentication scheme has been proposed in [46] with employs a Contactless Smart Card (CSC) that holds a multitude of biometric features. This authentication scheme is aimed to enhance airport security by securing logical access. The contactless smart card is useful for authentication because of its various features like the low mechanical complexity of the reader-writer unit, fast speed, reduced maintenance cost and secure physical access. The proposed authentication system makes use of fingerprint recognition and iris scanning for providing support to many fields such as airline passengers, border security, transportation security, law enforcement and logical access. In the proposed work, a Two Stage Random Number Generator (TSRG) has been employed that makes use of randomized encryption techniques to design a TSRG cryptosystem that is secure functionally. The fundamental biometric feature used in the proposed scheme is the fingerprint. For users who are not comfortable to enroll using fingerprint template, iris recognition has been recommended. Then, biometric authentication is performed to determine the identity of the user. This authentication scheme enrolls the biometric live features of the user. Moreover, the usage of Ferroelectric Random Access Memory (FRAM) technology in smart cards augments the efficiency of the proposed system as it consumes less power, has a higher write speed and greater rewrite endurance as compared to its counterpart i.e. EEPROM. Despite its numerous features, there lies further scope for improvement in the design paradigm of the proposed system by focusing on coordination, cooperation and interoperability.

In [47], the authors have re-examined the security claims of Predicate-based Authentication Service (PAS) and successfully indicated PAS was insecure against probabilistic attack and brute force attack. The PAS system claims security against three attacks: random guess, SAT (satisfiability solver) and brute force attacks that is highly over-estimated. The proposed system introduces probabilistic attack, which even with a small session of authentication breaks part of the password.

Further, it was also found that the PAS system is poor in terms of complexity and security against low complexity than Cognitive Authentication Scheme (CAS). The attack is also computationally efficient and reduces the PAS system to challenge-response based OTP system. Thus, it has less security as well as usability than OTP systems. However, the probabilistic attack introduced on PAS is unable to completely break the password or secret key shared between server and user.

The prominent pros and cons of a variety of Non-OTP based security techniques have been framed in Table II as follows:

TABLE II. REVIEW OF NON-OTP BASED SCHEMES

AUTHOR	CONTRIBUTION	RESULT OBTAINED	LIMITATIONS
(David et al., 2003)[46]	A new identity authentication scheme employing contactless smart card that holds biometric features aimed to enhance airport security	• High reliability provided by features of contactless smart card. • Usage of FRAM augments efficiency. • Contactless smart card widens the arena of its applications from logical access to physical access. • High security ensured by combining TSRG data with several cryptographic methods.	• Additional research needs to be carried out to examine its adoptability. • Further scope for improvement in the design paradigm of the system by focusing on coordination, cooperation and interoperability.
(Li et al., 2009)[47]	Re-examined the security claims of PAS by inducing Probabilistic attack over the same.	• Successfully illustrated weakness associated with PAS towards probabilistic and brute force attacks.	• Fails to provide satisfactory solution to tackle the security issues with PAS thus hampering adoptability.
(Kim et al., 2010)[42]	Proposed a federation oriented single sign-on authentication scheme based on network identity	• Eliminates issues like message overhead and latency by bundling together NACF and IMS authentications. • Higher security and reliability as authentication is being provided between web application service and NACF. • Proves beneficial when user has to access network with regard to multiple network service operators.	• Access identifiers need to be added to the profile of user equipment to identify the user. • Only one Proxy Call Session Control Functional Entity (P-CSC-FE) is connected to NACF that is not feasible in real time. • Vulnerable to spoofing attack.
(Moon et al., 2012)[44]	Proposed three solutions for fuzzy fingerprint vault to improve the security of biometric data	• Highly reliable as biometric data cannot be lost, copied, changed or guessed. • Suitable for crypto-biometric systems since it works with unordered sets. • Infeasibility of polynomial reconstruction ensures its security. • Resistant to correlation attack. • Improved performance of GAR is fetched without affecting FAR.	• Once compromised, fuzzy vault cannot be revoked. • Biometric information may be compromised by some sophisticated attacks. • Suitable for restricted applications as the system cannot scale well to large service pool.
(Shin et al., 2012)[43]	Proposed a lightweight multi-user authentication mechanism based on non-linear cellular automata for cloud based environments	• Use of CA-based PRNG makes the system fast and implementable architecture-wise. • Resistant to various attacks like replay attack, reflection attack and eavesdropping.	• Based on non-linear CA whereas linear CA could provide much higher degree of randomness.
(Ma et al., 2013)[5]	Designed an identity authentication mechanism based on speech features.	• Provides better user ergonomics.	• Can be compromised if intruder used recorded voice of authenticated user.
(Aboud, 2014)[45]	Presented a secure and efficient authentication system based on smart cards	• Resistant to multitude of attacks e.g., stolen attack, offline dictionary attack, user attack, server attack, etc. • System turns out to be more flexible to the user as user anonymity is taken care of in the same. • No reliance on registration centre for password change. • Cost-efficient as compared to existing systems.	• Usage of an extra token shall cause inconvenience to users and may even prove to be expensive solution for service provider.

III. OPEN ISSUES

After performing the study of numerous security techniques taken up in access management as have been conducted by researchers, it can be concluded that there are still some loopholes in the proposed security solutions. The limitations have been found to be about several areas ranging from technical adoptability to computational complexity to communication media employed for the said purpose. Finally, the progressive survey conducted in this paper ends with the inference of open issues as mentioned below:

- As can be observed from the study of various authentication schemes [33][37][45][30][40][39][22], special featured hardware tokens, smart cards and other chip modules have been projected that lead to user inconvenience and may prove expensive to the service provider. Thus, these systems lack user ergonomics and this impeding their technical adoptability.
- OTP distribution is a grave issue as ascertained from the review of prevalent authentication mechanisms. It is observed that there is a dependence on external parties such as GSM or some authorized individuals [39][41][27][28].
- Numerous issues have been uncovered with the existing authentication systems [36][22][26][44] like

high storage, elevated processing time and computational cost and degradation of system speed owing to public key operations, self-updating hash chains and fuzzy vault schemes.

- Although there are some systems like [38][40][42] that allow a user to access services from multiple service providers from a single token, but majority of the authentication mechanisms discussed in the literature fall dumpy in this aspect. As a result, users craving for amenities from multiple service providers shall have to maintain different tokens for each service provider.
- There are several systems for authentication [44] that are suitable for restricted applications only, given the fact that they cannot scale well to larger service pools.
- Majority of authentication mechanisms in place [28][38][34][40], offer weak security by employing password generation schemes like SHA-1, AES, MD-5 etc, that prove to be insecure and fail on the lines of continual existence.
- It has been found that some of the authentication and authorization systems discussed in the literature employ multiple communication channels [28][40] that becomes quite infeasible to catch up in real life scenarios and further will burden the user on lines of service charges.

IV. Conclusion

The paper presents the explicit discussion of the prior research works introduced in the past for the purpose of incorporating secure authentication and authorization for any legitimate members attempting to perform secure transactions. Both significant contributions and weaknesses of these security schemes have been elaborated upon overtly. It was observed that all the authentication mechanisms aim at providing complete security but have failed in one context or the other. Some of authentication schemes employ complex cryptography that degrades system performance by increasing complexity, and are even unable to conform upon user ergonomics in an effective manner. At the same time, these authentication schemes cannot provide continued support as they bow down to advancement in the computation. It can be safely concluded that although large volume of research has been conducted in the discussed domain, but there still exist some gaps that need to be crammed because of the advent of new hacking tools and techniques on the part of hackers.

Acknowledgment

We would like to express gratitude towards Prof. Abdul Rashid Trag (*Vice Chancellor Islamic University of Science & Technology*) for his support throughout the many years of our stay at *IUST*. We are sincerely thankful to Sheikh Asif (*Assistant Registrar Academics* at *IUST*) for his aspiring guidance, invaluably constructive criticism and friendly advice during this research.

References

[1] M.A. Thakur and R. Gaikwad, “User identity and Access Management trends in IT infrastructure-an overview”, in *Pervasive Computing (ICPC), 2015 International Conference on*, Pune, 2015, pp. 1 - 4.

[2] S. Xiaoling, “The study on computer network security and precaution”, in *Computer Science and Network Technology (ICCSNT), 2011 International Conference on (Volume:3)*, Harbin, 2011, pp. 1695 - 1698.

[3] J. Kizza, *Computer network security*. New York: Springer, 2005.

[4] C. Yan-ping, L. Dong-liang and G. Rui, “Security and precaution on computer network”, in *Future Information Technology and Management Engineering (FITME), 2010 International Conference on (Volume: 1)*, Changzhou, 2010, pp. 5 - 7.

[5] H. Ma, S. Yan, X. Bai and Y. Zhu, “The Research and Design of Identity Authentication Based On Speech Feature”, in *Sensor Network Security Technology and Privacy Communication System (SNS & PCS), 2013 International Conference on*, Nangang, 2013, pp. 166 - 169.

[6] W. Stallings, *Cryptography and Network Security*, 5th ed. India: Pearson Education, 2011.

[7] R.K. Banyal, P. Jain and V.K. Jain, “Multi-factor Authentication Framework for Cloud Computing”, in *Computational Intelligence, Modelling and Simulation (CIMSim), 2013 Fifth International Conference on*, Seoul, 2013, pp. 105 - 110.

[8] W. Highleyman and S. Associates, Inc, “Hacked AP Tweet Crashes Markets”, availabilitydigest, 2013. [Online]. Available: http://www.availabilitydigest.com/. [Accessed: 21- May- 2015].

[9] Z. Zhao, Z. Dong and Y. Wang, “Security analysis of a password-based authentication protocol proposed to IEEE 1363”, *Theoretical Computer Science*, vol. 352, no. 1-3, pp. 280-287, 2006.

[10] A. Conklin, G. Dietrich and D. Walz, “Password-Based Authentication: A System Perspective”, in *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, 2004.

[11] P. Elftmann, “Secure Alternatives to Password-based Authentication Mechanisms”, 1st ed. RWTH Aachen University Aachen, Germany, 2006.

[12] B. K. Marshall, “Tips for Avoiding Bad Authentication Challenge Questions”, 1st ed. White Paper, 2007.

[13] A. Narayanan and V. Shmatikov, “Fast Dictionary Attacks on Passwords Using Time-Space Trade-off”, in CCS '05 Proceedings of the 12th ACM conference on Computer and communications security, New York, NY, USA, 2005, pp. 364-372.

[14] C. Goel and G. ARYA, “Hacking of Passwords in Windows Environment”, International Journal of Computer Science & Communication Networks, vol. 2(3), pp. 430-435, 2012.

[15] N. Adhikary, R. .Shrivastava, A. Kumar, S. Verma, M. Bag and V. Singh, “Battering Keyloggers and Screen Recording Software by Fabricating Passwords”, International Journal of Computer Network and Information Security, vol. 5, pp. 13-21, 2012.

[16] www.bloggingstocks.com, “headline-reports-ebay-hacked”, 2007. [Online]. Available: http://www.bloggingstocks.com. [Accessed: 17- May- 2015].

[17] www.grahakseva.com/complaints, “online fraud happened hacking my icici bank credit card”, 2013. [Online]. Available:http://www.grahakseva.com/complaints/130310/online-fraud-happened-hacking-my-icici-bank-credit-card. [Accessed: 01- May- 2015].

[18] www.foxnews.com, “World Bank Under Cyber Siege in Unprecedented Crisis", 2008. [Online]. Available:http://www.foxnews.com/story/2008/10/13/world-bank-under-cyber-siege-in-unprecedented-crisis/. [Accessed: 27- May- 2015].

[19] J. Vacca, *Computer and information security handbook*. Amsterdam: Elsevier, 2009.

[20] J.C. Liou and S. Bhashyam, “A feasible and cost effective two-factor authentication for online transactions.”, in *Software Engineering and Data Mining (SEDM), 2010 2nd International Conference on*, Chengdu, China, 2010, pp. 47 - 51.

[21] F. Cheng, “A Novel Rubbing Encryption Algorithm and the Implementation of a Web Based One-time Password Token.” in *Computer Software and Applications Conference (COMPSAC), 2010 IEEE 34th Annual*, Seoul, 2010, pp. 147 - 154.

[22] K. Alghathbar and H. A. Mahmoud, “Noisy Password Scheme: A New One Time Password System”, in *Electrical and Computer Engineering,*

2009. CCECE '09. Canadian Conference on, St. John's, NL, 2009, pp. 841 - 846.

[23] L. Soares, D. Fernandes, M. Freire and P. Inacio, "Secure user authentication in cloud computing management interfaces", in *Performance Computing and Communications Conference (IPCCC), 2013 IEEE 32nd International*, San Diego, CA, 2013, pp. 1-2.

[24] P. Thiyagarajan, V. Venkatesan and G. Aghila, "Anti-phishing technique using automated challenge response method", in *Communication and Computational Intelligence (INCOCCI), 2010 International Conference on*, Erode, 2010, pp. 585 - 590.

[25] M. Bond, O. Choudary, S. Murdoch, S. Skorobogatov and R. Anderson, "Chip and Skim: cloning EMV cards with the pre-play attack", in *Security and Privacy (SP), 2014 IEEE Symposium on*, San Jose, CA, 2014, pp. 49-64.

[26] K. Bicakci and N. Baykal, "Infinite Length Hash Chains and Their Applications", in *Enabling Technologies: Infrastructure for Collaborative Enterprises, 2002. WET ICE 2002. Proceedings. Eleventh IEEE International Workshops on*, Ankara, Turkey, 2002, pp. 57 - 61.

[27] V. Srivastava, A. Keshri, A. Roy, V. Chaurasiya and R. Gupta, "Advanced Port Knocking Authentication Scheme with QRC Using AES", in *Emerging Trends in Networks and Computer Communications (ETNCC), 2011 International Conference on*, Udaipur, 2011, pp. 159 - 163.

[28] B. Davaanaym, Y. Lee, H. Lee and S. Lee, "A Ping-Pong Based One-Time-Passwords Authentication System", in *INC, IMS and IDC, 2009. NCM '09. Fifth International Joint Conference on*, Seoul, 2009, pp. 574 - 579.

[29] F. Tao and S. Ping, "Design of Two-Way One-Time-Password Authentication Scheme Based on True Random Numbers", *Computer Science and Engineering, 2009. WCSE '09. Second International Workshop on*, vol. 1, pp. 11 - 14, 2009.

[30] Y. Chang, C. Chang and J. Kuo, "A secure one-time password authentication scheme using smart cards without limiting login times", *SIGOPSOper. Syst. Rev.*, vol. 38, no. 4, pp. 80-90, 2004.

[31] A. Castiglione, A. De Santis, A. Castiglione and F. Palmieri, "An Efficient and Transparent One-Time Authentication Protocol with Non-Interactive Key Scheduling and Update", in *Advanced Information Networking and Applications (AINA), 2014 IEEE 28th International Conference on*, Victoria, BC, 2014, pp. 351 - 358.

[32] L. Li and Y. Zhu, "Authentication Scheme for Substation Information Security Based on Chaotic Theory", in *Power and Energy Engineering Conference, 2009. APPEEC 2009. Asia-Pacific*, Wuhan, 2009, pp. 1 - 3.

[33] W. Hsieh and J. Leu, "Design of a Time and Location Based One-Time Password Authentication Scheme", in *Wireless Communications and Mobile Computing Conference (IWCMC), 2011 7th International*, Istanbul, 2011, pp. 201 - 206.

[34] S. Liao, Q. Zhang, C. Chen and Y. Dai, "A unidirectional one-time password authentication scheme without counter desynchronization", in *Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on (Volume: 4)*, Sanya, 2009, pp. 361 - 364.

[35] X. Ren and X. Wu, "A Novel Dynamic User Authentication Scheme", in *Communications and Information Technologies (ISCIT), 2012 International Symposium on*, Gold Coast, QLD, 2012, pp. 713 - 717**.**

[36] Z. Min-Qing, D. Bin and Y. Xiao-Yuan, "A New Self-Updating Hash Chain Structure Scheme", in *Computational Intelligence and Security, 2009. CIS '09. International Conference on (Volume: 2)*, Beijing, 2009, pp. 315 - 318.

[37] J. Jeong, M. Young Chung and H. Choo, "Integrated OTP-Based User Authentication and Access Control Scheme in Home Networks", in *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, Waikoloa, HI, 2008, pp. 294.

[38] M. Long and U. Blumentthal, "Manageable One-Time Password for Consumer Applications", in *Consumer Electronics, 2007. ICCE 2007. Digest of Technical Papers. International Conference on*, Las Vegas, NV, 2007, pp. 1 – 2.

[39] S. Hallsteinsen, I. Jørstad and D. Van Thanh, "Using the mobile phone as a security token for unified authentication", in *Systems and Networks Communications, 2007. ICSNC 2007. Second International Conference on*, Cap Esterel, 2007, p. 68.

[40] M. Alzomai and A. Josang, "The Mobile Phone as a Multi OTP Device Using Trusted Computing", in *Network and System Security (NSS), 2010 4th International Conference on*, Melbourne, VIC, 2010, pp. 75 – 82.

[41] M. Borowski andM. Lesniewicz, "Modern Usage of "Old" One Time Pad", in *Communications and Information Systems Conference (MCC), 2012 Military*, Gdansk, 2012, pp. 1 - 5.

[42] K. Kim, S. Jo, H. Lee and W. Ryu, "Implementation for federated Single Sign-on based on network identity", in *Networked Computing (INC), 2010 6th International Conference on*, Gyeongju, Korea (South), 2010, pp. 1 - 3.

[43] S. Shin, D. Kim and K. Yoo, "A Light-Weight Multi-User Authentication Scheme Based On Cellular Automata in Cloud Environment", in *Cloud Networking (CLOUDNET), 2012 IEEE 1st International Conference on*, Paris, France, 2012, pp. 176 - 178.

[44] K. Moon, D. Moon, J. Yoo and H. Cho, "Biometrics Information Protection Using Fuzzy Vault Scheme", in *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on*, Naples, 2012, pp. 124 - 128.

[45] S. Aboud, "Secure Password Authentication System Using Smart Card", *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, vol. 3, no. 1, pp. 75-79, 2014.

[46] M. David, G. Hussein and K. Sakurai, "Secure Identity Authentication and Logical Access Control for Airport Information Systems", in *Security Technology, 2003. Proceedings. IEEE 37th Annual 2003 International Carnahan Conference on*, 2003, pp. 314 - 320.

[47] S. Li, H. Jameel Asghar, J. Pieprzk, A. Sadeghi, R. Schmitz, and H. Wang, "On the Security of PAS (Predicate-Based Authentication Service)", in Computer Security Applications Conference, 2009. ACSAC '09. Annual, Honolulu, HI, 2009, pp. 209 - 218.

Psychosocial Correlates of Software Designers' Professional Aptitude

Walery Susłow, Jacek Kowalczyk, Michał Statkiewicz
Koszalin University of Technology
Koszalin, Poland

Marta Boińska, Janina Nowak
The University of Gdansk
Gdansk, Poland

***Abstract*—This paper presents quantitative results of the first phase of empirical research carried out within the framework of the interdisciplinary project InfoPsycho that was initiated in 2013 at the Koszalin University of Technology and the University of Gdansk. The aim of the study was to identify the personality traits that characterize successful applicants for university studies in the field of software development. Synthetic indicators of quality and performance of their design tasks and exercises were selected as the criteria for candidates' professional skills. To measure personality traits, the NEO-FFI questionnaire was used, based on the five-factor model by Costa and McCrae. Preliminary results show that expected young designers (N=140) score high on neuroticism and introversion as compared with those designers whose design documentation is of poor quality. They also show a high degree of conscientiousness, which can be seen when their performance of exercises and programming tasks is being evaluated.**

Keywords—Software designer; Software quality; Professional skills; Early diagnosis; Psychological tests; NEO-FFI

I. INTRODUCTION

One of the major challenges facing the education system in Poland is to adapt the university educational offer to the market needs. The number of software developers is constantly increasing while the demand is not diminishing but it is rather growing consistently and steadily. Hence, an important task seems to be to examine the psychological correlates of career aptitude in this profession. These correlates should be good indicators of future optimum efficiency of the designer and of their software products commercial success.

These questions have been present in literature for a relatively short period. Despite increased research into psychological factors in the field of computer science in many countries, we cannot detect this wider trend in Polish literature. This seems all the more surprising given that IT experts in the area of Central and Eastern Europe are considered to be one of the most successful software engineers in the world.

The result of developers' teamwork is software, which is usually ordered from them by a customer who possesses relatively little knowledge of programming languages or the latest solutions and trends that emerge in the field. An average customer has little experience in ordering software solutions. The end-users only have limited knowledge of what they would like to order without knowing the vast possibilities of how the ultimate product can be specifically designed to suit their needs. Thus, from a practical point of view, it is important that the software designer (the leading specialist in the development team) possesses high-level communication skills and specific personality traits that support the product creation process during which the customer is a powerful decision-maker. One could say that the software designer should possess both remarkable interpersonal and technical skills to perform their job effectively. This is an assumption based on an important observation that led us to undertake joint efforts with researchers in the area of social sciences and information technology with the aim of designing an interdisciplinary assessment programmers' advanced academic and workplace performance using various psychosocial variables. The research should be interdisciplinary in nature for two reasons: a professional evaluation of developers' suitability can only be performed by highly qualified specialists in the field of computer science, while an accurate assessment of psychosocial variables can be carried out only by psychologists who are willing to draw up and analyze opinion questionnaires.

Modern methods of software development transform a set of user requirements related to data processing into a set of instructions (a computer program within the desired framework). The common feature of these methods is that they are based on a creative activity using a design conception, not on a ready template. Hence, software quality is significantly influenced by designer's intuition and experience. By using the expressions used in literature on project typology already published [1], we can conclude that designer's personality traits will be absolutely critical in the analytical and design stage of the project (according to the criterion of the life cycle phases) as well as the primary and development stage of the project (according to the criterion of changes).

Developing IT projects is a very demanding job because the designer ought to be an expert in information technology, know which efficient and reliable methodology to adopt, and possess a good understanding of the specific problem domain. In practice, hard systems methodologies are used where the developer first comes up with a technical design and the whole responsibility is shifted onto the designers or programmers who develop and launch a product in close cooperation with the customer. They gain the necessary knowledge about the problem area and search for a comprehensive solution that is technically achievable and most user-friendly. The most dramatic development of the software systems takes place in small IT groups which have a fairly limited budget and an insufficient number of expert advisors. Unfortunately, these bad conditions often force the programmer to act not only as the architect of the system but also as a specialist in other areas. Similar problems occur in major projects that are poorly defined and coordinated, where the Open Source projects are a leading example.

For commercial software to be successful, it should be a high quality product. That is not a question of understanding the quality in terms of the degree of perfection as Plato defined it. Good software must be reliable, stable, efficient, safe, ergonomic, and portable. However, software is of outstanding quality when it is in accordance with the requirements imposed by the project framework and it meets all the user's needs and expectations. Thus, the buyer's satisfaction with the software, which takes place at the stage of deployment of the finished product, is in essence an evaluation of the product's quality. This forces designers to confront their own personal standards and ideas with the attitudes, needs and expectations of a broadly understood consumer.

An analysis of the specificity of the software as a market product leads to the conclusion that in order to help students excel in one specialized area of software design, the academic training programme should involve teaching the know-how about dealing with service clients and product end-users. It is commonly believed that the profession of the IT specialist involves only the technical background knowledge and necessary problem-solving skills. Indeed, programming skills and good knowledge of modern information technology, and hardware platforms are a firm foundation for this vocational career. But this is not enough; the designer needs a good "interface" as well as communication and cooperation skills to be able to work in a team and with a wide range of stakeholders. He should be willing to learn about innovations, to widen his perspective, and adapt his point of view to the ever-changing market trends. In other words, apart from being equipped with excellent technical knowledge, the software programmer should master humanistic skills (social, psychological, emotional). Unfortunately, this aspect of the vocational training of software designers is missing from the curricula of technical universities. As to some of the humanistic skills, it can be assumed that they cannot be effectively acquired by all students of computer science without guidance. What is now a common practice is that university teachers can only try to choose students according to their psychosocial aptitudes, and assign them specializations that seem to most suitable for their future career.

II. Review of Existing Literature

Research on psychosocial variables among software programmers makes use of a variety of theoretical models and psychometrics with different effects. The sections to follow provide a brief description of psychometric instruments and models, deliver the theoretical framework of the study, and then summarize the current results of research on the personality of programmers.

A. *Models of personality and their operationalization in research on software developers' personality traits*

Recently, there has been a great interest in testing and evaluating psychosocial variables, including personality, among the group of programmers [2-6]. The techniques used in these studies, which are very diverse, are the effect of an operationalization of a particular theoretical model that investigates personality [7-8]. There is a need to deepen and replicate research for the population of programmers.

The most common method of measuring personality traits in studies on programmers has been so far the Myers-Briggs Type Indicator (MBTI) [9-11]. It is based on Carl Gustav Jung's theory of psychological types of individuals. The MBTI questionnaire covers four bipolar factors or dimensions: extraversion-introversion (EI), sensing-intuition (SN), thinking-feeling (TF), and judging-perception (JP). An equally widely-used personality assessment was the Keirsey Temperament Sorter (KTS) questionnaire [12]. However, more recent studies show a shift to a five-factor model of personality known in short as Big-Five [13-14]. Psychologists know that, for testing purposes, the questionnaire to measure personality traits should be selected based on a number of factors, which have been defined by the methodologists of psychological research. The most important parameters of a good questionnaire are reliability and validity. The MBTI questionnaire, based on the concept of Jung's psychodynamic approach to personality, has been criticized despite its popularity because of the above-mentioned main parameters [15]. The results obtained by this method provided an obscure picture of programmers' personality. Hence, it is recommended to explore a different theoretical model and, consequently, other research techniques [16-18]. Due to the above-mentioned psychometric limitations of the MBTI questionnaire, in presented study the five-factor model of personality (PMO) designed by Paul Costa and Robert McCrae is employed. This model is used more frequently; what is more, it has repeatedly been tested in academic research and clinical trials [19]. It differentiates individuals in a significant way, which is widely regarded [20-22]. It represents a hierarchically well-organized personality inventory of five characteristics: extraversion, agreeableness, conscientiousness, openness and neuroticism. Here is a brief description of these traits:

1) Extraversion (E) indicates the degree to which a person is sociable, assertive and active in a conversation. An extrovert feels good in social relationships and derives pleasure from these.

2) Agreeableness (U) refers to individual characteristics such as kindness, trust and warmth. A person with a low level of agreeableness is described as selfish and full of doubts towards society.

3) Conscientiousness (S) deals with one's orientation to achievements. People who receive high scores are hardworking, reliable and organized enough to perform their tasks on time. On the other hand, low scores for conscientiousness are characteristic of impulsive, disordered, and irresponsible people.

4) Openness (O) describes the readiness of individuals to a wide range of intellectual and cultural activities. A person with high openness, one that has broad horizons, is ready to take risks to stimulate. At the opposite extreme, there is a person with low openness to experience, showing little sensitivity to aesthetic or cultural stimulation.

5) Neuroticism (N) alternatively described as an emotional stability [23], is often correlated with a sense of *self-efficacy [24]. People with low score are confident; feel safe, relatively stable in mood, calm. High neuroticism is an indicator of frequently changing moods, nervousness, anxiety and uncertainty [25].*

The rationale for popular PMO methods is lexical hypothesis, which is one of the key assumptions of the model. It points out that those five attributes (factors) described are contained in a natural language, one which we use in our country. These factors can be considered as the elements of our own "universal consciousness" as they concern the most adaptive characteristics and individual differences between people [26]. As any theoretical model, the PMO has its limitations and it has received some critical opinions, mainly related to the amount of factors that describe the personality. Nevertheless, the PMO is well studied in the academic and practical field, what makes this model of personality a standard choice of operationalization to numerous studies [27]. PMO is intended for testing many different tools, such as NEO-PI-R, NEO-FFI, and IPIP. NEO-FFI is an instrument which is used in this study due to its reliability and validity and due to its common usage to measure the personality in academic field [28-29]. Careful analysis of the literature allowed to make a decision to input PMO as a theoretical framework, which is operationalized in the form of the NEO-FFI instrument. The section to follow describes a review of the studies published, ones which focused on relations between the personality and the programmer's achievements.

B. *Research on computer programmers' personality and performance*

There are many studies in literature that test relations between the programmer's psychosocial qualities and the indicators of its operation. The main areas of studies are related to the issues of programming in pairs (PIP), as well as the efficiency of development teams (EMPA). Unfortunately, the results of these studies do not provide clear answers about the relationship between programmer's personality and the quality of his programming outcomes [15].

In this section three studies are presented where researchers were testing the relations of the personality of programmers and the efficiency of programming outcome in PIP. The results obtained do not always confirm the hypotheses. Only one study has fully demonstrated that in the process of programming in pairs a proper fit of personalities has resulted in the improvement of effectiveness [17]. In this publication, Choi states that if two people, without any previously acquired programming experience, are similar in terms of major personality traits or the MBTI complementary model (ST-SF, NT-NF, ST-NT, and SF-NF), their level of performance (in terms of quality of the software created) will be much higher than for other combinations. Another study [31] suggests that different levels of conscientiousness (understood as a personality trait) do not affect the academic success of paired programmers-students in the course of joint programming. Research [12], which was used in KTS, suggests that pairs consisting of a heterogeneous personality better fall within the parameters of the programming rather than a pair of the same type of personality. Most studies have provided inconsistent results, or ones that do not confirm the existence of any significant differences in terms of personality in PIP.

Equally often tested is how programmer team is managed and what influences the efficiency of team members. The most important study in this area is one where an attempt was made to determine the impact of personality on the results of team programming. Research has shown that teams can work in a satisfactory manner, despite significant differences in the members' personalities and ethnic and religious backgrounds [33]. On the other hand, there are the results showing that in fact there is a significant correlation between personality factors and the satisfaction of teamwork [32]. In contrast, Chao and Atli [10] studied the personalities of 60 respondents, and the data obtained by them did not allow to confirm the thesis that there is a difference in the quality of code between different matching types of personality in pair programming tasks. During the literature review, it is difficult to find accurate data on the personality of programmers in the context of individual work. Few studies reports show inconsistent results. Capretz L. F. [7, 10] shows that according to the classification of the Jung's MTBI model, among programmers there are 57% introverts vs. 43%, extroverts 81% minded vs. 19% sentient, 58% judgmental vs. 42% followers. The predominant set of traits is an introvert personality ISTJ (introversion, sensing, thinking, judgment): 24% for the entire sample of the respondent programmers, where studies on the general population of computer ISTJ reach only 11.6%. People of this type are "systematic, robust, practical, realistic approaching to reality, keeping promises, respecting the duties, valuing facts, well-organized, selfless" [34].

The question is whether such a set of traits is beneficial for the quality of the software developed. Cunha and Greathead [11] attempted to find an answer to this question by examining students in terms of the severity of the characteristics of the MBTI model. It turned out that people with higher results for the dimension of intuition performed significantly better in a task requiring a review of 282 lines of the Java code. In turn, Acun with colleagues indicates that extraversion is significantly associated with a better software quality in software development in an agile methodology [35].

III. Report on the Progress of the Infopsycho Project

A. *Concept and hypothesis*

The aim of the present study was to examine whether and how psychological predispositions (personality traits) of young software developers, recorded in the early stages of their professional education, are associated with an ability to develop good, user-friendly client software, one that fulfills the customer's requirements. Since the literature does not allow us to pose specific research hypotheses, this study is of an exploratory (pilot) nature. We pose the following research question: "What personality traits characterize young designers who are successful in the field of software at university level?" We have set two general hypotheses:

1) There is a relationship between personality traits and the quality of software created by a designer.

2) It is possible to indicate those personality traits that are important for the quality of design of software before a young designer commences their professional career.

To verify the abovementioned hypotheses, we tested two different age groups of students-informaticians taking into account different performance indicators of students' work on the software exercise mode and the design mode. To ensure the anonymity of the test, we used codes that allowed us to combine the results of individual indicators without identifying any individual participants registered.

We posed one additional research question, which was not followed by a hypothesis, because the scientific literature does not provide a sufficient suggestion for a probable answer: Do high-performing and low-performing IT students differ from non-technical faculty students in personality traits? We investigated personality traits of Pedagogy students and compared their results with IT students achieving the lowest and the highest outcomes in project quality scale in the second group.

B. Groups of participants

The group examined in the project InfoPsycho included 128 men and 12 women who were undergraduate and postgraduate students of Computer Science in the Department of Electronics and Computer Science at the Koszalin University of Technology. The structure of pilot studies carried out in the period from November 2013 to February 2014 assumed a division of the respondents into two separate groups. The first subgroup included 65 people (one person was later excluded because of missing some data), where men constituted 89.4%. They were second-year students actively involved in computer programming and information technology, but still not involved in design work. The average age for this subgroup was 22 years old. The second subgroup included 73 people (one person was later excluded because of missing some data), where 93.2% were men. This subgroup included students of the final year of undergraduate and postgraduate studies. They were all project contractors and, during the experiment, they performed the role of software designers, analysts, and software project managers. The average age in this subgroup was 24 years old.

C. Research techniques and tools

NEO-FFI Costa and McCrae questionnaire in the Polish adaptation by Zawadzki, Strelau, Szczepaniak and Sliwinska was used as a psychological tool [32], measuring five dimensions of personality in the Costa and McCrae concept: neuroticism, extroversion, openness to experience, agreeableness and conscientiousness. The questionnaire included 60 positions – 12 for each of 5 subscales. The test optional answers were presented on the five-point Likert's scale ranging from "strongly disagree" to "strongly agree".

The abovementioned personality dimensions were measured in the form of a self-report inventory. The results represent a full description of the personality of participants and the anticipation of their adaptability to the professional environment [32]. Time given to fill in the questionnaire was approximately 15 minutes.

Indicators that characterized the quality of design[1] done by the students were brought to a three-level rating scale; 1-2-3 is ranging from the worst to the best. When the exercise programming, the software system design, and the quality of communication in the project's team was judged, the expert approach was used; judgment was aided by checklist concerning all necessary elements of the students' work.

The general description of the specific nature of development work and the idea of quality of software are presented in the introduction to this article. Quality of the students' design was assessed with respect to a set of guidelines:

- compliance with formal requirements (adequate description, sequence of activities, presence of diagrams) − a critical guideline;
- syntactic correctness − a critical guideline;
- conceptual correctness (consistency of design and functionality of the designed system) − a critical guideline;
- efficiency and originality of created solutions;
- readability of the design documentation (describing and diagramming style, presence of glossary);
- history of the project (presence of schedule the number of patches).

It is necessary that the critical guidelines are met. The degree to which students follow the guidelines determines the final result of the assessment. However, when assessing the programming exercises, we have taken into account the following aspects of the students' work:

- correctness of the realized task (solution to the problem in the form of a compilable source code) − a critical guideline;
- good understanding and correct analysis of the code − a critical guideline;
- use of good programming practices (object-oriented paradigms, readability, error handling);
- efficiency of technical solution.

Students' activity assessment was based on active participation and involvement in meetings with supervisors and student coordinators when students' solutions and decisions were consulted. The parameter of the periodic activity of each project group was calculated in proportion to the number of meetings and of the closed tasks, and then a numerical value of the range of 1-2-3 was assigned.

D. Results

The results are presented in two main stages. First, we show correlation coefficients between two groups: the grade,

[1] "Level of effectiveness of the design function in determining a product's operational requirements (and their incorporation into design requirements) that can be converted into a finished product in a production process", definition from http://www.businessdictionary.com.

activity during the lessons, and personality traits in the first subgroup of IT students and quality of design, communication quality, and personality traits in the second subgroup of IT students. Second, we compare three set of data: high-performing IT students from the second subgroup, low-performing IT students from the second subgroup, and Pedagogy students. In both stages, the alpha level of 0.05 was applied. Statistical analyses were conducted with IBM SPSS Statistics 21.

Table I presents the means, standard deviations, and correlation coefficients of the grade, activity during lessons, and the personality traits of the first subgroup of IT students. We use two different correlation coefficients to explore relationships between different types of scales: Kendall’s tau-b for personality traits (interval scale), grade (ordinal scale) and activity during the lessons (ordinal scale); Spearman’s Rho for grade (ordinal scale) and activity during lessons (ordinal scale).

TABLE I. Means, Standard Deviations, and Correlation Coefficients between the Grade, Activity During Lessons, and Personality, N=65

	M	**SD**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. Grade			1.00	0.70**	-0.05	0.38**	-0.16	0.99	0.01
2. Activity during lessons				1.00	-0.07	0.21*	0.01	0.00	0.10
3. Extraversion	39.51	7.41			1.00	0.16*	-0.29**	0.23**	0.09
4. Conscientiousness	42.28	7.95				1.00	-0.18*	0.13	0.01
5. Neuroticism	32.05	9.18					1.00	-0.20*	0.03
6. Openness to experience	38.57	6.87						1.00	-0.14
7. Agreeableness	38.68	4.52							1.00

* p < 0.05 ** p < 0.01

The grade correlates positively with conscientiousness (r = 0.38, p < 0.01). Also, activity during lessons correlates positively with conscientiousness (r = 0.21, p < 0.05). The grade correlates positively with the activity during lessons (r = 0.70, p < 0.01), which is consistent with the evaluation criterion: the common variance is the result of the number of correct answers given by each student. This is included in both variables.

The quality of design correlates positively with neuroticism (r = 0.21, p < 0.05) and negatively with extraversion (r = -0.25, p < 0.01). The communication quality correlates positively with agreeableness (r = 0.27, p < 0.01) and negatively with extraversion (r = -0.19, p < 0.05).

The remaining correlation coefficients failed to gain statistical significance. The quality of design correlates positively with the communication quality (r = 0.60, p < 0.01).

Table II presents the means, standard deviations, and correlation coefficients of the quality of design, the communication quality, and the personality traits of the second group of IT students. Similarly to the first group, we use two different correlation coefficients to explore the relationships between different types of scales: Kendall’s tau-b for personality traits (interval scale), the quality of design (ordinal scale), and communication quality (ordinal scale); Spearman’s Rho for the quality of design (ordinal scale) and the communication quality (ordinal scale).

In the next step, we compare high-performing IT students from the second subgroup, low-performing IT students from the second subgroup, and Pedagogy students. The groups do not differ significantly in age and gender. Means and standard deviations in personality traits in all three groups are presented in Table III.

TABLE II. Means, Standard Deviations, and Correlation Coefficients between Quality of Design, Communication Quality, and Personality, N=73

	M	**SD**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. Quality of design			1.00	0.60**	- 0.25**	0.08	0.21*	0.04	0.04
2. Communication quality				1.00	- 0.19*	0.12	0.11	0.03	0.27**
3. Extraversion	39.70	6.78			1.00	0.21**	- 0.29**	0.05	0.12
4. Conscientiousness	44.21	8.05				1.00	- 0.25**	0.14*	0.22**
5. Neuroticism	29.16	7.95					1.00	0.50	-0.13
6. Openness to experience	37.97	6.03						1.00	-0.00
7. Agreeableness	45.14	6.89							1.00

* p < 0.05 ** p < 0.01

TABLE III. Means and Standard Deviations in Personality Traits in Pedagogy Students, High-Performing IT Students, and Low-Performing IT Students

	Pedagogy students, N = 30		**High-performing IT students, N=31**		**Low-performing IT students, N = 18**	
	M	SD	M	SD	M	SD
Extraversion	42.50	7.09	38.13	6.94	42.78	5.91
Neuroticism	31.17	8.52	30.74	8.52	25.22	6.22
Conscientiousness	45.00	6.01	45.42	6.82	42.53	8.40
Agreeableness	40.83	5.72	42.00	6.03	42.33	5.51
Openness to Experience	37.63	6.22	37.96	6.40	37.33	5.85

A one-way comparison between ANOVA subjects was drawn to compare the personality traits of three groups of students. There is a significant difference between the groups in levels of extraversion [F(2, 76) = 4.11, p = 0.020], n2 = 0.43 and neuroticism [F(2, 76) = 3.54, p = 0.034], n2 = 0.39. Remaining comparisons are insignificant: conscientiousness [F(2, 76) = 1.03, p = 0.361], agreeableness [F(2, 76) = 0.479, p = 0.621], openness to experience [F(2, 76) = 0.062, p = 0.940]. Post hoc comparisons using the Tukey HSD test for neuroticism indicate that the mean score of the low-performing IT students is of minor significance as compared to the high performing IT students (I-J = -5.7, p = 0.054) and is significantly different than that of Pedagogy students (I-J = -5.94, p = 0.044). However, the high-performing IT students do not significantly differ from Pedagogy students (I-J = -0.23, p = 0.993). Post hoc comparisons using the Tukey HSD test for extraversion indicate that the mean score of the high-performing IT students is of a borderline significance with the low-performing IT students (I–J = -4.65, p = 0.060) and is significantly different than that of Pedagogy students (I–J = -4.37, p = 0.037). However, the low-performing IT students do not significantly differ from the Pedagogy students (I-J = 0.28, p = 0.990).

All in all, these results suggest that the high-performing IT students have a similar level of neuroticism and a lower level of extraversion than the Pedagogy students, whereas the low-performing IT students have a lower level of neuroticism and a similar level of extraversion as Pedagogy students.

E. Discussion of research findings

The results of the study confirm the first hypothesis. Achieving good grades in informatics correlates with specific dimensions of personality. A positive correlation was observed between the indicators of the quality of design and the level of neuroticism and introversion in terms of Costa and McCrae. Students with the highest degree of efficiency in the area of software design are more neurotic (5 sten) than students with a lower degree (3 sten). Differences become significant in intragroup comparisons; however, all the results fit in the given age norms. Intragroup differences in extraversion were also observed: students getting better results in quality assessment and communication were less extraverted (5 sten) than students getting worse results (6 sten). Students with the highest degree of efficiency in the area of software design are less emotionally stable than less predictive students in this field. Differences in extraversion were also observed: students getting better results in the assessment of the quality of design and communication were less extravert (5 sten) than students getting worse results (6 sten).

Many previous studies have shown that the level of neuroticism is correlated with the experience of negative emotions and emotional instability [33-35]. But higher levels of neuroticism also show that the individual is able to think in an unconventional way and comes up with very innovative ideas [36]. This seems especially desirable for future software designers whose significant advantage is the ability to "break" the generally applicable schemes (where appropriate; however, not ignoring reliable methodologies and solutions), even if such thinking seems at first incomprehensible or even illogical [37]. Much attention has been devoted to neuroticism by Karen Horney [38] who described a model of neurotic competition. The components of such neurotic competition are as follows: always comparing oneself with others, the desire to be unique, and an element of hostility. The best work results and most original ideas very often are produced when one works in solitude and avoids the distraction of other people. People who prefer to work individually than work in a team are often deeply convinced that only few will succeed in a given field, therefore they steer clear of any collective work which can be even harmful. Neurotic independence, mentioned above, should be considered as a trait of a future employee in their workplace. On the one hand, a competitive individual will get the project implemented quickly and efficiently, but on the other hand, he can, for the same reason, have an opinion of a difficult employee. Salgado meta-analysis [39], where counterproductive behavior was the focal point, showed that emotional stability, which is the opposite of neuroticism, is associated with a lack of rotation in the workplace.

The combination of the results mentioned above, related to neuroticism and introversion among future successful programmers, is consistent with similar studies carried out in this convention [40]. This clearly suggests that IT specialists are more timid in their dealings with other people. Human relations for people with a high degree of introversion are strenuous, and therefore they spend more time working on their own. They prefer solitude and seem to have less need for the entertainment and fun that is often associated with working with peer colleagues. Each of their statement is scrupulously analyzed before it is said out loud [41]. It should be noted that our target group in this study were students who are not sufficiently aware of the significance of interpersonal relationships. The curriculum of computer science studies does not include any activities aimed at developing soft (social) skills. The academic work of these students is evaluated by other technical specialists, which results in students functioning in a rather hermetic environment of the university campus.

One of the qualities that is conducive to developing essential communication skills is a high level of agreeableness. This is in line with the previous research results. The results achieved suggest that individuals with a high level of agreeableness are able to work with others and can compromise [42]. They are also empathetic, friendly, and sympathetic [43]. Agreeableness components include dimensions such as straightforwardness, altruism, submissiveness, humility, a tendency of self-pity [44]. People with a higher level of agreeableness have a natural tendency to neutralize conflicts and emphasize the benefits that stem from functioning in a group [47]. It is also worth mentioning that people with high scores in this area are able to control their anger, which can also have a temperamental basis and foster proper communication [45].

A correlation between conscientiousness and the assessment of activities is consistent with studies which indicate that academic achievements are positively related to this exact personality dimension. Furthermore, a significant amount of research demonstrates that high levels of conscientiousness allow one to predict the functioning of the individual in the workplace [46-47]. Such a person is perceived as responsible, persistent in pursuing goals, thoughtful in planning and undertaking new tasks. Low conscientiousness is associated with a tendency to do activities which may be described as procrastinative [48]. Conscientious employees are more credible, more motivated, they also have lower rates of absence and are less prone to harmful behavior at work, such as theft and aggressive behavior towards other workers [49].

Another issue, presented in the study, was a comparison of the students of Pedagogy with future software designers. The results showed that promising designers are characterized by a similar degree of neuroticism as the students of Pedagogy, while future IT specialists with a low degree achieve significantly poorer results in this area. This result leads to conclusion that a lower level of neuroticism is detrimental to the quality of their design. More neurotic individuals are more likely to interpret their mood adequately, analyze themselves, and look critically at reality, which may be associated with a lower self-esteem and a sense of helplessness. They are aware that improving their mood must have a real basis and is not the result of wishful thinking or the actual needs to be happy. Students of Pedagogy, who wish to follow the career of teachers and educators, are trained in the tasks of teaching others the values and traditions, as well as attitudes which tend to change over time. They learn how to teach openness and tolerance of cultural and individual differences which are of utmost importance in today’s world. This valuable training leaves them less opinionated and more accepting of others.

In addition, a more realistic and critical view of the reality can be as well a good side of those individuals whose work is based on tedious and time-consuming programming, which depends on the final result. Such job often involves revision, reorganization of work, or verification of the details that are unfamiliar (mysterious) to others; it can lead to a short-term or long-term reduction in self-esteem; the individual will be determined to do his job properly and accurately.

What is more, promising software designers are characterized by a lower degree of extraversion than those with lower rates, or than students of Pedagogy. The pedagogue's future job, independently of the specialization completed, is associated with numerous interpersonal relationships. Pedagogical studies are focused on the development of social competence. Therefore, a high level of extraversion is the key to their career success. However, it turns out that a high level of extraversion, which for pedagogues can be an important indicator of their suitability for the profession, for software designers is much less of a desirable trait and can even impair the quality of their design. As mentioned earlier, promising software designers prefer to operate on their own, away from various distractors. When one functions alone, work is more predictable and under control. If software needs to be fixed, then only the designer can fix it quickly and safely. His work results are verified on regular basis. Such an operating mode seems to be particularly desirable in two cases: when working on individual IT projects, as well as when doing work in the remote mode.

To sum up, the analysis above suggests that promising software designers are characterized by specific personality traits. They are more introvert than Pedagogy students and future IT specialists who are not considered to be the promising ones. They also score lower on neuroticism than future pedagogues, but higher than students who are not likely to specialize in the area of software design.

It should be emphasized that our study was of a pilot nature. Relatively small groups were a limitation to the study and, because of the university's educational process, these groups were broken down according to different evaluation criteria. In the next stages, our team will aim to set the evaluation criteria for the performance of IT specialists as accurately as possible. In addition to this, we will research more into the personality characteristics found in the dissertations of Costa and McCrae’s on neuroticism, conscientiousness, and extraversion. This will allow us to build a more complete and detailed model of efficient programmer’s personality. Apart from this, it will be necessary to expand the research to include also those designers who work with software professionally. This will provide evidence and be a starting point for discussion on how personality characteristics indicated by us, can be also diagnostic in the work environment.

IV. Conclusions

The results obtained in the pilot study indicate directly the existence of psychosocial correlates of professional predispositions of software developers. Following a preliminary research, it is already obvious how further research should be designed to lay a methodological basis for the practical use of psychological tests. As explained in previous passages, psychological tests support the learning process at the faculty of Computer Science. One would expect the implementation of the dynamic specializations’ information management at faculties based on the commitment to the professional aptitude, to have already been discovered by the students of the second year, to their preferred roles in industrial design teams. In fact, this requires appropriate planning expertise based solely on the mating segments of the software market or platform implementation. The ability of a pro-active psychological profiling of vocational students (dynamic specializations), based on the forecast of the demand for IT professionals, can become answer to the need of adapting the educational offer of higher education to the market needs. A practical implication of the study would improve the communication skills of students-specialists. One aspect to consider is a possibility of a more efficient use of teaching hours planned in the humanities module.

References

[1] S. Wrycza, Informatyka ekonomiczna. PWE, Warszawa 2010, s. 345.

[2] M. V. Kosti, R. Feldt, L. Angelis, Personality, emotional intelligence and work preferences in software engineering: An empirical study. Information and Software Technology, 56 (8): 973–990, 2014.

[3] R. Feldt, R. Torkar, L. Angelis, M. Samuelsson, Towards individualized software engineering: empirical studies should collect psychometrics. In: Proceedings of the 2008 International Workshop on Cooperative and Human Aspects of Software Engineering, pp. 49–52, ACM, May 2008.

[4] J. E. Hannay, E. Arisholm, H. Engvik, D. I. Sjoberg, Effects of personality on pair programming. IEEE Trans. Softw. Eng., 36 (1): 61–80, 2010.

[5] R. Feldt, L. Angelis, R. Torkar, M. Samuelsson, Links between the personalities, views and attitudes of software engineers. Inf. Softw. Technol., 52 (6): 611–624, 2010.

[6] L. G. Martínez, G. Licea, A. Rodríguez, J. R. Castro, O. Castillo, Using MatLab's fuzzy logic toolbox to create an application for RAMSET in software engineering courses. Computer Applications in Engineering Education, 21(4): 596–605, December 2013.

[7] L. F. Capretz, Personality types in software engineering. International Journal of Human–Computer Studies, 58 (2): 207–214, 2003.

[8] N. Salleh, E. Mendes, J. Grundy, Investigating the effects of personality traits on pair programming in a higher education setting through a family of experiments. Empirical Software Engineering, 19 (3):714–752, Jun 2014.

[9] N. Gorla, Y. W. Lam, Who should work with whom? Building effective software project teams. Communications of the ACM, 47(6): 79–82, 2004.

[10] J. Chao, G. Atli, Critical Personality Traits In Successful Pair Programming. In: Proceedings of Agile 2006 Conference, pp. 89–93, 2006.

[11] A. D. D. Cunha, D. Greathead, Does personality matter? An analysis of code-review ability. Communications of the ACM, 50(5): 109–112, 2007.

[12] P. Sfetsos, I. Stamelos, L. Angelis, I. Deligiannis, Investigating the impact of personality types on communication and collaboration-viability in pair programming – an empirical study. Proceedings of the 7th International Conference on Extreme Programming and Agile Processes in Software Engineering (XP 2006), pp. 43–52, 2006.

[13] G. J. Boyle, Myers-Briggs type indicator (MBTI): some psychometric limitations. Australian Psychologist, 30(1): 71–74, 1995.

[14] K. S. Choi, A discovery and analysis of influencing factors of pair programming. Unpublished Ph.D. Dissertation, New Jersey Institute of Technology, USA 2004.

[15] K. S. Choi, F. P. Deek, Im I., Exploring the underlying aspects of pair programming: the impact of personality. Information and Software Technology, 50(11): 1114–1126, 2008.

[16] N. Salleh, E. Mendes, J. Grundy, Empirical studies of pair programming for CS/SE teaching in higher education: A systematic literature review. IEEE Trans Software Eng, 37(4): 509–525, 2011.

[17] G. S. J. Burch, N. Anderson, Personality as a predictor of work-related behavior and performance: recent advances and directions for future research. In: Hodgkinson G. P., Ford J. K. (eds) International review of industrial and organizational psychology. Wiley UK, pp. 261–305, 2008.

[18] M. R. Barrick, M. K. Mount, The big five personality dimensions and job performance: a meta-analysis. Personality Psychology, 44: 1–26, 1991.

[19] P. Sfetsos, I. Stamelos, L. Angelis, I. Deligiannis, An experimental investigation of personality types impact on pair effectiveness in pair programming. Empir. Softw. Eng., 14 (2): 187–226, 2009.

[20] K. Lee, M. C. Ashton, Psychopathy, Machiavellianism, and narcissism in the Five-Factor Model and the HEXACO model of personality structure. Personality Individ. Differ., 38 (7): 1571–1582, 2005.

[21] R. R. McCrae, O. P. John, An introduction to the five-factor model and its applications. Journal of Personality, 60 (2): 175–215, 1992.

[22] N. Schmitt, The interaction of neuroticism and gender and its impact on self-efficacy and performance. Human Performance, 21: 49–61, 2008.

[23] J. E. Driskell, E. Salas, F. F. Goodwin, P. G. O'Shea, What makes a good team player? Personality and team effectiveness. Group Dynamics: Theory, Research, and Practice, 10(4): 249–271, 2006.

[24] A. L. Pervin, Psychologia osobowości. GWP, Gdańsk 2002, s. 64.

[25] M. A. Conard, Aptitude is not enough: how personality and behavior predict academic performance. J Res Pers, 40: 339–346, 2006.

[26] K. Matzler, B. Renzl, J. Muller, S. Herting, T. A. Mooradian, Personality traits and knowledge sharing. J Econ Psychol, 29: 301–313, 2008.

[27] B. De Raad, H. C. Schouwenburg, Personality in learning and education: a review. Eur J Pers, 10: 303–336, 1996.

[28] N. Salleh, E. Mendes, J. Grundy, Investigating the effects of personality traits on pair programming in a higher education setting through a family of experiments. Empir Software Eng, 19: 714–752, 2014.

[29] J. Karn, T. Cowling, A follow up study of the effect of personality on the performance of software engineering teams. In: Proceedings of the 2006 ACM/IEEE International Symposium on Empirical Software Engineering, ACM, pp. 232–241, 2006.

[30] S. T. Acuńa, M. Gómez, N. Juristo, How do personality, team processes and task characteristics relate to job satisfaction and software quality?, Inf Softw.Technol., 51 (3): 627–639, 2009.

[31] G. H. Grabowska, Bliskość emocjonalna w tworzeniu zespołów projektowych. Zeszyty Naukowe Wydziału Informatycznych Technik Zarządzania Wyższej Szkoły Informatyki Stosowanej i Zarządzania „Współczesne Problemy Zarządzania”, 1, 2013.

[32] B. Zawadzki, J. Strelau, P. Szczepaniak, M. Śliwińska, Inwentarz osobowości NEO-FFI Costy i McCrae. Pracownia Testów Psychologicznych PTP, Warszawa 1998, s. 7–34.

[33] E. Diener, E. M. Suh, R. E. Lucas, H. L. Smith, Subjective well-being: Three decades of progress, Psychological Bulletin, 125: 276–302, 1999.

[34] J. Suls, P. Green, S. Hillis, Emotional reactivity to everyday problems, affective inertia, and neuroticism. Personality and Social Psychology Bulletin, 24: 127–136, 1998.

[35] A. Lee, R. O. Pihl, Prefrontal cognitive ability, intelligence, Big Five personality and the prediction of advanced academic and workplace performance. Journal of Personality and Social Psychology, 93: 298–319, 2007.

[36] S. Rothmann, E. P. Coetzer, The big five personality dimensions and job performance, SA Journal of Industrial Psychology, 29(1): 68–74, 2003.

[37] M. Tamir, M. D. Robinson, Knowing good from bad: The paradox of neuroticism, negative affect, and evaluative processing. Journal of Personality and Social Psychology, 87: 913–935, 2004.

[38] K. Horney, Neurotyczne osobowości naszych czasów. Rebis, Poznań 2011, s. 62–68.

[39] J. F. Salgado, The big five personality dimensions and counterproductive behaviours. International Journal of Selection and Assessment, 10(1/2): 117–125, 2002.

[40] C. G. Cegielski, D. J. Hall, What makes a good programmer? Comm. ACM, 49 (10): 73–75, Oct. 2006.

[41] M. O. Laney, The Introvert Advantage: How to Thrive in an Extrovert World. NY: Workman Publishing, 2002.

[42] J. M. Digman, N. K. Takemoto-Chock, Factors in the Natural Language of Personality: Re-analysis, Comparison, and Interpretation of Six Major Studies, Multivariate Behavioral Research, 16: 149–170, 1981.

[43] R. Hogan, Socioanalytic theory of personality. In M. M. Page (Ed.), 1982 Nebraska Symposium on Motivation: Personality—current theory and research, pp.55–89. Lincoln: University of Nebraska Press, 1983.

[44] P. T. Jr. Costa, R. R. McCrae, Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people. Journal of Personality and Social Psychology, 38: 668–678, 1980.

[45] M. K. Rothbart, J. E. Bates, Temperament. In W. Damon (Series Ed.), N. Eisenberg (Vol. Ed.), Handbook of child psychology: Vol. 3. Social, emotional and personality development, (5th Ed), New York: Wiley, 1998, pp. 105–176.

[46] E. K. Gray, D. Watson, General and specific traits of personality and their relation to sleep and academic performance, Journal of Personality, 70: 177–206, 2002.

[47] D. M. Higgins, J. B. Peterson, A. Lee, R. O. Pihl, Prefrontal cognitive ability, intelligence, Big Five personality and the prediction of advanced academic and workplace performance, Journal of Personality and Social Psychology, 93: 298–319, 2007.

[48] S. Dewitt, H. C. Schouwenburg, Procrastination, temptations, and incentives: The struggle between the present and the future in procrastinators and the punctual. European Journal of Personality, 16 (6): 469–489, 2002.

[49] M. K. Mount, M. R. Barrick, G. L. Stewart, Five-factor model of personality and performance in jobs involving interpersonal interactions. Human Performance, 11 (2): 145–165, 1998.

Map Reduce: A Survey Paper on Recent Expansion

Shafali Agarwal
JSS Academy of Technical Education, Noida, 201301, India

Zeba Khanam
JSS Academy of Technical Education, Noida, 201301, India

***Abstract*—A rapid growth of data in recent time, Industries and academia required an intelligent data analysis tool that would be helpful to satisfy the need to analysis a huge amount of data. MapReduce framework is basically designed to compute data intensive applications to support effective decision making. Since its introduction, remarkable research efforts have been put to make it more familiar to the users subsequently utilized to support the execution of massive data intensive applications.**

Our survey paper emphasizes the state of the art in improving the performance of various applications using recent MapReduce models and how it is useful to process large scale dataset. A comparative study of given models corresponds to Apache Hadoop and Phoenix will be discussed primarily based on execution time and fault tolerance. At the end, a high-level discussion will be done about the enhancement of the MapReduce computation in specific problem area such as Iterative computation, continuous query processing, hybrid database etc.

Keywords—Map Reduce; Hadoop; Iterative Computation; Phoenix; Databases

I. Introduction

In the present days, a voluminous data handling is a prime concern topic for researchers. Many applications like data mining, Image processing, data analytic etc are required processing of huge amount of data. In 2004, Google [1] had invented a MapReduce framework suitable for parallel data processing in distributed computing environment. MapReduce is a processing paradigm of executing data with partitioning and aggregation of intermediate results. It works to process data in parallel in which splitting of data, distribution, synchronization and fault tolerance are handled automatically by the framework. Map reduce framework is famous for large scale data processing and analysis of voluminous datasets in clusters of machines.

A MapReduce framework can be categorized into mainly two steps such as [2]:

Map Phase:

- Initially split the data into key value pair and fed into mapper which in turn process each key value pair and generate intermediate output.

Reduce Phase:

- The Intermediate key value pair first collected, sorted and grouped by key and generate values associated with each key.
- The receiver produces final output based on some calculation and stores it in an output file.

Despite being featured such as scalability in clusters, ensuring availability, handling failures Google's MapReduce has been unusable for certain kind of applications requires iterative computation, execution of high-level language such as SQL and work on an Internet desktop grid. Since the MapReduce introduced, numerous MapRduce frameworks have been developed by several companies including Google's MapReduce [1], Apache's Hadoop MapReduce [3], AMPLab's spark [4], SASReduce [5], Disco [6] etc. A lot of research has been done to address the issues highlighted above and some recent MapReduce implementation helps to overcome the limitations of the prior framework. While we consider databases, an author described salient features of MapReduce implementation and its performance comparison with the parallel database. According to report, MapReduce works well in different storage systems and provide a good framework to fault tolerance for large jobs [7].

Initially the paper describes the MapReduce classification as well as an introductory explanation of its applications such as distributed pattern based searching, geospatial query processing, web link graph traversal, distributed sort, machine learning applications etc. The primary focus of this survey paper is to highlight some MapReduce implementation worked well to accomplish a specific purpose and compared with previously available frameworks. A remarkable performance improvement over the existing system seems after comparison. Later we discussed the recent enhancements which help to solve the issues related to iterative computation, efficient continuous queries execution and hybrid database.

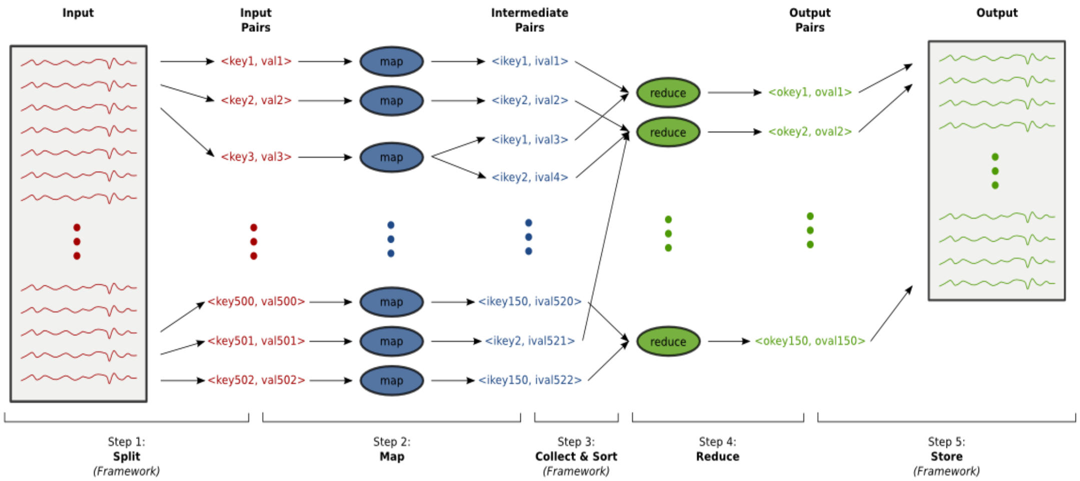


Fig. 1. Map Reduce framework

II. Map Reduce Classification

Map Reduce data analytic applications are categorized on the basis of their functions [8]:

A. *Clustering based algorithm*

These algorithms are memory sensitive as cluster-based algorithm required a large amount of storage. To measure the parameter values of multiple clusters, a massive computation makes it compute intensive method. for eg. K-means, Fuzzy K-means, canopy clustering etc.

B. *Classification Algorithm*

This algorithm works on a training set and query set to compute k nearest values which required a sufficient memory space to store the data. It is also compute intensive method because a vector product is carried out to calculate the similarity between two vectors. For eg. K-nearest neighbor etc

Author [9] analyzed the different mechanism to improve the memory utilization on the multi-core machine for MapReduce. Author had also explored three given applications with respect to efficient memory utilization.

1) Hash Join- It is a variant of broadcast join by Blanas et al [10]. In the join operation, only Map function is used to join two tables i.e. data table (S) and reference table (R). A hash join is not compute intensive application and its time complexity is O(|S|).

2) KMeans- K-means application is used to partition a set of n sample objects into K clusters for input parameter K. This algorithm is memory intensive and compute-intensive which in turn limiting the number of clusters K-means can generate. The time complexity is O(|n||k|).*

3) K-nearest neighbors- K-nearest neighbors is a classification algorithm that uses a large in-memory data set. KNN method uses two data sets, a query set Q and a training set T. It chooses K closet elements in T based on a computed distance between data points in both sets. The time complexity of the method is O(|Q||T|) because it calculates the distance between every point in Q and in T. So the KNN is compute intensive as well as memory intensive application.*

III. Map Reduce Applications

Map Reduce implementation is used in various data intensive computation because of the functionality of parallel processing of massive data. A short introduction of related applications is given below:

A. *Distributed pattern based searching*

Distributed grep command is used to search a pattern in the given text distributed over a network. Here map function searches for the pattern and produces the output so no intermediate result writes. Hence reduce function is just copied the intermediate result to output in distributed pattern searching [1].

Example: A big data of medical health record is analyzed using parallelization and pattern searching property of MapReduce taking into consideration [11]:

1) Public dataset- It consists of various reports of patients from US Food and Drug administration.

2) Biometric Datasets- It is having human characteristics like images [12].

3) Bioinformatics Signal datasets- This dataset represents the recording of vital signs of a patient. e.g. Electrocardiography ECG

4) Biomedical Image datasets- A dataset having a collection of scanning of medical images such as ultrasound images.

B. *Geospatial Query Processing*

With the technological advancement in location based service, MapReduce helps to find out the shortest path in Google map for a given location. Here Map function searches all connecting paths from source to destination with distance value. After sorting the keys, the Reduce function emits the path which is of shortest distance.

An algorithm LoNARS [13] has implemented to improve Reduce task scheduling by considering data locality and network traffic. Even author achieved 15% gain in data shuffling time and up to 3-4% improvement in job completion time.

C. *Distributed Sort*

Distributed sort is used to arrange the data in sorted manner split across multiple sites. In Map Reduce implementation, initially input data is given to map function to convert it into intermediate data which is stored in a local disk buffer. In next step, data is transmitted to the appropriate reducer function over the network. A number of reduce functions sort the data according to given key value and writes the output [14].

Author represents massive data sorting using Apache Hadoop open source software framework with the help of three map reduce functions [15]:

- Teragen: used to generate input data to be sort.
- Terasort: Sample the input data and used them with Map Reduce to sort the data.
- Teravalidate: At last sorted output data is validated.

This method is I/O intensive as it works on data input/output.

D. *Web Link Graph Traversal*

A large-scale graph is also known as web graph. For eg. According to a survey Facebook is having more than 1 billions of users (vertices) and more than 140 billions of relationships (edges) among them in 2012 [16].

Basically Map Reduce model is not suitable for iterative data analysis application that’s why it is assumed to be inadequate for graph traversal. In order to accomplish large scale graph processing, Surfer and GBASE are used as an extension of Map Reduce that are proposed to make it suitable for graph processing.

Surfer- Surfer is an engine used in graph processing. It works with two components i.e. Map Reduce and propagation. Map Reduce processes data parallel in terms of key/value pair whereas propagation is an iterative computational pattern that propagate data from a vertex to its neighbors in the graph.

GBASE- GBASE [17] executes block compression to store homogeneous region of the graph. When a graph traversal query is fired, GBASE selects the grid having a block that is relevant to query. Therefore only relevant required data is fed into Hadoop jobs.

E. *Term Vector per Host*

This term refers to summarize the important words of a document or multiple documents. A map function finds out the term vector for a particular host name as (host name, term vector) pair and pass this data to reduce function for a given host. Now reduce function add these term vectors and produces a final output in terms of (host name, term vector) [18].

F. *Machine Learning Applications*

Machine learning is a branch of artificial Intelligence which deals with the building of systems that learn from data without need of explicit programming for all the possible conditions.

Author [19] discussed the case of Netflix prize data which is an online DVD rental company. Netflix wants to predict the user preferences of movies based on their rating. In order to get the data map function is used to generate a table which contains information regarding users and their movie preferences. After completing this process, reduce function derive a contingency table for each group of intermediate results depicts user preferences about movies.

G. *Data Clustering*

Data clustering is a fascinating field for researchers involved in Image processing, data mining and document retrieval area. Data clustering is used to solve the computational complexity arises due to the voluminous data used in processing by dividing complete data set into small data subsets based on certain criteria.

Author [20] used parallel K-means clustering using map reduce to minimize the efforts make to handle a large data sets. A key feature of this algorithm is the use of combiner used to partially combine the intermediate values of map function with the same key.

H. *Inverted Index*

It is an index data structure storing mapping from contents such as words or numbers to its locations in the database file or in a document. Inverted Index is used in data retrieval in a large database management system. This process receives a list of document as input and produces word to document indexing. Alternatively it is used to track the position of words in a given document.

A map function parsed each document and retrieved its document Id with the word. Later reduce function accepts all pair of given words and emits corresponding word with its list of relevant documents. Hence complete output pairs represent an Inverted Index of the database.

IV. Map Reduce Models and their Comparison

A. *Hadoop vs. Phoenics++*

Many map reduce implementations have been discussed in the previous section. From which Hadoop is given by Apache and support distributed memory clusters. Similarly phoenix++ works on shared memory multicore systems [21]. Author [22] compared the performances for word count problem running on Amazon elastic compute cloud (Amazon EC2) of both systems and concluded that phoenix++ is superior to Hadoop in terms of execution time. According to him phoenix++ is faster than Hadoop by 28:5 on four virtual CPUs for 7.4 seconds versus 211 seconds.

B. *Phoenix vs. Phoenix2*

Phoenix was introduced as a Map Reduce model which can work on shared memory machine and symmetric multiprocessors with scalability [23][24]. This model was not appropriate for many types of workloads because of certain functionalities. Author described the revised version of phoenix2 as phoenix++ with the introduction of containers which eventually reduces the memory requirement. It hides task scheduling details and represents a basic map reduce model. Containers are used to store emitted key-value pair by key and storing them in combiners which stored all emitted values with the same key. This increases the necessity of writing high-performance code which eventually improves scalability over phoenix2.

C. Hadoop vs. BitDew Map Reduce

Google invent a new map reduce programming for Internet Desktop Grids using BitDew middleware. The main feature of this implementation highlights a firewall friendly protocol, fault tolerance, result certification, two level schedulers and more. The Author presented new optimizations to BitDew MapReduce in terms of aggressive task backup, intermediate result backup, task re-execution, mitigation and network failure hiding.

A new framework is proposed by the authors [25] which emulated key aspects of Internet Desktop Grid and as well as compared it with apache Hadoop framework. According to their report BitDew Map Reduce framework is able to handle all stress tests whereas Hadoop is not suitable with wide area network topology which includes PC hidden behind firewall and NAT. Additionally BitDew Map Reduce is more successful in terms of fairness, resilience to node failures and network disconnections.

D. Map Reduce Parallel Computation vs. PRAM

The author compared Map Reduce parallel computation model to PRAM (Parallel Random Access Machine) model and analyzed that parallelization of computation on the relatively small number of machines makes Map Reduce model more efficient than PRAM model. However, complete running time for mapper & reducer reaches polynomial time rather than linear. In the paper, authors explained the idea to compute a minimum spanning tree of a dense graph in only two rounds whereas PRAM model requires $\Omega(\log n)$ rounds [26].

V. MAP REDUCE ENHANCEMENT

A. Peacock: An improved version of Phoenix

Phoenix++ is a Map Reduce implementation best worked with shared memory multicore platform. An application distributed sort is efficiently carried out with the introduction of built-in containers. Initially, Map Reduce starts with partitioning the complete data set into equal size portion, each of which is processed by map workers. Further in next step, containers invoked to group the emitted values with the same key and stored them in combiners.

Combiner object passes the data to reduce phase after a run on all cross-thread emitted values. At last reduce phase parse the data and produce the final result stored in result buffer array. With the help of container phoenix++ implementation reduces the overhead occurred in intermediate data storage.

Later author had described a refined version of phoenix++ known as peacock. Peacock is a MapReduce system with workflow customization execution flow which reduced the overhead of intermediate data which is having only one emitted value per key [27].

B. HaLoop and Spark for Iterative Computation

An extended version of MapReduce known as HaLoop used for data-intensive applications also work well for Iterative computation. Author devices Iterative task with three iterations that have two features-

1) Data source of each iteration is having two parts, one is variant and another is invariant.

2) Convergence of iterative procedure to a fixed point might need a progress check at the end of each iteration.

In the iterative computation, additional functions Add Map and Add Reduce of HaLoop works for efficient processing of data. Here different units of HaLoop functions work constantly for variant part of data and stored the Intermediate value of invariant data locally. Hence reduces unnecessary scanning of invariant data.

The reducer just compared the data that has been catched from the previous iteration with the newly generated results to check whether a fixed point is achieved. This strategy helps in time saving with the advent of local storage of invariant data.

Spark is another implementation of Map Reduce, useful for performing iterative computation. A storage abstraction called resilient distributed dataset (RDD), which is a collection of tuples across a set of machines inputted to the map function. A usual processing of map function takes place with the tuples of each partition of RDD and further reduce function is used for aggregation of the resulted tuples. A key feature in spark implementation is the use of intermediate data of RDD stored locally in memory and reused it in subsequent iteration computation. Hence a faster processing of iterative function is carried out [28].

C. Hadoop Online Prototype (HOP)

A new improved version of Hadoop MapReduce framework was proposed by the authors [29] which supports intermediate data to be pipelined between operators and named it Hadoop online prototype (HOP). HOP helped to widen the range of the domain of the problems like a continuous queries execution. According to his study, MapReduce framework can be used for event monitoring and stream processing.

D. Reduced Input size to solve graphs

We know that MapReduce is known for parallel processing of peta byte scale data. An idea of the author is to apply some filtering technique so that the input size can be reduced in distributed manner, resulting to a much smaller problem instance can be solved on a single machine.

Author [30] mainly emphasized on the related graph problems such as for minimum spanning tree, maximal matching, approximate weighted matching, approximate vertex and edge covers and minimum cuts. The given algorithm represents the trade-off between available memories on the machine and numbers of map reduce rounds. Later to proven his idea, the author depicted the implementation of the maximal matching algorithm and represents that how to compute a maximal matching in three map reduce rounds in the model of [31]. Finally author concluded that if the machine have memory O(n) then this algorithm required O(log n) rounds.

E. *HOG: Hadoop on Grid*

The author proposed a Hadoop Map Reduce framework executed on open science grid which covers all institutions span in USA. The framework is different in terms of data availability and detection and resolution of the zombie datanode problem from those which are dedicated to a cluster or cloud. It creates multi institutions failure domain and also provides wide area data analysis as well as map data centers across U.S. This proposed system has experimented with 1100 nodes on grid and provided comparable performance than cluster [32].

F. *Cloud Data Management System*

Map Reduce is a programming model which implements applications over cloud data storage system. Various service providers provided data management systems over cloud such as Google's Bigtable [33], Yahoo's PNUTS/Sherpa, Amazon's Dynamo, Microsoft's Dryad ets.

G. *Summarizing Large Text Based On Map Reduce Framework*

Author proposed a technique to summarize large collection of text using semantic similarity based clustering and topic modeling using Latent Dirichlet Allocation (LDA) over Map Reduce framework [34]. The proposed method is evaluated in terms of scalability, compression ratio, retention ratio, ROUGE and pyramid score. Experiment results have shown the better scalability and reduced time complexity of summarization of large text data over Map Reduce framework. Author also proposed a multilingual text summarization over Map Reduce framework as his future work.

VI. Map Reduce and Data Processing Tools

A. *HadoopDB*

The author suggested a hybrid system of parallel database and Map Reduce based system named HadoopDB to utilize performance and efficiency of parallel database as well as scalability, flexibility and fault tolerance of Hadoop. The ability of HadoopDB makes extensible support for performing data analysis at the large scale of workloads. [35]

B. *Hive*

Hive- an open source data warehousing system used by various companies like Yahoo, facebook etc to store and process huge data sets on commodity hardware [36]. Hive works on a SQL like declarative language- HiveQL to execute queries. Hive contains a system catalog - Metastore – which includes schemas and statistics, useful in data exploration, query optimization and query compilation. Authors are aiming to develop methods for multi-query optimization techniques and generic n-way joins process in a single map-reduce job.

C. *Apache Pig*

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs with a salient property that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets [37]. There is a compiler in Pig's infrastructure layer that produces sequence of map reduce programs. Pig is used a declarative language i.e. Pig Latin which has the properties 1). Ease of programming; 2). Optimization opportunities; 3). Extensibility.

D. *SCOPE (Structured Computations Optimized for Parallel Execution)*

Scope is a declarative and highly extensible language for web scale data analysis on large clusters. This is much like to SQL so users don't require training to use it. Users can easily develop their own functions to solve problems by implementing their own functions and versions of operators: extractors (parsing and constructing rows from a file), processors (row-wise processing), reducers (group-wise processing), and combiners (combining rows from two inputs). SCOPE compiler generates a parallel execution plans which is further optimize by its optimizer [38].

VII. Related Work

The most related work associated with the introduction of various MapReduce models and its relation with the database processing [39]. This tutorial provides the insight about how to improve the performance by increasing availability of the system in case of failure, reduced network communication overhead, process scheduling etc. [40] Authors performed a detailed study about its open source implementation-Hadoop and few factors such as 1) I/O mode, the way of a reader retrieving data from the storage system, 2) data parsing, the scheme of a reader parsing the format of records, and 3) indexing, which is used to speeding up data processing. According to the given study [41] in case of complex analytical task, Hadoop is slower by a factor of 3,1 to 6.5 as compare to parallel data base systems. Later it has been notified that by tuning the above given factors, performance of Hadoop system is improved by a factor of 2.5 to 3.5 for the same benchmark. A critical comparison is carried out between parallel database and MapReduce that criticize the performance of MapReduce [42] for large data bases. According to survey parallel DBMS is more suitable for large scale data processing whereas MR excels in complex analytics and ETL. Basically an interface is required between parallel DBMS and MapReduce to gain the performance excellence of both systems. A large scale data management arise the interest about cloud environment. Author described the concept of cloud computing, related research and its implementation based on VCL (Virtual Computing Laboratory) [43].

VIII. Conclusion

MapReduce provides a distributed parallel computing across multiple nodes and return result on a particular node. MapReduce plays a vital role in parallel data processing because of its salient features such as scalability, flexibility and fault tolerance. Previous Research showed that Map Reduce framework is not sufficient to handle some specific kind of applications. It raised a question regarding improvement and enhancement of the Map Reduce architecture to address those issues and challenges. In this survey paper, our focus was on the extended Map Reduce framework with additional functionalities to support some specific kind of tasks. Initially, we reviewed Google invented Map Reduce architecture and its various applications. Many organizations have invented various Map Reduce frameworks with additional features after Google's invention. We had compared the design and functionalities of frameworks with Apache Hadoop and Phoenix.

A lot of research work has been done on the extension of Map Reduce carried out with new functionalities and mechanism to optimizing it for a new set of problems. We reviewed the extended version of Mapreduce for more data intensive applications such as HaLoop and Spark Map Reduce work well for Iterative computation. Another improved version of Hadoop is known as hadoop online prototype (HOP) designed to support continuous query execution & event handling concluding with the introductory description of HadoopDB which helps to improve the performance of the system with combined features of parallel database and Hadoop database. At last a brief introduction of different data processing tool such as HadoopDB, Hive, Apache Pig and SCOPE used with Map Reduce has been discussed.

IX. Future Research Directions

Map Reduce was initiated by Google to handle big data analysis which is unstructured data such as web document. We have discussed a number of Map Reduce models still researchers can develop a more efficient Map Reduce with improved functionalities. Similarly a new user friendly data processing language can be introduced to make data handling easier.

References

[1] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters" In ACM OSDI, 2004.

[2] A. Elsayed, O. Ismail, and M. E. El-Sharkawi, MapReduce: State-of-the-Art and Research Directions, International Journal of Computer and Electrical Engineering, Vol. 6, No. 1, February 2014.

[3] "Hadoop," available at http://hadoop.apache.org/.

[4] M. Zaharia, M. Chowdhury, Michael J. Franklin, Scott Shenker and Ion Stoica, Spark: Cluster Computing with Working Sets University of California, Berkeley, May, 2010.

[5] D. Moors, Whitehound Limited, UK, "SASReduce An implementation of MapReduce in BASE/SAS", Paper 1507-2014

[6] E. Bugnion, S. Devine, K. Govil, and M. Rosenblum, Disco: Running Commodity Operating Systems on Scalable Multiprocessors (1997).

[7] J. Dean and S. Ghemawat, MapReduce: A flexible data processing tool, Communications of the ACM, Vol. 53 No. 1, Pages 72-77 10.1145/1629175.1629198

[8] K. Ericson and S. Pallickara, On the Performance of High Dimensional Data Clustering and Classification Algorithms (2013).

[9] Y. Zhang, "Optimized Runtime Systems for MapReduce applications in Multi-core clusters", A thesis in Houston Texas, May 2014.

[10] S. Blanas, J. M. Patel, V. Ercegovac, J. Rao, E. J. Shekita, and Y. Tian, "A Comparison of Join Algorithms for Log Processing in MapReduce," in Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, SIGMOD '10, (New York, NY, USA), pp. 975–986, ACM, 2010.

[11] E. A Mohammed, B. H Far and C. Naugler, Application of the MapReduce programming framework to clinical big data analysis:Current landscape and future trends, BioData Mining. Vol. 7, 22. Oct 29, 2014.

[12] M Jonas, S Solangasenathirajan and D Hett. Annual Update in Intensive Care and Emergency Medicine 2014. New York – USA: Springer. Patient Identification, A Review of the Use of Biometrics in the ICU; pp. 679–688. (2014).

[13] E. Arslan, M. Shekhar & T. Kosar, "Locality and Network-aware reduce task scheduling for data intensive applications", published in Proceedings DataCloud'14 Proceedings of the 5th International workshop on Data Intensive Computing in the Clouds, page 17-24, ISBN:978-1-4799-7034-6

[14] D. Gillick, A. Faria and J. DeNero, "Map Reduce: Distributed Computing and Machine Learning", Dec-2006

[15] Owen O'Malley, "TeraByte Sort on Apache Hadoop", Yahoo! owen@yahoo-inc.com May 2008.

[16] S. Sakr, Processing Large Scale Graph Data: A Guide to Current Technology, National ICT Australia, June 2013.

[17] U Kang et al., GBASE: A Scalable and General Graph Management System, San Diego, California, U.S.A. ACM978-1-4503-0813-7/11/08. Aug-2011.

[18] J. Dean, "Experience with MapReduce, An Abstraction for Large Scale Computation", Google, Inc., proceedings of the 15th international conference on parallel architecture and compilation techniques, ACM New york, US, ISBN:1-59593-264-X doi>10.1145/1152154.1152155

[19] S. Chen and S. W. Schlosser,"Map reduce meets wider varieties of applications", IPR-TR-08-05, Research at Intel (2008).

[20] W. Zhao, H. Ma & Q. He, "Parallel K-means clustering based on mapreduce", cloudcom, LNCS 5931, pp 674-679 © springer-verlag Berlin Heidelberg 2009.

[21] J. Talbot, R. M. Yoo, and C. Kozyrakis, "Phoenix++: Modular mapreduce for shared-memory systems," In Proc. of the second international workshop on MapReduce and its applications, pp. 9–16, 2011.

[22] C. Cao, F. Song, D. G. Waddington, "Implementing a high performance recommendation system using Phoenix++", In Proc. of Internet Technology and Secured Transactions, 8th International Conference for, DOI 10.1109/ICITST.2013.6750200, pages 252-257, dec 2013.

[23] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis. Evaluating MapReduce for multi-core and multiprocessor systems. In Proc. of the 13th Int'l Symposium on High Performance Computer Architecture, pages 13–24, 2007

[24] R. M. Yoo, A. Romano, and C. Kozyrakis. Phoenix rebirth: Scalable MapReduce on a large-scale shared-memory system. In Proc. of the 2009 IEEE Int'l Symposium on Workload Characterization, pages 198–207, 2009.

[25] L. Lu, H. Jim, X. Shi, Fedak G.,"Assessing MapReduce for Internet Computing: A Comparison of Hadoop and BitDew-MapReduce", In Proc. of the Grid Computing (GRID), ACM/IEEE 13th International Conference on Grid Computing, DOI:10.1109/Grid.2012.31, ISBN: 978-0-7695-4815-9, pp. 76-84, Sept 2012.

[26] H. Karloff, S. Suri and S. Vassilvitskii, A Model of Computation for MapReduce, at AT & T Labs and Yahoo! Research (2010).

[27] S. Wu, Y. Peng, H. Jin, J. Zhang,"Peacock: a customizable Map Reduce for Multicore plateform, In Proc. of the Journal of Supercomputing, DOI 10.1007/s11227-014-1238-2, Springer Science + Business Media New York pages:1496-1513, June 2014

[28] Li, F., Ooi, B-C., Özsu, M. T., Wu, S., Distributed Data Management Using MapReduce. ACM Comput. Surv. 0, 0, Article A (0), 41 pages. DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000 (2013).

[29] T. Condie, N. Conway, P. Alvaro, J. M. Hellerstein,"MapReduce online" in UC Berkeley, 2010.

[30] G. D. F. Morales, A. Gionis and M. Sozio. Social Content Matching in MapReduce. 37th International conference on very large databases, Seattle Washington, Proceedings of the VLDB Endowment, Vol. 4, No. 7 Copyright 2011 VLDB Endowment 2150-8097/11/04.

[31] A. V. Goldberg and S. Rao. "Beyond the flow decomposition barrier", JACM, 45(5):783–797, 1998.

[32] C. He, D. Weitzel, D. Swanson, Y. Lu. HOG: Distributed Hadoop MapReduce on the Grid Published by SC Companion: High Performance Computing, Networking Storage and Analysis (2012).

[33] Sakr S, Liu A, Batista DM, Alomari M, A survey of large scale data management approaches in cloud environments. IEEE Commun Survey Tutorials 13(3):311-336 , 2011.

[34] N K Nagwani, Summarizing large text collection using topic modeling and clustering based on MapReduce framework, Journal of Big Data 2:6 DOI 10.1186/s40537-015-0020-5, 2015.

[35] A. Abouzeid, K. B. Pawlikowski, D. Abadi, A. Silberschatz and A. Rasin, HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads, VLDB Endowment '09' August Lyon France, 24-28, 2009

[36] A. Thusoo, J.S. Sarma, N. Jain, Z. Shao, P. Chakka, N. Zhang, S. Antony, H. Liu, and R. Murthy, Hive – A Petabyte Scale Data Warehouse Using Hadoop. Proc. of ICDE, 2010.

[37] Hadoop Pig. Available at http://hadoop.apache.org/pig

[38] R. Chaiken, et. al. Scope: Easy and Efficient Parallel Processing of Massive Data Sets. In Proc. of VLDB, 2008.

[39] J. Zhao, J. Pjesivac-Grbovic, MapReduce: The Programming Model And Practice, 2009.

[40] D. Jiang, B. C. Ooi, L. Shi and S. Wu, The performance of mapreduce: An in-depth study. Proc. VLDB Endow., 3 pp. 472–483 (Sept 2010),

[41] A. Pavlo, E. Paulson, A. Rasin, D.J. Abadi, D.J. DeWitt, S. Madden, and M. Stonebraker. A comparison of approaches to large-scale data analysis. In Proceedings of the 35th SIGMOD international conference on Management of data, SIGMOD '09, pages 165–178. ACM, 2009.

[42] M. Stonebraker, D. J. Abadi, D. J. DeWitt, S. Madden, E. Paulson, A. Pavlo, and A. Rasin. MapReduce and parallel DBMSs: friends or foes? Communictions of the ACM, 53(1):64{71, 2010.

[43] M. A. Vouk, Cloud Computing-Issues, Research and Implementations, Journal of Computing and Information Technology, CIT 16, 4, 235-246, doi-10.2498/cit.1001391, (2008).

AUTHOR PROFILES

Shafali Agarwal is associated as an assistant professor with JSSATE, Noida, formerly she worked with NIET, Greater Noida. She has received her Ph.D. from Singhania University, Rajasthan. Her research areas are MapReduce Implementation and fractal analysis which is a part of Image processing. She has published papers in national conference, International conference and International journal which are indexed by ACM, Springer, Citeseer, ProQuest, Index Copernicus, EBSCO, Scribd and many more. She has done graduation in 2001, master in computer applications in 2004 and after that MPhil in 2007. She got published a book titled "Data Structure using C" for engineering students. She is an active member of IEEE Computer society.

Zeba Khanam is presently working at the department of Computer Science, JSSATE,Noida .She has received her PhD degree from Jamia Millia Islamia (Central University),New Delhi. Her research focuses on evolving the legacy systems using refactoring techniques with aspect oriented programming.

Her research interests are Map Reduce Framework, Software Re engineering and Reverse Engineering. Her recent publications include articles in IEEE Xplore, Wseas Transactions, Procedia Engineering,WORLDComp'11 proceeding

Security Issues Model on Cloud Computing: A Case of Malaysia

Komeil Raisian
Faculty of Information Science and Technology, the National University of Malaysia 43600 UKM Bangi
Selangor, Malaysia

Jamaiah Yahaya
Faculty of Information Science and Technology, the National University of Malaysia 43600 UKM Bangi
Selangor, Malaysia

Abstract—**By developing the cloud computing, viewpoint of many people regarding the infrastructure architectures, software distribution and improvement model changed significantly. Cloud computing associates with the pioneering deployment architecture, which could be done through grid calculating, effectiveness calculating and autonomic calculating. The fast transition towards that, has increased the worries regarding a critical issue for the effective transition of cloud computing. From the security viewpoint, several issues and problems were discussed regarding the cloud transfer. The goal of this study is to represent a general security viewpoint of cloud computing through signifying the security problems that must be focused and accomplished appropriately for identifying a better perspective about current world of cloud computing. This research also is to clarify the particular interrelationships existing between cloud computing security and other associated variables such as data security, virtual machine security, application security and privacy. In addition, a model of cloud computing security which depends on the investigation regarding previous studies has been developed. To examine the model a type of descriptive survey is applied. The survey sample population is selected from employers and managers IT companies in Malaysia. By testing the correlation, the results of study indicated that there are those identified security challenges in current world of cloud computing. Furthermore, the results showed that the cloud computing security correlation with data security, virtual machine security, application security and privacy is positive.**

Keywords—Cloud Computing; Security Issues; Security Viewpoint; Grid Computing

I. INTRODUCTION

Cloud computing refers to the next-generation architecture of IT enterprise. This approach usually focuses on financial utility model by developing several existing methods and computing technology that include distributed application, facilities, and IT infrastructures of computers, networks, and storing capitals [1]. Currently, associated administrations, particularly in small and medium business (SMB) enterprise, assume cloud computing to improve the efficacy and success of their organization as well as decreasing the cost of buying and preserving the organization [2]. Because of the high interest about toward the cloud computing, there is an important worry about the evaluation of the current tendencies in security of that technology. Internet considered as a driving force of the improved technologies and possibly, one of the most important topics in this regard is Cloud Computing. Cloud computing considered as a tendency of the current day scenario, with nearly all the administrations that tried to make it as an entry. Followings are considered as the benefits of using cloud computing: i) decreasing the price of hardware and related maintenance charges, ii) it is easily accessible all over the world, and iii) it has flexible and the extremely automatic procedure, in that the clients are not usually worried about upgrading the software which is a daily manner [3]. Security control events in cloud are identical to old IT settings. Regarding the multi-tenant features, service delivery and deploy models of cloud computing usually associates with the previous IT settings, though, cloud computing might have several risks and problems too [2].The most important goal of this study is to classify, categorize and establish the most important security problems associated with cloud computing which that were discussed in the related literature.

II. CLOUD COMPUTING TYPES AND SERVICES MODEL

To provide a better secure cloud computing clarification, a significant decision relates to the kind of cloud to be applied. Currently three usual utilization approaches of cloud were presented [4]. Initially, private cloud which refers to the private cloud infrastructure set up inside an internal enterprise information center. Regarding this type, it should be noted that, this type is easier to be aligned with security, agreement, and controlling and it can offer a better control over arrangement and practice [5]. Secondly, Public cloud which refers to the cloud infrastructure is usually applied for general public or a big manufacturing groups. Public cloud considered less secure comparing to other related cloud models [6]. And third one refers to the Hybrid cloud which is an alignment of private cloud associated to one or more public cloud services that were restricted through a secure network. Hybrid delivers more secure controlling of the over data [7].

III. CLOUD COMPUTING TYPES AND SERVICE MODEL OF DELIVERY

The second security issue that follows the cloud organization models and should be unpacked by the enterprise management refers to the cloud examination delivery models [8]. The architecture regarding the service delivery models of cloud computing might be characterized in three categories. Firstly, Infrastructure as a Service (IaaS) which considered as the essential computing capitals e.g. storing, network, servers were applied for providing facilities to the final clients. Secondly, Software as a Service (SaaS) which deliver a remote access using a web for operating virtualized and pay-per-use software over the cloud structure .Thirdly, Platform as a service (PaaS) that refers to the use of equipment and capitals which were delivered by the cloud structure for supporting the end client needs.

IV. Cloud Computing Security Issues

The advantageous of cloud computing that were presented in the related literature covers both fast and easy arrangement, the pay-per-use ideal, and reducing the in-house IT charge. Nevertheless, they also believed that the security considered as the most important subject which should be discussed in the related literature to improve the security of cloud computing application [9]. Main security subjects about cloud computing are as follows:

A. Data Security

Liu [10] believed that cloud computing refers to the quick developments of the users for essay access to the required hardware, software and facilities and other associated capitals in several times. Furthermore, the data security seems to be more important while it is applied for the cloud computing in the SPI background. Cloud computing also faced with several problems, if the experts will not resolve them well, its fast improvement could be affected too. Data security is common in several applications and among the associated problems, it can create many problems for the operators while they keep sensitive data in different cloud servers. These problems refers to the cloud servers which are typically functioned by commercial providers that are very probable to be used outside of the reliable territory of the operators [11].

1) Availability

Certifying timely and consistent contact for using the related data. An obtainability problem refers to the interruption of contact for using the related data or for the data system [12]. Bowers et al. [13] presented HAIL (High-Availability and Integrity Layer), a distributed cryptographic scheme which permits servers to prove to the client that a stored file is complete and retrievable. HAIL strengthens, officially unifies, and streamlines are considered as distinct methods about the cryptographic and distributed-systems groups. Proofs in HAIL are well quantifiable by servers and extremely compact typically tens or hundreds of bytes, regardless of considering the size of the file. HAIL cryptographically confirms and reactively changes file shares. It is strong against a dynamic, mobile opposition, i.e., one which might increasingly corrupt the full different servers. Bowers et al. [13] suggest a perfect, official adversarial approach for HAIL, and difficult examination of element choices, in which the researcher showed how HAIL progresses the safety and efficacy of current equipment, such as Proofs of Retrievability (PORs) that is organized on separate servers.

2) Confidentiality

Preserving authorized limitations about the data accessibility and release, like the tools for keeping the individual privacy and exclusive data, a loss of privacy is the illegal disclosure of data [12]. About some useful application arrangements, the privacy of the information is not considered merely a security/privacy subject, but a juristic problem. As an example, in healthcare application projects using the Protected Health Information (PHI) must follow the necessities of Health Insurance Portability and Accountability Act (HIPAA), and making the user information private in the storage servers is not just considered as a possibility. Data confidentiality might likewise attained while Cloud Servers cannot learn the plaintext of related data file in the system [11].

3) Integrity

Guarding against unsuitable data adjustment or destruction and safeguarding the data, non-repudiation and truthfulness, a loss of truthfulness is the unauthorized change or destruction of data [12]. One big problem about the cloud data storage refers to the data integrity confirmation at untrusted servers. As an example, the storage service supplier, that experiences Byzantine problem infrequently, might decide to hide the information errors of the users for their own benefits. The more serious issue is that by saving money and the storage space, the service provider may ignore keeping or deliberately delete the infrequently retrieved information files that belongs to a normal user. By considering the large dimension of the outsourced electronic information and the customer's forced resource competence, the core of the problem might be widespread as how can the customer find an effective method for performing periodical integrity confirmations without copying the local information files [14].

4) Data location

Several regulations for managing the information might vary from country to country. Consequently, transporting confidential information among the countries could be considered as a challenging task. Regarding the cloud setting, the position of the information centers and backups should be understood perfectly to ensure that legal problems wouldn't happen [15]. By using the Cloud Computing, users will have the chance of using data mobility capabilities to a high extent and customers do not typically know the location of their information and in many cases, it is not considered as big challenge for the users. As an example, emails and photographs that were uploaded to the Facebook might exist all over the world and Facebook users are commonly not concerned about this matter. Nevertheless, once an enterprise has some sensitive information which is kept on a storage device of the Cloud, they might want to see its location too. Moreover, they might also want to identify a favored location (e.g. information to be reserved in the UK), then they needs a contractual contract among the Cloud service providers and customers, in that information must stay in a specific position or exist in on a specified recognized server [16].

5) Data Recovery

Data Recovery considered as an important section of each Business Continuity Planning. By applying an unrestricted cloud provider, it should be noted that the Business Continuity Planning and the Data recovery could be expanded to contain catastrophes which affects the public cloud provider. About natural problems or related disasters, a cloud service provider information center might be inaccessible. About this possibility, it is important to apply a well-thought out disaster retrieval strategy [17]. An event like a server breakdown might cause injury or loss about the users' information. To avoid this issue, users should do a backup from the data for the recovering in the future. Additionally, cloud users can save backup of important data on a local computer [1].

6) Retention

How long could the personal data which was transported to the cloud retained? Who applies the preservation strategy about the cloud, and how we can manage litigation holds [18]. How long we can retain the personal information which were transferred by the cloud? Who imposes related retention policies in the cloud environment, and how we can manage our exceptions regarding this policy (like the litigation holds). Logs typically contain timestamps and timing information considered important for the compliance of laws and policies about the data retention, so it seems important to have a data retention and destruction strategy for all related data storing schemes. Timing activates might also decrease the data which should be recorded as the temporary information which is merely kept for doing current transaction and formerly deleted that has minimal confidentiality implications [19].

7) Ownership

Typically, workers or administrations have accessibility to the information and they can manage them well. Once the information moved to the cloud, we have to consider how we can maintain the Information possession [20]. Data ownership refers to the clouded initial move of the cloud, with queries about what happens to information while it moves to the cloud? What occurs while a cloud provider goes out of industry? In addition, what occurs if cloud clients could not pay their bills? [20]. Cloud computing did investigations about the virtualization, distributed computing, utility computing, and during the recent years on networking, web and software facilities. It suggests a service-oriented architecture, limited data technology overhead for the end-client, having high flexibility, decreased total cost of possession, on request facilities and many related matters.

8) Access control

Regarding the cloud setting, the association between capitals and operators considered very commercial hoc and active resource workers and customers are not usually located in a similar security field while clients are typically recognized by their features or qualities, not predefined characteristics. Consequently, the old-style character based access control models are not effectual, and access decisions should be made according to the qualities. Diplomas delivered by a PKI facility might be applied to enforce admission control in the Web setting [21]. Access Control permits one application to trust the individuality of related application. The old model for accessing control is application-centric access control, where per application keeps related tracks of its user collection and manages them which is not practical in cloud founded architectures, as in this approach we need lots of memories to store the details of the users like their username and password. Consequently, cloud needs a user centric access control while every operator request to the related service provider that is bundled with the user character and right data [22].

9) Data lock-in

In the other word, the clients cannot move easily from a SaaS or IaaS vendor to the other one. The client data could be destroyed, that stop users to adopt cloud Computing. Coghead recognized as an example of a cloud platform whose shutdown left clients scrambling, to reword their requests for running on the other platform and the solution is to regulate cloud Application Programming Interface (API) [1]. Weiss [23] believed that software considered as a service in the cloud that might recover doubts about the vendor information lock-in as an important concern in the processer era. Assuming a cloud worker and a thin-client seller partner together, it is likely that per half will need the other. Services about the cloud might be unreachable to those without an access tool from a single brand. Some believe that the cloud could inspire the development of walled-gardens, a potential step back associated to the comparatively open internet of today.

B. Virtual machine level security

Virtual setting contains different VMs, which deliver self-governing security areas. It is hard to manage several VMs effectively, working on a similar physical organization. The most important purpose of Virtualization refers to the way it certifies several VM examples working on a similar physical engine that are separated from each other [24, 25]. Virtual machines (VMs) considered as the most common form to provide the computational capitals of cloud operators at this layer, where the operators get finer-granularity flexibility as they generally get super-user access to their VMs, and may use it to modify the software stack on their VM for presentation and efficacy and frequently, such facilities are dubbed Infrastructure as a Service (IaaS). Virtualization considered as an enabler technology for this cloud component that permits users of unparalleled flexibility to arrange their locations while protecting the physical organization of the providers' information center. Recent progresses in OS Virtualization made the IaaS concept believable. This was exactly enabled by two virtualization methods namely; par virtualization and hardware-assisted virtualization. Though both virtualization skills concerned with the performance separation among virtual machineries opposing on shared resources, performance interference among VMs shares, and similar cache and TLB hierarchy cannot yet be evaded [26].

1) Hypervisor security

Hypervisor considered as a key software constituent of Virtualization. It usually affects all VMs acts working with the Virtualization host. While an attacker totally controls a hypervisor, then he may apply any activity to the VMs on the host scheme. Two stages in security administration of hypervisor were proposed [1, 24 and 18]. Regarding a real hypervisor product, like Xen, OpenVZ or VMware, the attacker usually attempt to exploit the security holes to modify the hypervisor, so that he may install a rootkit on it. The solution is to update and patch the hypervisor product and other virtualization products regularly. Furthermore, the investigation of how different components in the hypervisor architecture work, like monitoring the actions of the guest VMs and intercommunication amongst different infrastructure machineries, might contribute improving the security of cloud system. Two stages in security administration of hypervisor were proposed that will follow [1, 24 and 18].

2) Authorization and Authentication

Authorization and Authentication are considered as the most significant features of managing a virtual host reviewing goal. Authorization confirms that clients should be authorized and have consent to do their required tasks. For the authentication, suitable values and existing instruments should be applied to validate related account correctly [27]. Youseff [26] believed that before persuading customers for migrating from desktop to cloud applications, cloud applications' providers should consider different users' concerns regarding both security and safety of keeping private information on the cloud, users' verification and approval, up-time and presentation, backing up the data and problems for recovering and providing reliable SLAs about their cloud applications.

3) Networking

Network communications and arrangements are considered as the important security subjects about the cloud computing organizations. Cloud computing embraces cyber infrastructure, and shapes upon periods of investigation in virtualization, dispersed computing, “grid computing”, utility computing, and more lately, networking, web and software amenities. It usually refers to the service oriented architecture, reduced information technology designed for the users, being more flexible, reduced entire charge of ownership, on request facilities and several other issues [20]. It is significant to deliver a mechanism for the assurance of secure assembly of the organization in the safety zone that has three instruments [24]. Firstly, transfer security that it is Cloud computing circulated architectures contain a huge resource sharing and virtual engine instance synchronization. Therefore, it needs VPN machineries to protect the cloud scheme against sniffing, spoofing and side-channel problems. Secondly, firewalling which refers to the protection of the provider's interior cloud substructure, Firewalls can deliver protection from insider and outsider and permit VM isolation, fine-grained filtering about the addresses and ports and preventing the Denial-of-service (DoS). It is significant to improve a reliable firewall and other safeties regarding the cloud contexts. Thirdly, Security configuration that usually focuses on the formation of protocols, schemes and skills for meeting predictable level of security and confidentiality without cooperating performance of efficacy.

4) Isolation

In the virtual setting of cloud computing, although it reasonably isolated, all VMs have the same hardware and therefore the similar capitals. This might clue to exploit of data leaks and cross-VM attack. For better protection the notion of separation might also be used for a better fine-grained properties, like computational capitals, storing and memory [1]. The most important feature of virtualization refers to the ensuring of VM instances running on a similar physical mechanism that are separated from each other. Though, in the isolation technologies current VMMs offer and the control of manager on host and guest working schemes are not considered good that leads to several security matters of virtualization [1]. Virtual machine technology delivers strong isolation among virtual areas. As an example, security isolation avoids a malicious application to attack applications or retrieving information in other areas. Fault isolation avoids one fault application to bring down the entire system. Environment isolation permits several operating schemes for running on a similar machine, accommodating legacy applications and cutting-edge software, each with a distinct set of arrangements and elements [28].

C. application security

Application security refers to the use of system capitals, such as software and hardware for secure requests which holds them in contradiction of malicious saturation that attacks the cloud. Though there is a security program in cloud computing like Quad Core Intel Xeon Processors and IP address [30], but still there are several problems in security at application step which might permit unlawful clients to have access. Consequently, cloud is usually insecure the application due to the security holes like unconfident software Connectors or APIs interrelating with cloud facilities. There are different threats regarding the security program of the cloud [29].

1) Cloud browser security

In Saas module Customers computing tasks are allocated to the remote server. The client system is usually applied for the IO take and sends instructions to the cloud. Though there are several security matters in cloud, but browser security is very significant, particularly in cloud computing [31]. Browser might only be applied in the encryption and signature Transport Layer Security (TLS) that usually have enough security to define the malicious attacks. Solution provided Simultaneous application of TLS and XML is founded encryption at the central of the browser [32]. Web browsers might not openly apply XML Signature or XML Encryption and information might merely be encoded over TLS, and signatures are merely applied inside the TLS handshake. For all other cryptographic information circles inside WS-Security, the browser merely serves as the passive information store. Several simple workarounds were planned to be used e.g. TLS encryption instead of XML Encryption, but the main security challenges with this method were elaborated in the literature and working attacks were applied as proofs-of concept (cf. 3.2.2). Our purpose is to suggest provably secure solutions applying TLS, but at the same time inspire the browser community to adopt XML founded cryptography to be included in the browser core [32].

2) Cloud malware attack

This kind of attack injects VM malicious or implementation service to cloud computing organization and its goal it is to vary extensively, either stopping, eavesdropping or adapting information by adapting the delicate about overall Capability variations. The aggressive make VM destructive instance of model Implementation Services like, Saas, Iaas and add it to cloud computing. And its answer refers to the implementation of integrated review, like Services before using it for received desires along the cloud scheme [33]. Bhadauria et al. [29] believed that apart from the above stated network, related problems are regarded to dissimilar security problems in a mobile cloud computing setting. By using the applications lying over the cloud, it is likely for the hackers to corrupt an application and gain access to the mobile device once opening that application. To avoid those conditions, strong virus scanning and malware security software should be installed to prevent any kind of virus/malware check into the mobile scheme. Likewise, by inserting device identity guard, like permitting access to the authorized operator according to some form to identity check feature that will let blocking unauthorized admission.

3) Backdoor and debug option

The majority of the designers write code which are backdoor requested or unwanted. They might likewise stop some debugging choices for testing or revising the website again. In Saas and Paas models, though backdoors are in these contexts, but some hackers can simply enter website and use the important related data. These concepts must be resolved at the advanced level [34]. A usual habit of the designers is to permit the debug choice when publishing a web-site. This allows them to make developing variations in the code and getting them executed in the web-site. Since these debug options are considered ease backend admission to the inventers, and occasionally these debug choices are left allowed unnoticed, it might deliver an easy possibility for the hacker to enter the web-site and allows him/her for making variations at the web-site level [29].

4) Cookie poisoning

It points to the illegal contact or web requests by Identify bases of cookie. In Saas model, Cookies defense of data permits requests for detecting illegal user identification and these cookies are obtainable. They might be invented for shaping the individuality of an illegal user [29]. The threats to application level security contain XSS attacks, Cookie Poisoning, Hidden field manipulation, SQL injection attacks, DoS attacks, Backdoor and Debug Options, CAPTCHA Breaking etc subsequent from the illegal usage of the applications. It includes altering or adapting the contents of cookie for making unauthorized accessibility to an application or to a webpage. Cookies essentially include the identity of the user related credentials and when these cookies are available, the content of those cookies could be forged to imitate an authorized operator. This might be evaded either by doing regular cookie cleanup or applying an encryption system about the cookie information [29].

5) Privacy

The data of operators are usually sored in the data center, then the cloud provider allocate them among hundreds of servers that wish to have risk possibility. These facilities are applying internet as announcement, presenting online software, so cloud providers particularly [Iaas] would be involved with risks [35]. Once the users attempt to use their hidden data from cloud provider examination in this period, they would lose the information. This is considered, as good chance for attackers, they would examine to submit data by operators [34]. Some data refers to the name, address, religion, race, well-being job performance, credit card number which depends on the type of cloud provider facilities [36]. Cloud providers particularly [Iaas] offer to their users that the data storage frequently bring a frictionless of the procedure of registration, as it lets someone to use cloud service and there are several indications that hackers instigated to target [Iaas] retailers [37]. Based on the cloud-based amenities customer's information kept in third-party part [38]. Consequently, service provider should measure the amount of information security precisely for ensuring the privacy of the information. One way for enhancing the safety is incorporation of information encryption. In the other words, incorporation information encryption with data could be done to protect the information of the user against hackers, and it will be helpful to limit the accountability of service providers. Wen and Xiang [39] believed that the protections against malicious hackers who might have access to the service provider's scheme considered as the final goal which is not sufficient. We might face several dangers once providers attempt to recover the related information. Consequently, how providers may improve the customer's information? It seems to be easy that user only find the cloud provider that he/she can trust. This method considered appropriate once the data is not so significant. This approach is suitable to recover the data in small company for finding the reliable provider. Surly, it might be a problem in that company, but for medium-sized to maximum-sized firm, it is logical to find a solution than finding reliable provider for information retrieval. They must expand techniques and approaches over the information encryption to ensure the privacy of cloud provider or apply private cloud could be a better clarification in these businesses.

V. Proposed Model

The quantitative study method had been useful for this research. Figure 1 shows the proposed model of this research to know the effects of the factors (data security, virtual machine security, application security and privacy) on cloud computing security in current world of cloud computing based on enlists main studies indicating main factors in Table 1. This model presents the relationships between Main Security factors on Cloud Computing and indicates how those components are positively associated with it. The final results involving SPSS tested the correlation of those factors with cloud computing security based on sample population is selected equal with 150 that is randomly obtained from employers and managers IT companies in Malaysia. Figure 1 exhibits the particular associations concerning these factors and Cloud Computing security as well. Questionnaires were arranged to be determined by these factors. This section supplied the reason on the four factors independently and together with their investigation from the questions gotten by respondents of the research. This reason was taken by the researcher to come up with a model as it is presented in Figure 1. The model combines those factors that had not really been connected from the previous researcher. The researcher examined impacts of each component in suggested model through conducting an additional study in the various firms from the previous one.

TABLE I. CLUSTERING SORTS OF CLOUD COMPUTING SECURITY

Articles	**Cloud computing security**			
	Data security	**Virtual Machine Security**	**Application Security**	**Privacy**
[10]	*			
[13]	*			
[12]	*			
[11]	*	*		
[14]	*			
[16]	*			
[17]	*			
[18]	*	*		
[19]	*			
[20]	*	*		
[21]	*			
[22]	*			
[23]	*			
[24]		*		
[25]		*		
[30]			*	
[27]		*		
[26]		*		
[28]		*		
[29]			*	
[32]			*	
[33]			*	
[34]			*	
[36]				*
[35]				*
[37]				*
[38]				*
[39]				*
[15]	*	*		
[30]			*	
[31]			*	
Total	14	9	8	5

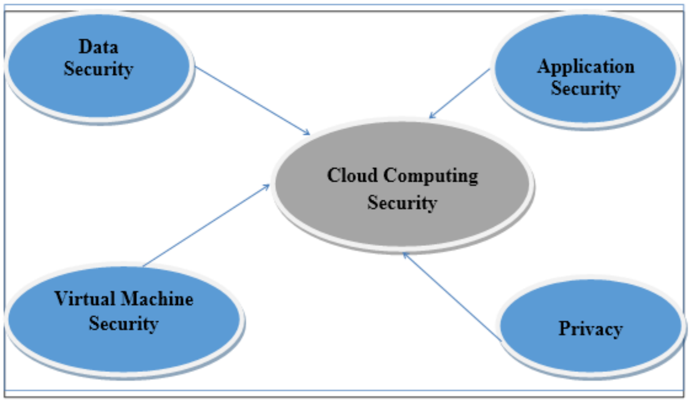


Fig. 1. Research Proposed Model

VI. THE RELATIONSHIP OF SECURITY CLOUD COMPUTING FACTORS

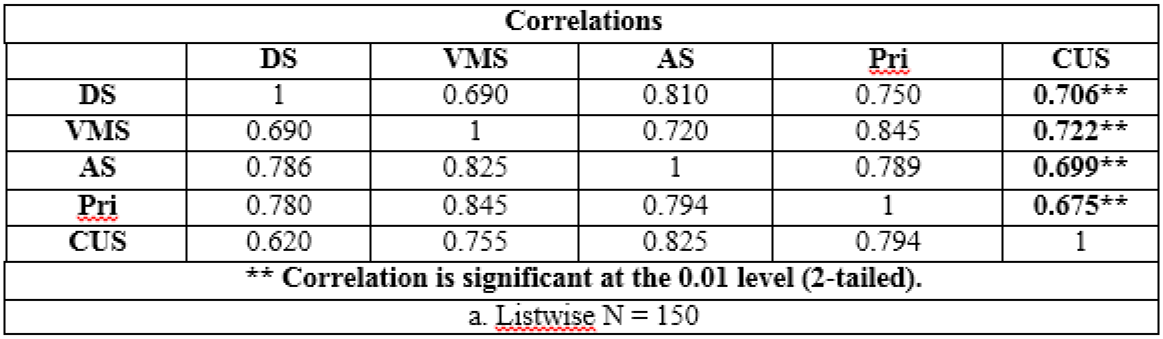
This study had recently mentioned data security, virtual machine security, application security and privacy as main factors of proposed cloud computing security model. Therefore, this section aims to examine the correlation of these seven factors with cloud computing security. Common produced variances are all above the advised 0.5 degrees, Anderson and. [40] stated that supporting the discriminate validity of measurement scales supporting the discriminate validity of measurement scales. Correlation indicates the strong of a linear relationship between two variables. The created correlation coefficients that symbolize the strength of these relationships relating to the research variables are shown in Table 2.

It is obvious from Table 2 that the correlation coefficients on the connections between research variables can be found strong. In addition, within Table 2, the correlation coefficient value is 0.706 between Data Security (DS) and Cloud Computing Security (CUS) variables that indicates these variables are strongly correlated since it is greater than 0.5. Virtual Machine Security (VMS) is found to be strongly related to CUS and the correlation coefficient value is 0.722. Application Security (AS) is found to be strongly related to CUS with correlation coefficient equal with 0.699. Furthermore, Privacy (Pri) is strongly related to CUS with the correlation coefficient value of 0.675. The results still support our proposed model. While our results show that whole, factors on current world of cloud computing are indeed distinct constructs and it also appears that all factors are well correlated with Cloud Computing Security.

VII. CONCLUSION

Cloud computing ensures having an extensive effect on the schemes and networks of organization and other initiatives and it focuses on cost decrease, high performance and benefit of cloud computing in the administrations .One of the gorgeous trait in cloud computing might refer to the difference between classic security plan and control. Classifying the safety of complicated computer system that joint together is a long time security subjects regarding the computing in overall cloud computing. Access to high qualities considered as the main purpose in applying the cloud computing security experts and workers. Public cloud computing considered as a critical factor that enterprises needs for combining their Information as a solution package. The enterprises must be ensured that related activities about security and confidentiality is happening correctly in their business. Assessing management risk in cloud computing systems could be changed in several organizations.

TABLE II. INTER-ITEM CORRELATION



Correlations					
	DS	VMS	AS	Pri	CUS
DS	1	0.690	0.810	0.750	**0.706****
VMS	0.690	1	0.720	0.845	**0.722****
AS	0.786	0.825	1	0.789	**0.699****
Pri	0.780	0.845	0.794	1	**0.675****
CUS	0.620	0.755	0.825	0.794	1

**** Correlation is significant at the 0.01 level (2-tailed).**

a. Listwise N = 150

Likewise the system must have a balance against obtainability of privacy and security's control. Administrations must evaluate the fit balance among the number and strength of the control and hazards associated with the cloud computing solutions. This research also clarifies this purpose and its relationship with cloud computing security within four crucial constructs in cloud computing. By testing the correlation, the results of study indicated that the results of study indicated that there are those identified security challenges in current world of cloud computing. In addition, the results showed that the cloud computing security correlation with data security, virtual machine security, application security and privacy is positive. Between cloud computing security and other associated variables such as data security, virtual machine security, application security and privacy.

ACKNOWLEDGMENT

First and foremost praise be to Almighty Allah for all his blessings for giving me patience and good health throughout the duration of this research. I wish to thank my supervisor and my family for their love, support, encouragement and sacrifice throughout the course of this study. Last, but not least, I would like to thank my beloved wife for standing by me no matter how rough the sea and the journey has been. She is the only one that can attest to all sacrifices we both had to make, so that this manuscript can become a reality.

REFERENCES

[1] You, P., Peng, Y., Liu, W., & Xue, S. (2012, June). Security issues and solutions in cloud computing. In Distributed Computing Systems Workshops (ICDCSW), 2012 32nd International Conference on (pp. 573-577). IEEE.

[2] Chen, D., & Zhao, H. (2012, March). Data security and privacy protection issues in cloud computing. In Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on (Vol. 1, pp. 647-651). IEEE.

[3] Maggiani, R. (2009, July). Cloud computing is changing how we communicate. In Professional communication conference, 2009. IPCC 2009. IEEE international (pp. 1-4). IEEE.

[4] Ramgovind, S., Eloff, M. M., & Smith, E. (2010, August). The management of security in cloud computing. In Information Security for South Africa (ISSA), 2010 (pp. 1-7). IEEE.

[5] Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: state-of-the-art and research challenges. Journal of internet services and applications, 1(1), 7-18.

[6] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., & Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50-58.

[7] Krutz, R. L., & Vines, R. D. (2010). Cloud security: A comprehensive guide to secure cloud computing. John Wiley & Sons.

[8] Mell, P., & Grance, T. (2011). The NIST definition of cloud computing.

[9] Zissis, D., & Lekkas, D. (2012). Addressing cloud computing security issues. Future Generation computer systems, 28(3), 583-592.

[10] Liu, X. (2014, January). Data Security in Cloud Computing. In Proceedings of the 2012 International Conference on Cybernetics and Informatics (pp. 801-806). Springer New York.

[11] Yu, S., Wang, C., Ren, K., & Lou, W. (2010, March). Achieving secure, scalable, and fine-grained data access control in cloud computing. In INFOCOM, 2010 Proceedings IEEE (pp. 1-9). Ieee.

[12] Winkler, V. (2011a). Chapter 1 - Introduction to Cloud Computing and Security Securing the Cloud (pp. 1-27). Boston: Syngress.

[13] Bowers, K. D., Juels, A., & Oprea, A. (2009, November). HAIL: a high-availability and integrity layer for cloud storage. In Proceedings of the 16th ACM conference on Computer and communications security (pp. 187-198). ACM.

[14] Wang, C., Wang, Q., Ren, K., & Lou, W. (2010, March). Privacy-preserving public auditing for data storage security in cloud computing. In INFOCOM, 2010 Proceedings IEEE (pp. 1-9). Ieee.

[15] Kumar, P., & Arri, H. S. (2013). Data Location in Cloud Computing. International Journal for Science and Emerging Technologies with Latest Trends, 5(1), 24-27.

[16] Mahmood, Z. (2011, September). Data location and security issues in cloud computing. In Emerging Intelligent Data and Web Technologies (EIDWT), 2011 International Conference on (pp. 49-54). IEEE.

[17] Sitaram, D., & Manjunath, G. (2012). Chapter 7 - Designing Cloud Security Moving To The Cloud (pp. 307-328). Boston: Syngress.

[18] Popovic, K., & Hocenski, Z. (2010, May). Cloud computing security issues and challenges. In MIPRO, 2010 proceedings of the 33rd international convention (pp. 344-349). IEEE.

[19] Ko, R. K., Jagadpramana, P., Mowbray, M., Pearson, S., Kirchberg, M., Liang, Q., & Lee, B. S. (2011, July). TrustCloud: A framework for accountability and trust in cloud computing. In Services (SERVICES), 2011 IEEE World Congress on (pp. 584-588). IEEE.

[20] A Vouk, M. (2008). Cloud computing–issues, research and implementations. CIT. Journal of Computing and Information Technology, 16(4), 235-246.

[21] Zissis, D., & Lekkas, D. (2012). Addressing cloud computing security issues. Future Generation Computer Systems, 28(3), 583-592. pp.163-275.

[22] Onankunju, B. K. Access Control in Cloud Computing. International Journal of Scientific and Research Publications, Volume 3, Issue 9, September 2013 1 ISSN 2250-3153

[23] Weiss, A. (2007). Computing in the clouds. networker, 11(4).

[24] Dorey, P. G., & Leite, A. (2011). Commentary: Cloud computing–A security problem or solution?. information security technical report, 16(3), 89-96.

[25] Gonzalez, N., Miers, C., Redigolo, F., Simplicio, M., Carvalho, T., Näslund, M., & Pourzandi, M. (2012). A quantitative analysis of current security concerns and solutions for cloud computing. Journal of Cloud Computing, 1(1), 1-18.

[26] Youseff, L., Butrico, M., & Da Silva, D. (2008, November). Toward a unified ontology of cloud computing. In Grid Computing Environments Workshop, 2008. GCE'08 (pp. 1-10). IEEE.

[27] Takabi, H., Joshi, J. B., & Ahn, G. J. (2010). Security and Privacy Challenges in Cloud Computing Environments. IEEE Security & Privacy, 8(6), 24-31.

[28] Koh, Y., Knauerhase, R. C., Brett, P., Bowman, M., Wen, Z., & Pu, C. (2007, April). An Analysis of Performance Interference Effects in Virtual Environments. In ISPASS (pp. 200-209).

[29] Bhadauria, R., Chaki, R., Chaki, N., & Sanyal, S. (2011). A survey on security issues in cloud computing. arXiv preprint arXiv:1109.5388.

[30] Intel Corporation, “Delivering Application-Level Security at Data Centre Performance Levels,” http://download.intel.com/netcomms/technologies/security/320923.pdf , 2008.

[31] Google, “Browser security handbook,” http://code.google.com/p/browsersec /, 2009.

[32] Jensen, M., Schwenk, J., Gruschka, N., & Iacono, L. L. (2009, September). On technical security issues in cloud computing. In Cloud Computing, 2009. CLOUD'09. IEEE International Conference on (pp. 109-116). IEEE.

[33] Subashini, S., & Kavitha, V. (2011). A survey on security issues in service delivery models of cloud computing. Journal of Network and Computer Applications, 34(1), 1-11.

[34] Hacker4Lease, “Backdoor and Debug Options,” http://www.hacker4lease.com/attack-methods/backdoor / , 2011.

[35] Almond, C. (2009). A practical guide to cloud computing security. A white paper from Accenture and Microsoft.

[36] Yang, J., & Chen, Z. (2010, December). Cloud computing research and security issues. In Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on (pp. 1-3). IEEE.

[37] Mead, N. R., & Stehney, T. (2005). Security quality requirements engineering (SQUARE) methodology (Vol. 30, No. 4, pp. 1-7). ACM.

[38] Lombardi, F., & Di Pietro, R. (2011). Secure virtualization for cloud computing. Journal of Network and Computer Applications, 34(4), 1113-1122.

[39] Wen, H., Hai-ying, Z., Chuang, L., & Yang, Y. (2011, August). Effective load balancing for cloud-based multimedia system. In Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference on (Vol. 1, pp. 165-168). IEEE.

[40] Anderson, E. W., Fornell, C., & Lehmann, D. R. (1994). Customer satisfaction, market share, and profitability: findings from Sweden. The Journal of Marketing, 53-66.

A Hybrid Heuristic/Deterministic Dynamic Programing Technique for Fast Sequence Alignment

Talal Bonny
Department of Electrical and Computer Engineering
College of Engineering
University of Sharjah, UAE

***Abstract*—Dynamic programming seeks to solve complex problems by breaking them down into multiple smaller problems. The solutions of these smaller problems are then combined to reach the overall solution. Deterministic algorithms have the advantage of accuracy but they need large computational power requirements. Heuristic algorithms have the advantage of speed but they provide less accuracy. This paper presents a hybrid design of dynamic programing technique that is used for sequence alignment. Our technique combines the advantages of deterministic and heuristic algorithms by delivering the optimal solution in suitable time. we implement our design on a Xilinx Zynq-7000 Artix-7 FPGA and show that our implementation improves the performance of sequence alignment by 63% for in comparison to the traditional known methods.**

***Keywords*—Sequence alignment, dynamic programing, Performance, optimization, FPGA**

I. Introduction

A large computing problem can be handled using dynamic programing. The programing that is a method for solving a complex problem by breaking it down into a collection of simpler sub-problems. This method appears to be both very precise and efficient. However, it does require a very large amount of computational power. An example of using dynamic programing is sequence comparison in computational linguistics [5], [9], [15]. In this field, researchers have established that almost all European languages are related and belong to a single family. Genetically related languages originate from a common proto-language. In the absence of historical records, proto-languages have to be reconstructed from surviving cognates. Sequence comparison is used to find the cognates in large database of divergent languages to reconstruct their proto-form and consequently, to reconstruct an entire proto-language which is an extremely time-consuming process that has yet to be accomplished for many language families.

Image processing is also using dynamic programing for Sequence comparison to retrieve information on handwritten document images [23]. Each word image is represented as a sequence of graphs and the similarity between word images is measured.

In Biology [10], [13], [14], dynamic programing is used to search a large database of sequences for close matches to particular sequence of interest, typically a recently discovered protein. If correlations are found, new drugs may be developed or better techniques invented to treat the disease.

Deterministic algorithms, such as Needleman-Wunsch [3] (for global alignment) and Smith-Waterman [4] (for local alignment), guarantee the return of the optimal alignment of two sequences. The first one is called query sequence (Q) and the second one is called database sequence (D). These kinds of algorithms take a long time to find the highest similarity score as the computing and memory requirements grow proportionally to the product of the lengths of the two sequences being compared, i.e, if n is the length of the query sequence and m is the length of the database sequence, then the previous algorithms provide the optimal alignment (highest similarity score) in n x m steps. Therefore when searching a whole database the computation time grows linearly with the size of the database. Heuristic algorithms, such as FASTA [1] and BLAST [2], provide an approximated solution by comparing query sequence to database sequence and calculating the statistical significance of matches. An approximation is obviously faster than an optimal solution provided by deterministic algorithms since less and easier calculations need to be performed, but it is less accurate because it might miss one or more unexpected but important homologies that would be found in the exact solution.

Various methods and techniques have been proposed to improve the speed of implementations of such algorithms [7]:

In [11], the authors implemented the Recursive Variable Expansion (RVE) based technique, which is proved to give better speedup than any best dataflow approach at the cost of extra area. The authors computed a block of (k x k) elements in parallel instead of computing one element on the FPGA. Compared to dataflow approach, their implementation was 2.29 times faster at the expense of 2.82 times more area.

In [12], the authors developed new tool, called SWIPE, for sequence alignment based on the Smith-Waterman algorithm. In SWIPE, residues from 16 different database sequences are processed in parallel and compared simultaneously to the same query residue. The operations are carried out using vectors consisting of 16 independent bytes. The 16 residues are fed into sixteen independent channels. When the first of these sixteen database sequences ends, the first residue of the next database sequence is loaded into the channel. SWIPE was found to be performing at a speed of 106 GCUPS with a 375 residue query sequence on a dual Intel Xeon X5650 six-core processor system. The authors achieved over six times more rapid than software based on Farrar's 'striped' approach [6].

In [16], the authors used GPU and CPU to improve the performance of aligning the sequences by running the long sequences

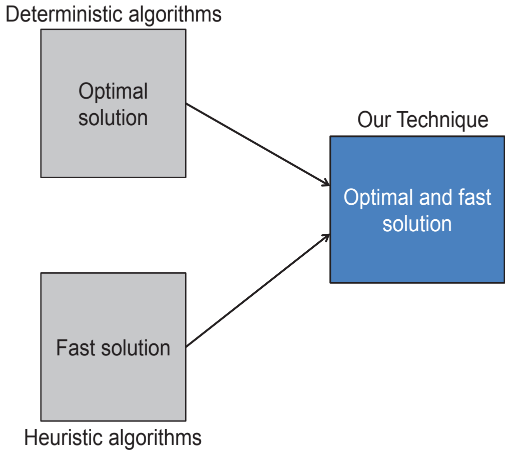


Fig. 1: Our technique combines the advantages of deterministic and heuristic algorithms

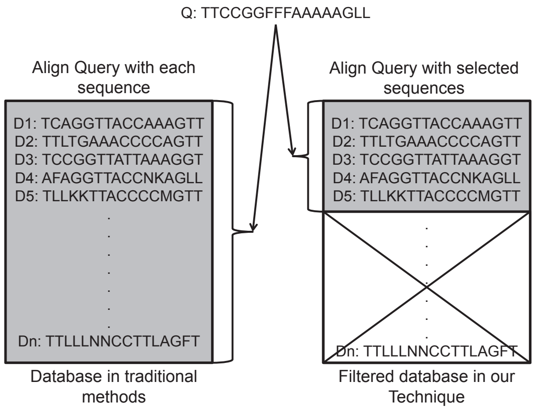


Fig. 2: Sequence alignment in traditional methods and in our technique

on the GPU and the short ones on the CPU. Another different methods were introduced for heuristic sequence alignment.
In [18], the authors proposed, a parallel algorithm that uses GPU to align huge sequences. Their algorithm (CUDAlign 2.1) combined the Smith-Waterman algorithm with linear space complexity. In order to achieve that, they proposed optimizations which are able to reduce significantly the amount of data processed, while enforcing full parallelism most of the time. Using the NVIDIA GTX 560 Ti board and comparing real DNA sequences that range from 162 KBP (Thousand Base Pairs) to 59 MBP (Million Base Pairs), they showed that CUDAlign 2.1 is able to produce the optimal alignment between the chimpanzee chromosome 22 (33 MBP) and the human chromosome 21 (47 MBP) in 8.4 hours.

In all previous work and applications, one object has to be searched/compared/aligned with all objects in the database to find the most closest one by using deterministic or heuristic algorithm. The object might be database sequence, string file, video stream, website page, etc. We have found that to get the optimal alignment, we do not need to apply the alignment algorithms on the whole database sequences. Instead, many database sequences may be excluded from the searching process. In this case, the alignment algorithms may only be applied on the remaining sequences of the database. This will reduce the searching scope and consequently the time required to find the optimal alignment.

In this work, an efficient hybrid design of dynamic programing technique is introduced. It has the advantage of deterministic algorithms, which is delivering optimal solution, and the advantage of heuristic algorithms, which is delivering fast solution, to provide the optimal solution in suitable time (see Fig. 1). Our technique uses new criteria to measure the difference score for each sequence of the database. It excludes the sequences which have high difference scores from the searching process and applies the dynamic programing algorithm (Needleman-Wunsch or Smith-Waterman) on part of the database and not on the whole of it.
Using our technique, we explicitly improve the time performance of the database sequence comparison applications by 63% in comparison to the traditional methods used. Applying our technique in conjunction with the previous state-of-the-art methods (as in [17], [20], [22]) will further improve their time performance.

The rest of the paper is organized as follows. In Section 2, we introduce the traditional methods for database sequence computing applications using Needleman-Wunsch algorithm and then introduce our technique. The complexity of our technique is introduced in Section 3. In Section 4, we present the hardware implementation of our technique on a FPGA. Experimental results are presented in Section 5. We conclude this paper in Section 6.

II. TYPICAL DATABASE SEQUENCE COMPUTING APPLICATIONS

In any database sequence computing application, the data must be coded into a set of sequences using a finite alphabet of states. Then, a cost matrix must be defined on these states and a gap cost scheme must be chosen. This matrix contains the cost of each operation required to transform one of a pair of sequences into the other. The operations used for transformation are insertion, deletion and replacement. An alignment algorithm is then applied on the database sequence and the compared one, resulting in a matrix of distances between all pairs of sequences. The database sequence alignment method is described in details, in the next sections, using the traditional methods and using our technique.

A. SEQUENCE ALIGNMENT USING TRADITIONAL METHODS

Traditional methods for aligning sequences are based on align the query sequence [1] ‘Q’ with each sequence of the database starting from the first sequence ‘D1’ of the database till the last one ‘D_n’ (see left part of Fig. 2). In each alignment process, a score to each cell comparison between the two sequences is computed. The score is based on the result of the comparison, which is either match, or mismatch. If the sequences are mismatched, then one of three operations may be

[1] the query is the searched sequence in the database or the compared sequence with other sequences in the database

done: insertion, deletion, or substitution. Gaps may be added to one or both sequences to make them close to each others Each of these operations has a previously defined score.
For each alignment process between the query and the database sequence, an alignment score (AS) is computed as following:

$$\begin{aligned} AS \;=\; & \textit{(\# of matches} \times \textit{match\_score)} \\ & + \textit{(\# of gaps} \times \textit{gap\_score)} \\ & + \textit{(\# of mismatches} \times \textit{mismatch\_score)} \end{aligned} \tag{1}$$

Usually, the match score is positive but the mismatch and the gap scores are negative. Therefore, more number of matches increases the alignment score but more number of gaps or mismatches decreases the alignment score. The scores of match, mismatch and gap are given as input parameters of the alignment algorithm.
The optimal number of matches, mismatches and gaps are computed using the Needleman-Wunsch algorithm [3] or Smith-Waterman [4] algorithm. In any algorithm, a scoring matrix of size ”m x n” (m is the length of the query sequence and n is the length of the database sequence) is first formed. The optimal score at each matrix element is calculated by adding the current match score to previously scored positions and subtracting gap penalties. Each matrix element may have a positive, negative or 0 value. For two sequences, query (Q) and database (D)
$Q = a_1a_2..................a_m$
$D = b_1b_2.................b_n$
where $H_{ij} = T(a_1a_2.............a_m, b_1b_2............b_n)$ then the element at the (i,j)th position of the matrix H_{ij} is given by

$$H_{i,j} = max \begin{cases} H_{i-1,j-1} + S \\ H_{i-1,j} - G_x \\ H_{i,j-1} - G_y \end{cases} \tag{2}$$

Where H_{ij} is the score at position i in the sequence Q and position j in the sequence D. S is the score of match or mismatch. Gx is the penalty for a gap of length x in the sequence Q and Gy is the penalty for a gap of length y in the sequence D. After the matrix is filled up, to determine an optimal alignment of the sequences from scoring matrix, a method called trace back is used. The trace back keeps track of the position in the scoring matrix that contributed to the highest overall score found. The positions may align or may be next to a gap, depending on the information in the trace back matrix. There may exist multiple maximal alignments.
The time required to get the optimal alignment for two sequences (the query sequence and just one sequence of the database) is proportional to the product of the lengths of the two sequences being compared, i.e, n x m steps.

B. SEQUENCE ALIGNMENT USING OUR TECHNIQUE

The database of the sequence computing applications contains large number of sequences (as explained in Section I). To align the query sequence ‘Q’ with each sequence of the database ‘D’, we need to apply Needlman-Wunsch algorithm on each pair of sequences (as explained in Section II-A). Our technique is based on filtering the database such that the database sequences which are not similar (not close) to the query are excluded from the searching and the Neddleman-Wunsch algorithm is applied only on the database sequences which are similar (close) to the query (see right part of Fig. 2).

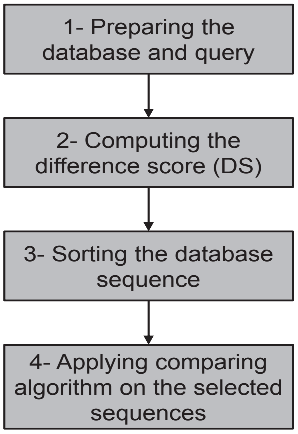

1- Preparing the database and query
2- Computing the difference score (DS)
3- Sorting the database sequence
4- Applying comparing algorithm on the selected sequences

Fig. 3: Steps of our technique

Our technique passes through the following four steps (see Fig. 3):

1- Preparing the database for the comparison:
To prepare the database, we propose our new similarity measure, we call it similarity function. This function is based on the mathematical parameters: frequency and standard deviation of the alphabet codes for each database sequence. These codes have been created previously (as explained in Section II).
The frequency similarity function (Freq) is the number of repeated code in the sequence. It is an indicator for the similarity between two sequences. For instance, if the frequency of the code in a sequence is close to its frequency in another sequence, this is a good indicator that the two sequences might be similar.
To find the frequency for each code in the sequence, we scan the sequence starting from the first code till the last one using number of counters equal to the different codes. One counter for each code. Each counter is incremented by one when it meets new code of the same type. By the end of the scanning, all counters save their values beside the database sequence. The counter values beside any sequence refer to the frequency of codes types for that sequence.
The frequency similarity function does not give always correct results. For example, if the two sequences have the same (or close) number of codes frequencies but the codes are distributed in different way between the two sequences. In this case, the frequency score is not correct score to measure the similarity.
Therefore, we propose another similarity function which is the standard deviation.
The standard deviation function (STD) shows the distribution of the code in a sequence. It gives an idea of how close the entire code of a sequence to the average value. Code with large standard deviation has data spread out over a wide range of values. The standard deviation (STD) is defined

mathematically as:

$$STD = \sqrt{\frac{\sum_{i=1}^{i=n}(X_i - \bar{X})^2}{n-1}}$$

Where $\bar{X}$ is the the average value.

For each database sequence, the frequency and standard deviation functions for each code are computed and stored beside it. This step may take long time as the database includes large number of sequences, but it is done off-line, i.e., independent from the comparison process. Therefore, it does not matter how long time it takes because we do it only one time and prepare the database for future comparison process.

2- Computing the difference score (DS):
Once a query sequence needs to be searched in (aligned with) all database sequences, the technique computes the similarity functions ”Freq” and ”STD” of all different codes for it.
Usually, the sequence has more than one different codes. To find if there is similarity between two sequences each have different codes, the technique computes the frequency difference score (FDS) and the standard deviation difference score (DDS) and add them together to compute the difference score (DS) which reflects the similarity between two sequences.
Difference Score (DS) = frequency difference score (FDS) + standard deviation difference score (DDS)
The frequency difference score (FDS) is the sum of the absolute values of the differences between the two sequences for each code type. Mathematically, the frequency difference score (FDS) between the query sequence ‘Q’ and the database sequence ‘D’ is defined as following considering that both sequences have ‘n’ alphabet codes:

$$\begin{aligned} \textit{FDS} \;=\;& |Freq\_code\_1(Q) - Freq\_code\_1(D)| \\ & +|Freq\_code\_2(Q) - Freq\_code\_2(D)| \\ & +|Freq\_code\_3(Q) - Freq\_code\_3(D)| \\ & +\ldots \\ & +\ldots \\ & +|Freq\_code\_n(Q) - Freq\_code\_n(D)| \end{aligned} \tag{3}$$

Where ”$Freq\_code\_1(Q)$” is the frequency of code 1 in the query sequence. ‘”$Freq\_code\_1(D)$” is the frequency of code 1 in the database sequence, etc.
The standard deviation difference score (DDS) is the sum of the absolute values of the differences between the two sequences for each code type. Mathematically, the standard deviation difference score (DDS) between the query sequence ‘Q’ and the database sequence ‘D’ is defined as following considering that both sequences have ‘n’ alphabet codes:

$$\begin{aligned} \textit{DDS} \;=\;& |Dev\_code\_1(Q) - Dev\_code\_1(D)| \\ & +|Dev\_code\_2(Q) - Dev\_code\_2(D)| \\ & +|Dev\_code\_3(Q) - Dev\_code\_3(D)| \\ & +\ldots \\ & +\ldots \\ & +|Dev\_code\_n(Q) - Dev\_code\_n(D)| \end{aligned} \tag{4}$$

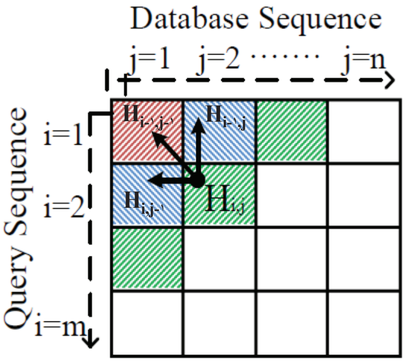


Fig. 4: Computation sequence of the similarity matrix. the score of the cells which have the same color are computed together

Where ”$Dev\_code\_1(Q)$” is the standard deviation of code 1 in the query sequence. ”$Dev\_code\_1(D)$” is the standard deviation of code 1 in the database sequence, etc.
Our technique is based on filtering the database such that the database sequences which are not similar (not close) to the query are excluded from the searching and the optimal or the heuristic alignment algorithm is applied only on the database sequences which are similar (close) to the query.

3- Sorting the database sequences:
In this step, the database sequences are sorted according to their difference score (DS), i.e. (FDS + DDS), such that the sequences which have low difference scores (more close to the query sequence) are shifted to the top of the database.

4- Applying the Alignment Algorithm:
In the last step, our technique applies the alignment algorithm (in our case the Needleman-Wunsch Algorithm) only on the sequences, which have the low difference scores (selected in the previous step). The sequences which have high difference scores will be excluded. This will provide the alignment in reasonable time because the alignment algorithm is applied only on a part of the database and not on whole of it.

III. COMPLEXITY OF OUR TECHNIQUE

In this section, we compare between the complexity of the traditional methods and our technique. In case of the traditional methods, when Needleman-Wunsch Algorithm is used, the complexity is based on the length of the two sequences being compared and the number of sequences in the database. Let m, n are the lengths of the query sequence Q and the database sequence D, respectively. As the Needleman-Wunsch Algorithm is based on dynamic programing, then the complexity to perform the alignment for one sequence is $O(m \times n)$. If s is the number of sequences in the database, then the total complexity will be $O(m \times n \times s)$.
In case of our technique, assuming we have c different codes. to compute the distribution of the c codes in the query sequence, we need to scan the query along its length. If the length of the query sequence is m, then we need m step to perform the scan. To compute the difference score (DS) between the query and one database sequence, we need c steps to perform the subtraction for the c codes and $c-1$ steps to sum up the results. If s is the number of sequences in the database,

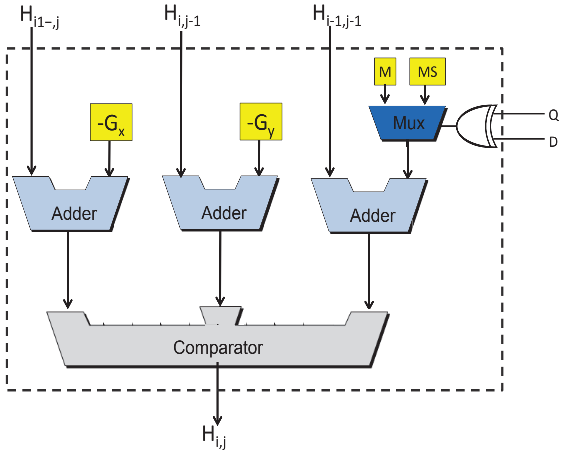


Fig. 5: The design implementation of our processing element for sequence alignment

then we need $((2c-1) \times s)$ steps to compute the difference score. To sort the s difference scores from the smallest to the largest one using Heap sort or merge sort algorithm, we need $(s \times logs)$ steps.

Assuming that 50% of the database sequences are selected to apply Needleman- Wunsch Algorithm on them. To perform this step we need $m \times n \times s/2$ steps.

Consequently, the total steps of our technique is $((m + (2c - 1) \times s + s \times logs) + (m \times n \times s/2))$, i.e., the complexity is $O(m \times n \times s/2)$.

For instance, if the length of the query m = 500, and the number of the database sequences s = 10000. Each sequence of the database has length n = 500. To align the query sequence with the database sequences using the traditional method, we need: 500 x 500 x 10000 = 2500000000 steps (2500 Million steps). Using our technique, and assuming that the data are codded with 4 different codes, we need:

500 + (7 x 10000) + (10000 x log 10000)+ (500 x 500 x 5000) = ≈ 1250 Million steps.

Using our technique we save 50% of the time required to align the sequences using the traditional methods.

IV. Hardware Implementation

As explained in Section II-A, to implement the alignment algorithm, we need first to form a similarity matrix of size ”m x n” using the Equation 2 and then we need to trace the scores in the matrix back. The time complexity of the alignment algorithm is O(MN). To reduce this complexity, multiple entries of the matrix are calculated in parallel. From the equation 2, we can determine that the score of any cell, H(i,j), in the matrix depends on the scores of the three other elements (see Fig. 4): The left neighbor, H(i,j-1), the up neighbor, H(i-1,j), and the up-left neighbor, H(i-1,j-1). This means that the score of any cell can not be computed before computing the scores of cells located on the anti-diagonal positions. Fig. 4 shows the sequence of the cell computation in the similarity matrix. All the cells located on the same anti-diagonal positions (have the same color) are computed together (simultaneously) because they are independent of each other.

To measure the performance of the alignment implementation, the Cell Updates per Second (CUPS) metric is commonly used [8], which represents the time required to complete the computation for one cell of the similarity matrix. The total number of cell updates gives the implementation performance of the sequence alignment algorithm:

$$\mathit{Performance\ (CUPS)} = \frac{\mathit{size(Query)\ x\ size(Database)}}{\mathit{Time\ to\ complete\ the\ computation}} \tag{5}$$

We implement our technique using a FPGA-based linear systolic array to reduce the complexity order of the computation. A linear systolic array [21] is an array of processing cores where each cell shares its data with the other cells in the array. Each processing core solves a subproblem and shares the solution to all other cells in the array to prevent calculation of the same problem twice. Each anti-diagonal has M cells, and each cell can be updated in parallel, so the systolic array consists of M Processing Elements (PEs) that each computes a new value for the cell in the row that they are responsible for [19].

We design the processing element (PE) of systolic array using the FPGA logic components. The PE is used to build a systolic array architecture of any size. Fig 5 shows the design implementation of our processing element. Assuming the sequences we are going to align are DNA sequences. Each sequence consists of four codes (letters): A, C, G, T. In this case, the two sequences Q and D are encoded using two bits for each code (letter) as following:

A: 00, C: 01, G: 10, T: 11.

Based on the equation 2, the PE receives the values of the three neighbor elements, the left neighbor, H(i,j-1), the up neighbor, H(i-1,j), and the up-left neighbor, H(i-1,j-1). Each PE also receives one code (2 bits) from each of the two sequences (Q and D). The PE has four given parameters (marked in yellow blocks in the figure) which are fixed for all PEs. These parameters are the scores of match (M), mismatch (MS), gap in sequence Q (G_x), and gap in sequence D (G$_y$). The PE calculates the outcome of the matrix element H(i,j) using XOR gate, one multiplexer, three adders, and one comparator. The multiplexer chooses match or mismatch score based on the similarity of the two codes compared, and sends the score to the adder. The comparator selects the maximum value of the three adders according to Eq. 2.

We implement our design on a Xilinx Zynq-7000 All Programmable SoC (XC7Z020-CLG484) Artix-7 FPGA from Digilent [26]. Each PE utilizes 19 slices. The XC7Z020 has 13300 slices (1.3 M ASIC gates), i.e., we can fit a maximum of 700 PEs on this FPGA chip. In our experimental results, we utilize only 400 PEs as the sequences we align are of length 400 nucleotides, as explained in the next section.

V. EXPERIMENTAL RESULTS

In this section, the experimental results of our technique are presented. To evaluate our technique, DNA sequences of the database DNA Data Bank of Japan (ddbj) [24] are used. 100 sequences of BCT and CON divisions are selected as case study. Each sequence has length of 400 nucleotides. The accession numbers of the selected sequences start with the ID ”AB”, ”AF”, ”AJ”, ”AM”, ”AY” and ”DQ”. We compare our technique with the traditional methods which use the first widely used program for optimal sequence alignment

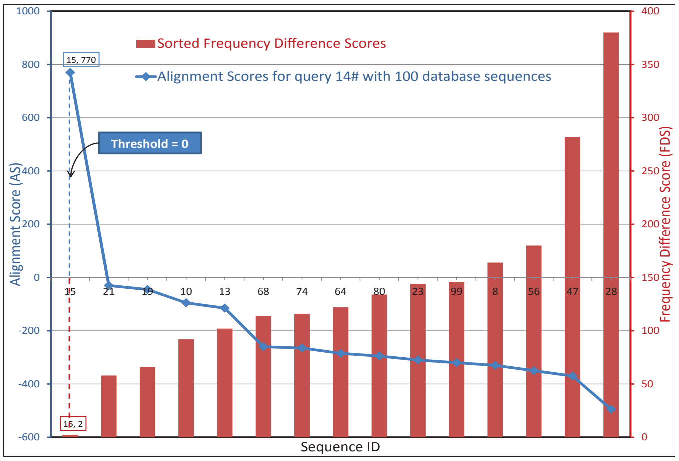


Fig. 6: Alignment and frequency difference scores for the query (sequence number 14) with 100 database sequences

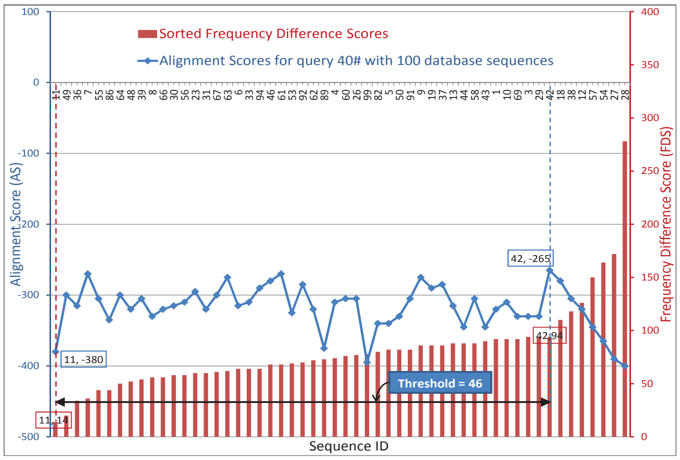


Fig. 8: Alignment and frequency difference scores for the query (sequence number 40) with 100 database sequences

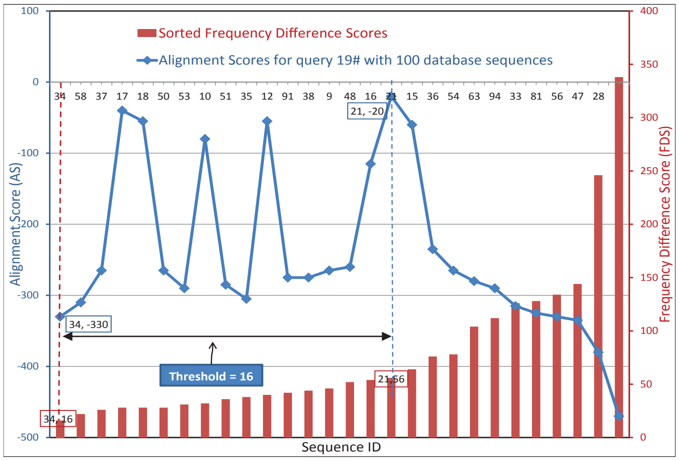


Fig. 7: Alignment and freuency difference scores for the query (sequence number 19) with 100 database sequences

Needleman-Wunsch Algorithm [25]. The score of match, mismatch, gap open, and gap extend are selected to be +2, -3, 0, -4, respectively.

As our database has 100 sequences, 100 cases are tested considering different query sequence for each case, i.e., in the first case, we consider the first sequence as query sequence and the remaining sequences as database. In the second case, the second sequence is considered as query sequence and the remaining sequences are considered as database, and so on. The results are presented for the three similarity functions, frequency (Freq), Standard Deviation (STD) and combination of both functions (Freq + STD):

1- Using the frequency similarity function (Freq):

The experimental results when our technique uses (Freq) are presented in Figures 6-9.

Figures 6, 7, and 8 show the results for three selected tested cases, case 14, case 19 and case 40, respectively. In these figures, the left y-axis shows the alignment score (AS) using Needleman-Wunsch Algorithm. The right y-axis shows the frequency difference score (FDS) of our technique. The x-axis shows the ID of the database sequences.

The alignment score (AS) and the frequency difference score (FDS) in these figures are computed for the query sequence (sequence number 14 in Fig. 6, sequence number 19 in Fig. 7, and number 40 in Fig. 8) with each sequence of the database. Then, The sequences are sorted based on their frequency difference score in ascending form, i.e., from the sequence which has the smallest difference score to the one which has the highest score.

The common result in these three figures is, when the frequency difference score (FDS) increases across the sequences, the alignment score (AS) decreases (or vice versa). This result shows that the criterion we use in our technique, for selecting the sequences to which we may apply Needleman-Wunsch Algorithm on instead of the whole database sequences, is correct. This is because the sequences which have low difference scores have high alignment scores. In figures 6, 7, and 8, we define new parameter called ”Threshold”. The threshold for any test case is the number of sorted sequences between the sequence which has the lowest frequency difference score (FDS) and the highest alignment score (AS). In other words. The threshold refers to the number of sequences we need to apply Needlman-Wunsch Algorithm on them (using our technique) instead of applying it on the whole database sequences (using the traditional methods).

In Fig. 6, the frequency difference score (FDS) curve is marked with a red label ”15,2”. This label means that the minimum frequency difference score, ‘2’, occurs at the sequence number ‘15’. The alignment score curve is marked with the blue label ”15,770”. It means that the maximum alignment score ‘770’ occurs at the sequence number ‘15’. In other words, the sequence number ‘15’ has the lowest frequency difference score (FDS) and the highest alignment score (AS) and the threshold in this case is ‘0’. From this figure, we conclude that using our technique implies that applying Needlman-Wunsch Algorithm on the sequence number 15 is enough to get the optimal alignment instead of applying it on the whole database sequences (using the traditional methods). This is the best case we get, but unfortunately it is not always the same for all cases.

In Fig. 7, the lowest frequency difference score, ‘16’, occurs at the sequence number ‘34’ which has alignment score equal to ‘-330’ (this alignment score is not the highest one). On the

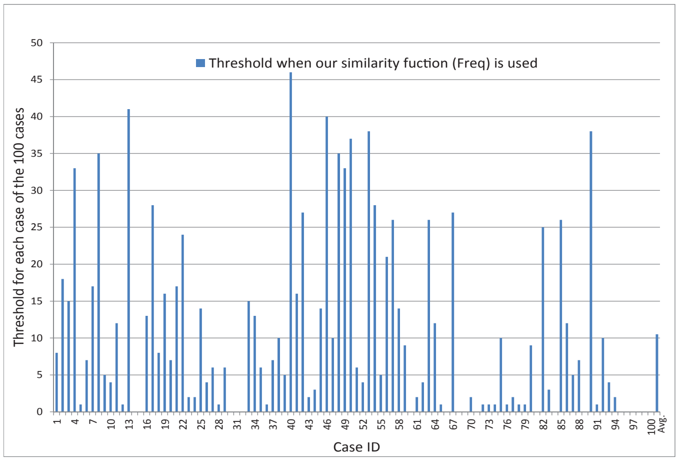


Fig. 9: The threshold for each case of the 100 cases when our similarity function (Freq) is used

other hand, the highest alignment score, '-20', occurs at the sequence number '21' which has frequency difference score equal to '56'. The threshold in this case is '16'. In other words, we need to apply Needlman-Wunsch Algorithm on the lowest 16 frequency difference score sequences instead of applying it on the whole database sequences. In Fig. 8, the lowest frequency difference score, '14', occurs at the sequence number '11' which has alignment score equal to '-380' (this alignment score is not the highest one). On the other hand, the highest alignment score '-265' occurs at the sequence number '42' which has frequency difference score equal to '94'. The threshold in this case is '46', i.e., we need to apply Needlman-Wunsch Algorithm on the lowest 46 frequency difference score sequences instead of applying it on the whole database sequences. From the previous three cases, we notice that the threshold is not fixed. To find the maximum threshold, the test has to be done for all cases. Fig. 9 shows the thresholds for all 100 cases. The last bar shows the average threshold through all cases which is equal to '10.5'. In this figure, all 100 cases are tested to find the highest alignment score and the lowest frequency difference score for each case (one case for each different query), and then the threshold were computed. From Fig. 9, we notice that the threshold differs from one case to another one based on the query sequence. And, the maximum threshold among all other cases is '46' which appears in the case number 40 (as shown in Fig. 8). This is the worst case in which we need to apply Needleman-Wunsch Algorithm on 46 sequences. In other words, when our technique is applied only on the top 46% of the database sequences (or in general case 50%), then the maximum AS, in each case, will be included in this top part, i.e., applying Needleman-Wunsch Algorithm on this 50% of the database sequences will be enough to find the maximum alignment score instead of applying the algorithm on the whole database sequences as done by traditional methods.

2- Using the standard deviation similarity function (STD):
If the two sequences being compared/aligned have the same (or close) number of codes frequency but the codes are distributed in different way between the two sequences, then, the frequency difference score (FDS) will not be correct score to measure the similarity. Therefore, we use the standard deviation similarity function (STD).

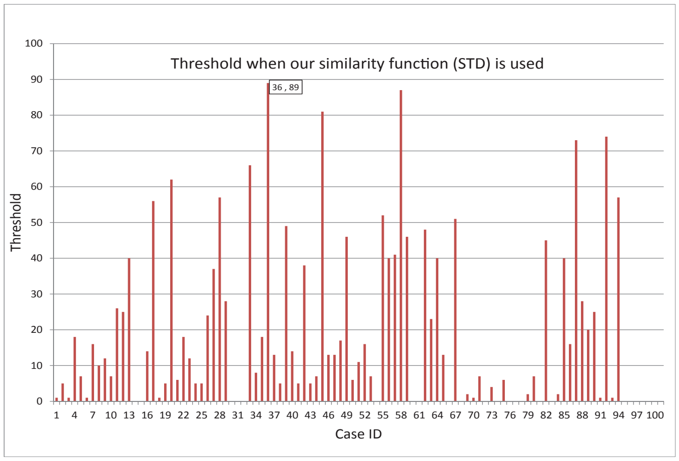


Fig. 10: The threshold for each case of the 100 cases when our similarity function (STD) is used

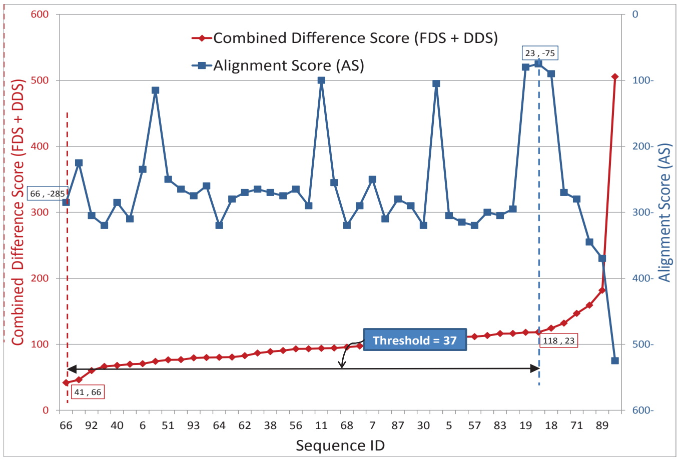


Fig. 11: The worst case threshold (the query is the sequence number 13) when our combined similarity functions (Freq + STD) is used

Figure 10 shows the thresholds for all 100 cases when our technique uses the STD similarity function. In this figure, the maximum threshold (worst case) among all other cases is '89' which appears in the case number 36. This is the worst case in which we need to apply Needleman-Wunsch Algorithm on 89 sequences which is not worthy. From this figure, we can conclude that using the standard deviation similarity function (STD) to measure the similarity is not also a good idea because the two sequences may have the same (or close) standard deviation (STD) but the code frequency for each sequence is different. Therefore, a combination of the two similarity functions (Freq) and (STD) can give better results as shown in the following section.

3- Using the combination of frequency (Freq) and standard deviation similarity functions (STD):
The experimental results when our technique uses the combination of the two similarity functions (Freq+STD) are presented in Figures 11-15.

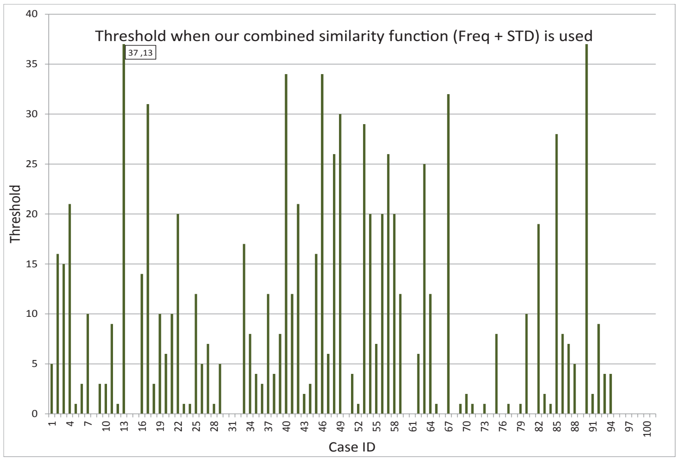


Fig. 12: The threshold for each case of the 100 cases when our combined similarity function (Freq + STD) is used

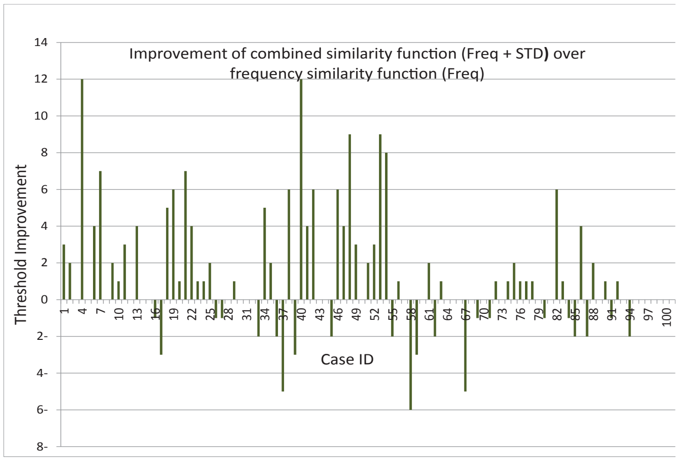


Fig. 13: Improvement of the threshold when the combined similarity functions (Freq + STD) is used instead of the sole frequency similarity function (Freq)

Figure 11 shows The results, for the query (sequence number 14) with 100 database sequences, when our combined similarity function (Freq + STD) is used. In this figure, the right y-axis shows the alignment score (AS) using Needleman-Wunsch Algorithm. The left y-axis shows the combined difference score (FDS + DDS) of our technique. The x-axis shows the ID of the database sequences. The combined difference score (FDS + DDS) curve, in this figure, is marked with a red label ”66, 41”. This label means that the minimum (FDS + DDS) difference score, ‘41’, occurs at the sequence ‘66’. The alignment score (AS) curve is marked with a blue label ”23,-75”. It means that the maximum alignment score ‘-75’ occurs at the sequence ‘32’. The number of sorted sequences on the x-axis which are located between the minimum frequency difference score sequence and the maximum alignment sequence (i.e. ”Threshold”) is 37.

Figure 12 shows the thresholds for all 100 cases when our technique uses the combined similarity function (Freq + STD). In this figure, the maximum threshold (worst case) among all other cases is ‘37’, which appears in the case number 13. This means, when our technique is applied only on the top 37% of the database sequences, then the maximum AS, in any query case, will be included in this top part, i.e., applying Needleman-Wunsch Algorithm on this 37% of the database sequences will be enough to find the maximum alignment score instead of applying the algorithm on the whole database sequences as done by traditional methods. Using the combined similarity function (Freq + STD) gives better results than using the sole frequency function (Freq). Figure 13 shows the threshold improvement when the combination is used. In this figure, 47 cases are improved (bars located in the positive area). The maximum improvement is ‘12’ which appears in the case number 40 (the threshold of this case number is ‘46’ for frequency similarity function (Freq) and it becomes ‘34’ for the combined similarity function (Freq + STD)). There are 32 cases where the threshold is ’0’ using the (Freq) function and they remain the same in the (Freq + STD) function. The remaining 21 cases are changed negatively (bars located in the negative area). When our technique applies Needleman-Wunsch Algorithm on less than 37 sequences of the database and repeated for 100 cases (each case with different query), then the result will not be correct for all the 100 cases, i.e., the sequence which has the lowest difference score is not the same as the sequence which has highest alignment score. The results differ based on the number of removed sequences from the database.

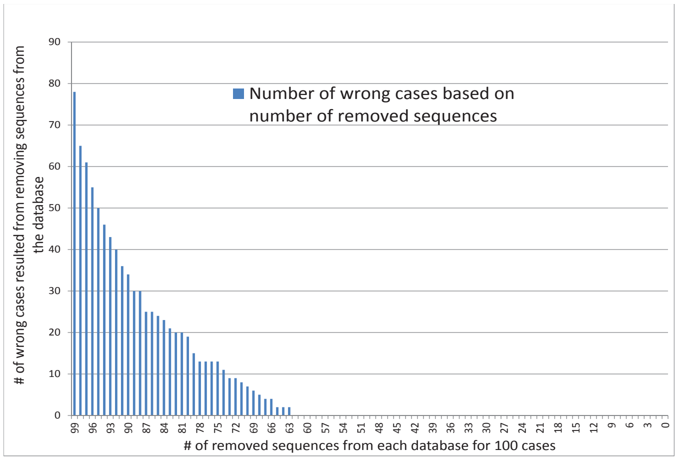


Fig. 14: The error rate resulted from removing sequences from the database

Fig. 14 shows the error rate resulted from removing sequences from the database. The x-axis shows the number of removed sequences from each database for 100 cases (for clarity, we do not show all 100 cases). The y-axis shows the number of wrong cases resulted from removing sequences from the database. For example, when 99 sequences are removed from the database and our technique is repeated for the 100 cases, there will be ‘78’ wrong cases and only ‘12’ cases will have correct results, i.e., in each case of the 12 cases, the the sequence, which has the lowest difference score, has the highest alignment score. When the number of removed sequences decreases, the error rate will be decreased and the number of correct cases will be increased. When the number of removed sequences is ‘63’, i.e., only ‘37’ sequences are remained in the database, there will be no wrong cases. This is the best case in terms of the

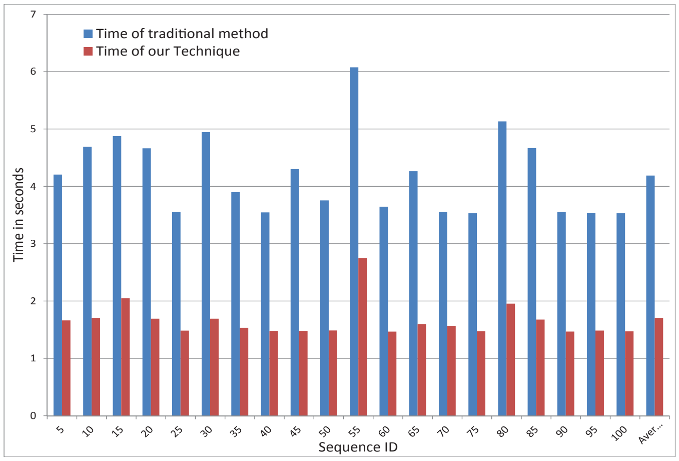


Fig. 15: Execution time comparison between traditional methods and our technique

size of the database and the execution time. Removing less number of sequences will not effect on the result but negatively will increase the size of the database and consequently the time required to analyze it. Fig. 15 shows comparison between the execution time of traditional methods and our technique. In this figure, The x-axis shows the 100 cases (not all cases are shown for clarity purpose). For each case, different query sequence used to be aligned with the remaining sequences of the database. The y-axis shows the execution time required for each case. The blue bar shows the time for traditional methods which apply NW algorithm on whole database sequences while the red one shows the time for our technique which applies NW algorithm on selected 37% of the database sequences. The last bars show the average time through all 100 cases.
In this figure, the execution time using our technique is 63% improved in comparison to the execution time required using the traditional methods. (the average time for traditional methods is 4.18 sec. while the average time for our technique is 1.8 sec.). This result we got because we have excluded selected 63% of the sequences from the process of applying Needleman-Wunsch Algoritm by using our technique.

VI. CONCLUSIONS

We have presented new technique to accelerate the database sequence alignment. Our technique has the advantage of the heuristic and deterministic algorithms that can delivers optimal and fast solution We compared our technique with the traditional methods which apply the alignment algorithms on the whole database sequences and showed that our technique saves almost 63% of the time required to perform the sequence comparing. Merging our technique with the state-of-the-art database computing technique may further improve the execution time.

REFERENCES

[1] European Bioinformatics Institute Home Page, FASTA searching program, 2003. http://www.ebi.ac.uk/fasta33/.

[2] National Center for Biotechnology Information. NCBI BLAST home page, 2003. http://www.ncbi.nlm.nih.gov/blast.

[3] S. Needleman and C. A. Wunsch. General method applicable to the search for similarities in the amino acid sequence of two sequences. Journal of Molecular Biology. Pages 443453, 1970

[4] T. F. Smith and M. S. Watermann. Identification of common molecular subsequence. Journal of Molecular Biology. Pages 196197, 1981

[5] G. Kondrak. Algorithms for Language Reconstruction. Ph.D. thesis, University of Toronto. 2002

[6] Farrar M: Striped Smith-Waterman speeds database searches six times over other SIMD implementations. Bioinformatics 2007, 23:156-161

[7] A. Stivala, P.J. Stuckey, M.G. de la Banda, M. Hermenegildo, and A. Wirth. Lockfree parallel dynamic programming. Journal of Parallel and Distributed Computing, 70(8):839-848, 2010.

[8] Lukasz Ligowski and Witold Rudnicki. An efficient implementation of Smith-Waterman algorithm on GPU using CUDA, for massively parallel scanning of sequence databases. In IEEE International Symposium on Parallel & Distributed Processing. pp. 1-8. 2009.

[9] S. J. Greenhill, Q. D. Atkinson, A. Meade and R. D. Gray. The shape and tempo of language evolution. In proceedings of the Royal Society. Pages 2443-2450, 2010.

[10] Bonny, T., M. A. Z. and Salama, K. N. An adaptive hybrid multiprocessor technique for bioinformatics sequence alignment. In the International Conference on Biomedical Engineering. pages 112-115, 2010

[11] Z. Nawaz, M. Nadeem, J. van Someren, and K.L.M. Bertels. A parallel fpga design of the smith-waterman traceback. In Field-Programmable Technology (FPT), 2010 International Conference on, pages 454-459, Beijing, China, December 2010.

[12] T. Rognes. Faster smith-waterman database searches with inter-sequence simd parallelisation. BMC Bioinformatics, 12(1):221, 2011.

[13] Talal Bonny and Khaled Salama, Fast Global Sequence Alignment Algorithm in the Asilomar Conference on Signals, Systems, and Computers to be held in PACIFIC GROVE, CA, in November 6-9th, 2011

[14] Talal Bonny and Khaled Salama, ABS: Sequence Alignment By Scanning in the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'11), on August 30 - September 3, 2011. Boston, MA, USA

[15] S. Nelson-Sathi, J. List, et. al. Networks uncover hidden lexical borrowing in Indo-European language evolution. In proceedings of the Royal Society. Pages 1794-1803, 2011

[16] M. Affan Zidan, T. B. and Salama, K. N. High performance technique for database applications using a hybrid gpu/cpu platform. IEEE/ACM 21st Great Lake Symposium on VLSI. pages 8590, 2011

[17] A. Chakraborty and S. Bandyopadhyay. Clustering of web sessions by FOGSAA. In IEEE Recent Advances in Intelligent Computational Systems (RAICS). Pages 282-287. 2013

[18] E. F. de O.Sandes and A.C.M.A. de Melo. Retrieving smith-waterman alignments with optimizations for megabase biological sequences using gpu. Parallel and Dis- tributed Systems, IEEE Transactions on, 24(5):1009-1021, 2013.

[19] H. Shah, L. Hasan, and N. Ahmad. An Optimized and Low-cost FPGA-based DNA Sequence Alignment. In the 35th Annual International Conference of the IEEE EMBS. Pages 2696-2699. 2013

[20] S. Kim, Y. J. Yoo, J. So, J. G. Lee and J. Kim., Design and Implementation of Binary File Similarity Evaluation System. International Journal of Multimedia and Ubiquitous Engineering, Vol.9, No.1. Pages 1-10, 2014

[21] J.M. Marmolejo-Tejada, V. Trujillo-Olaya, C.P. Renteria-Mejia and J. Velasco-Medina. Hardware implementation of the Smith-Waterman algorithm using a systolic architecture. In IEEE 5th Latin American Symposium on Circuits and Systems (LASCAS). Pages 1-4, 2014

[22] Manal Al Ghamdi and Yoshihiko Gotoh. Alignment of nearly-repetitive contents in a video stream with manifold embedding. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Pages 1255-1259, 2014

[23] P. Wang, V. Eglin, C. Largeron, J. Llads, A. Fornes And C. Garcia. A Novel Learning-free Word Spotting Approach Based On Graph Representation. In 11th IAPR International Workshop on Document Analysis Systems (DAS). Pages 207-211, 2014

[24] Website: http://www.ddbj.nig.ac.jp/

[25] Website: blast.ncbi.nlm.nih.gov/Blast.cgi

[26] Website: www.digilentinc.com/

AdviseMe: An Intelligent Web-Based Application for Academic Advising

Lawrence Keston Henderson
Department of Computing and Information Technology
University of the West Indies
St. Augustine, Trinidad

Wayne Goodridge
Department of Computing and Information Technology
University of the West Indies
St. Augustine, Trinidad

Abstract—**The traditional academic advising process in many tertiary-level institutions today possess significant inefficiencies, which often account for high levels of student dissatisfaction. Common issues include high student-advisor loads, long waiting periods at advisory offices and the need for advisors to handle a significant number of redundant cases, among others.**

Utilizing semantic web expert system technologies, a solution was proposed that would complement the traditional advising process, alleviating its issues and inefficiencies where possible. The solution coined 'AdviseMe', an intelligent web-based application, provides a reliable, user-friendly interface for the handling of general advisory cases in special degree programmes offered by the Faculty of Science and Technology (FST) at the University of the West Indies (UWI), St. Augustine campus. In addition to providing information on handling basic student issues, the system's core features include course advising, as well as information of graduation status and oral exam qualifications. This paper produces an overview of the solution, with special attention being paid to the its inference system exposed via its RESTful Java Web Server (JWS).

The system was able to provide sufficient accurate advice for the sample set presented and showed high levels of acceptability by both students and advisors. Furthermore, its successful implementation demonstrated its ability to enhance the advisory process of any tertiary-level institution with programmes similar to that of FST.

Keywords—Web-Based Academic Advising; Academic Advising; Ontology; Jena; Expert Systems

I. Introduction

A. Background - Academic Advising

In the realm of tertiary education, academic advising is a student-advisor collaborative process [1] designed to enhance a student's overall educational experience by lending academic decision support to them. This is done by analysing the student's academic records and external factors (academic capabilities, interests, daily schedules and financial constraints) in order to produce customized advice [2]. Such advice would then allow the student to make informed decisions so that they can develop an academic plan to complement their personal life goals and complete their course of study within the prescribed period, or with minimal excess from that date.

The advising process is long-term and iterative due to the continuous change of the environment it operates within [3]. Such changes include the addition and removal of courses from programmes as well as modification of prerequisite rules. It also has timely limitations, as advisors cannot lend advice for future semesters since it is difficult to predict which courses a student will pass, if any during the course of any semester [4]. As a result, student advising should be made available, at minimum, once per semester to ensure that students are guided based on the latest versions of their transcript and the rules that govern their study programmes.

Academic advising can be categorized into four major systematic models: prescriptive, developmental, integrated and engagement. In prescriptive advising, students succumb to the direct advice given by advisors, making advisors solely responsible for the decision making process. With developmental advising however, the advisor directs the student to the proper resources and the decision making process is shared between both parties with more responsibility being placed on the student, thus fostering a higher level of 'student-independence' [5]. Integrated advising is a fusion of formerly discussed methods and engagement advising is typically a type of developmental advising, with increased student-advisor meetings [6]. It was noted however that intuitive students typically endorsed a developmental advising model while others seldom valued a collaborative relationship and hence seemed more content with that of a prescriptive advising model [6].

There are two core methodologies associated with selecting courses based on interests and prerequisites completed. The 'bottom-up' approach is used in most structured programmes whereby an initial set of core courses (with no prerequisites) is selected and course selection per semester continues as these courses are successfully completed. Conversely, in the 'top-down' approach, advanced courses in which the student has interest is initially analysed. The prerequisites of these courses are then derived and the student takes these in order to reach his goal [7]. This approach is used more in unstructured programmes such as general degrees with specializations in particular fields.

It is crucial that a student receive proper advising as poor or no advising can have severe repercussions on his progress throughout his course of study, possibly resulting in delayed graduation [8]. In some cases, the advising process may demand more qualitative judgements before a reasonable decision can be attained. This can be due to personal student issues external to the university's academic context. As a result, academic advising does not only entail course advising, but is also designed to support and motivate students throughout their academic life so that they can comfortably accomplish their educational goals [9]. In this light, advisors must have full knowledge of the student's background, academics, plans and goals in order to give effective academic advice. This may therefore require frequent 'one-on-one' student-advisor meetings so that a relationship can be forged between both parties whereby they understand the student's unique needs. Such advising however can be difficult to achieve mainly due to the availability of experienced and committed persons to undertake the task [8].

B. The Traditional Academic Advising Process

At the University of the West Indies (UWI), St. Augustine campus, integrated academic advising is handled per faculty, with sub-advisory units at each department. In the faculty of Science and Technology (FST), this manual process is spearheaded by the Deputy Dean who has assigned one advisor to each department. Each advisor is expected to handle all student matters under his purview. Also twenty-five peer advisors have been trained to handle general advisory matters inclusive of how to request overrides and determine qualification for oral assessments, among other things. Up to the first month into the semester, advisors are expected to handle student cases for extended periods of time, ranging on average between four to six hours per day. At other times, advisors would conduct advising services generally around one to three hours per day.

The advisory process begins with the student filling (or updating) a paper-based form outlining the courses already taken (if any) by semester, with the corresponding grades obtained. This represents their advising profile and therefore must be kept and maintained throughout their entire academic life. The form would also have comments made by past advisors indicating what advice was given. On meeting with a student, the advisor retrieves the student's transcript from the Student Information System (SIS) to verify that the information on the form is accurate. He then uses a 'bottom-up' methodology, mapping the student's completed courses against those required for the student's programme of study. The course listings are usually taken from references such as faculty handbooks and departmental handouts and are used to deduce what courses the student should take in the upcoming semester.

For cases where students have completed a semester of courses and simply need a new set to attempt for the upcoming semester, the advising process ends following course suggestion. For special cases influenced by external factors however, additional time is required for advisors to learn the student's situation and perform more qualitative analysis before relating sufficient advice to the student.

Advising is also a platform to handle student issues that may arise during the semester. This can include, but is not limited to, learning how to handle override and exemptions, getting information on graduation status, determining qualifications for oral assessments and learning how to handle rescindment of 'Required to Withdraw' status. In an attempt to handle these minor issues, static forms of information that lend solutions were created in an attempt to curb some of the load faced by advisors. These range from paper-based handouts to bulletin boards strategically placed in department offices relaying information on programmes as well as steps to request overrides / exemptions and rescindment of RTW status. Attempts to broadcast such information via social media and departmental websites were also being made.

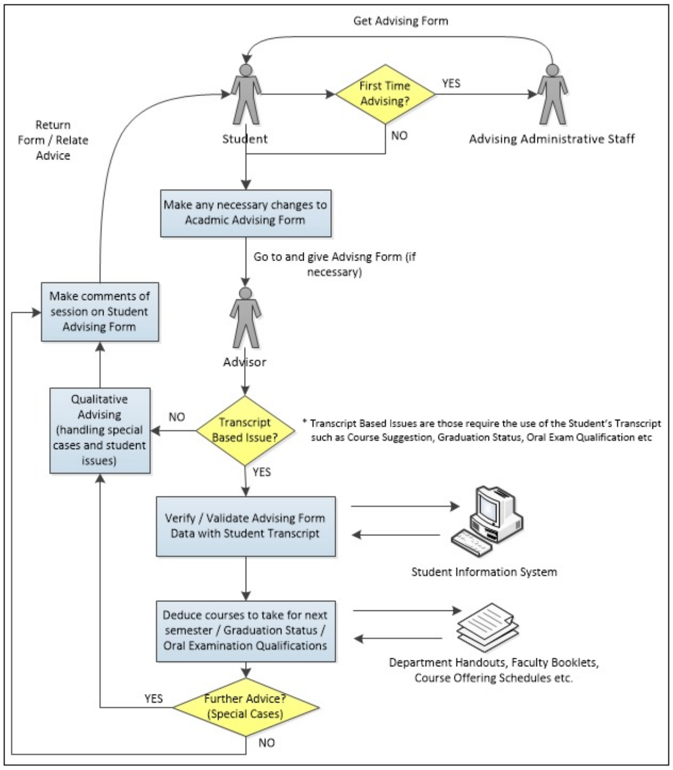

Get Advising Form
Student
First Time Advising?
YES
NO
Advising Administrative Staff
Return Form / Relate Advice
Make any necessary changes to Acadmic Advising Form
Go to and give Advisng Form (if necessary)
Advisor
Make comments of session on Student Advising Form
Qualitative Advising (handling special cases and student issues)
NO
Transcript Based Issue?
* Transcript Based Issues are those require the use of the Student's Transcript such as Course Suggestion, Graduation Status, Oral Exam Qualification etc
YES
Verify / Validate Advising Form Data with Student Transcript
Student Information System
Deduce courses to take for next semester / Graduation Status / Oral Examination Qualifications
Department Handouts, Faculty Booklets, Course Offering Schedules etc.
YES
Further Advice? (Special Cases)
NO

Fig. 1: The Traditional Advisory Process at FST

C. Issues faced in the traditional advising process

Although the process described in section I(B) may seem straightforward, further analysis show several issues that can be termed inefficient or problematic for the overall advising process.

Quality academic advising requires dedicated personnel to be available to handle the task. Financial constraints often make it unfeasible to hire staff solely for advising and hence existing staff are usually assigned this extra task, making their overall duties labour intensive. This is the case in FST as both staff and students are assigned advisory roles within their jurisdiction. While this can cut departmental costs, it raises availability issues as advising times may sometimes clash with advisors primary responsibilities [10] such as teaching and attending classes, thus forcing them to be unavailable for advising.

The quality of academic advising is also affected by the length of time that students are able to meet with advisors. Due to the high number of students per advisor, advisors tend to spend inadequate time with students or rush the advising process so that they can facilitate the load of students faced during the allotted advising sessions [1], [3]. This is often seen during the registration periods where there is often a backlog of students at advisor's offices.

On further analysis of student's issues, it was found that the majority of the advisor's time was spent answering recurrent questions [1], [10] pertaining to handling basic issues such as overrides as well as solving trivial course scheduling issues [2]. Also, when advisors try to spend sufficient time handling special cases, it often results in students having long waiting periods for advising, possibly even getting turned away and having to come again at another date. This can be quite unappealing to students, who may then resort to taking hearsay advice from peers which often leads to students making poor academic choices and then having to meet with an advisor to fix possibly more complicated student issues [10].

Even with tolerable advising loads there still exist many issues that lend to the inefficiency of the advising process at UWI. For example students are not assigned to any particular advisor and hence the possibility exists that they can interact with different advisors at different instances of their academic life. Such advisors would not have full knowledge of the student's background and hence would only be able to give advice based on the information provided to them at the point of advising. While the student should possess their paper-based profile which should have the student's academic history, the reality is that these papers are often lost or forgotten at the time of advising, forcing advisors to work with whatever student information is readily available. This can often result in students being led down 'blurred' academic paths as their past is not fully known. While this can possibly be resolved by the advisor simply asking the student to expand on his advising history, the process can be quite time consuming as well as inaccurate as the student might forget to relate critical information that can determine the advice given at the present meeting.

Improper representation of information can also cause unnecessary hiccups in the advising process. Advisors can be forced to work with multiple documents at a time, making the process more tedious than necessary when having to switch between them [2]. Also hand filled forms such as the one used in UWI's advisory process suffer a high possibility of having incorrect information due to human error by student or even past advisors. As a result, the advisor would have to validate form information against the student data within the SIS, which can usurp useful minutes from the advising session. In light of such redundancy, some institutions have chosen to eliminate the use of paper-based forms and simply work off the online student transcript generated from the SIS. The problem then is that such systems lack the ability to provide decision making capabilities based on student data and hence the advisor would still need to thoroughly analyse the student information before making a reasonable conclusion.

Administrative issues are also a factor in determining the quality of advising received by students. Advisors must be well equipped with knowledge on degree requirements, study plans and other rules pertaining to their advising scope to ensure that valid advice be given. In many cases, programmes are often under review and as such advisors are not always up-to-date with the changes made due to lack of dissemination of information from higher administration. Such shortcomings can also result in inaccurate advice being dealt to students causing them dissatisfaction and frustration, since it can possibly lead to delayed graduation [1].

These aforementioned issues, along with the issue of students not being able to attend advising sessions due to geographical constraints have all forced FST administration to explore ways to enhance the process altogether.

D. Academic Advising and Computer Science: Integration and Possible benefits

With the perpetual evolution of computer technology today, it is clear that at some point, institutions would seek to somehow computerize their advising process in an attempt to solve its underlying issues. Academic advising programmes should make full use of all existing modern technology, if necessary, to deliver the advising process.However, technology as a means of offering advisory services can be viewed by some to be cold and impersonal. It is therefore encouraged that technology be harnessed not to fully replace, but to improve the efficiency of the overall advising process, still allowing students to physically meet a human advisor if necessary.

As a result, technology can enhance the academic advising experience by assisting in the making of better informed decisions as well as providing improved services by migrating repetitive tasks on software. This would allow any student-advisor time to be dedicated to helping a student select the most appropriate path or handling any non-academic issues that may have an impact on the student's performance [5]. Such semi-automated advising would also significantly reduce the time for student-advisor interaction since students would only meet human advisors if their needs were not satisfied by the automated advisor. Furthermore, such systems would reduce the workload of staff that had to previously take on the extra job of advising, allowing them to focus on their primary areas of work, and by extension alleviate the issue of having too few advisors within the institution.

Such systems would typically eradicate the need for multiple hard-copy documents as all information could be available via a single interface, making the analysis of student data, if required, possibly easier than switching between physical documents. In addition, automation would help to remove inconsistencies in student information, especially if some sort of student profile is maintained, so that a history of all academic records and advising comments are available to the current advisor. The information would also be 'perpetually' available for the student, who can easily view his information when needed, as opposed to taking notes at advisory meetings or forgetting suggestions that was orally offered to him.

Finally, although providing a remote alternative to students who are unable to physically meet with advisors, an automated advising solution would seek, not to replace the human advisor altogether, but to alleviate his workload and cognitive stress [8]. This would be achieved by handling all student data and making optimized deductions so that the human advisor can focus on what he can do best, which is taking care of qualitative issues that the student may possess; thus improving the quality of academic advising. With such possible improvements it is clear that institutions would opt to transition to some form of an automated system, in an attempt to reap some of the many benefits that technology can provide.

E. Solution: AdviseMe : Student Advising Services

On scrutiny of the issues outlined in FST's academic advising programme, a computerized solution coined 'AdviseMe', an intelligent, web-based application for academic advising was proposed. This came as a result of the desire to raise the quality of the faculty's advising, eliminating inefficiencies where possible as well as migrating from a paper-based system. The solution is designed to complement the current process and hence is expected to work alongside the traditional human advisory system. It serves as an effort to enhance the efficiency, integrity and transparency of any tertiary-based advising system, similar to that of FST.

The system's core features, with the help of its rule-based inference engine, utilizes a student's transcript information and maps it against a set of configurable rules pertaining to programme information and university regulations. Results are then generated , rendering course suggestions for the upcoming semester, as well as information about the student's graduation status and eligibility for qualifications for oral assessments. It also provides students with reference material that provide answers to common issues and questions such as how to request overrides and apply for rescindment of RTW status, among others.

In the event that the advice generated for the student is insufficient, of that he has a special case that requires human attention, the system allows the student to remotely interact with an advisor associated with the student's course of study via email. At this point, it is left to the discretion of the student and human advisor, as to whether the issue can be handled remotely, or if a face to face appointment needs to be made.

AdviseMe also offers human advisor support, allowing them to see via a single interface, all previously mentioned student information as well the ability to place comments on student profiles which can then be used by the student as well as other advisors for future reference. This removes the need for a paper-based form to be maintained by the student and ensures that future advisors have a clear history of all student information before lending necessary advice. Advisors also have access to all reference manuals, eliminating the need for managing multiple documents and thus relieving some of the drudgery associated with the advising process.

An easy to use administration interface is included whereby from a web browser, administrators can manage student and course information, set customized prerequisite and exemption rules for courses, manage university regulations and maintain the overall health of the system, among other features. It also promotes the easy dissemination of rule changes to advisors by sending broadcasts when changes are made to rules within the system. This ensures that advisors are always up-to-date with the latest versions of rules and system settings as they are made in real time. All of these features, backed by an appropriately secure, efficient, scalable system with a simple, user-friendly interface create a compact but effective suite of services designed to enhance the overall advising process.

II. Literature Review

A. Description of Past Solutions

Surveying existing literature, we see that many institutions have implemented computerized solutions in order to enhance their overall advising experience. We also note that most institutions tend to write their own system, not only so that the solution is tailored to suit their direct needs, but also to eliminate the cost of licensing multiple copies of commercial software such as expensive expert system shells [11]. From our research, we see that solutions can be classified based on the level of automation they apply to the overall advising process. We define these systems as Basic Computerized Systems (BCS), Intelligent Interactive Automated Systems (IIAS), Advanced Automated Systems (AAS) and Intelligent Advanced Automated Systems (IAAS).

1) Basic Computerized Systems (BCS): Systems which either facilitate simple remote communication for the advisory process or those which migrate from the paper-based approach, but simply represent data in a computerized form or perform simple calculations are termed Basic Computerized Systems. In such systems, human advisors are still required to analyse information before any advice can be generated.

Reference [12] gives an example of a BCS as their institution uses an online 'Virtual Classroom' where web technologies are used to foster student-advisor communication. Such a system simply facilitates conventional advising, without the need for a fixed geographical location. A more technologically inclined BCS is presented by reference [2]. Created using VBA scripts and Microsoft Excel, the system automates some repetitive tasks in the advising process by performing functions such as GPA calculation. It should be noted however, that system operation requires two excel documents to be provided by the department; the first being a four-year schedule of the study programme and the other being a translation of the student transcript, since the system is not integrated in any way with the SIS. A web-based tool coined "The Online Advisor" however, utilizes existing data within their SIS (inclusive of prerequisite rules and graduation requirements) in order to produce an organized, colour coded representation of all advising-relevant information, centralized in a single display [5]. This is then used to complement the advising process, making it easy for advisors to create academic schedules by semester or year. It was designed to eliminate the use of multiple documents in the advising process by consolidating all information at a single interface.

2) Intelligent Interactive Automated Systems (IIAS): While BCS introduces technology into the advising process and also alleviates some of the drudgery associated with handling paper-based documents and making manual calculations, we see that introduction of higher processing capabilities and expert system technology can significantly reduce the human advisor's responsibility and student load in the overall process by directly handling advisory issues such as course suggestions. Intelligent Interactive Automated Systems seek to use such technology to emulate a real life student-advisor conversation, in order to gather sufficient data to generate substantial advice for students.

"A WWW Delivered Advising System" using the Exsys CORVID Professional Expert System Shell seeks to deliver a 'perpetually available' academic advisor specifically designed to handle cases of students who are unable to physically meet with a human advisor [11]. A similar system was put forward by reference [10] using Java Expert System Shell (JESS) and XML.

The drawback of both systems however is that students are required to supply solutions with a significant amount of information for results to be generated. This can sometimes result in the possibility of students entering inaccurate information and hence getting inaccurate advice, or students becoming disinclined to use such a system as they would prefer a system less demanding of them.

3) Advanced Automated Systems (AAS): In an attempt to make solutions more 'student friendly', we sought to turn attention to systems that produced similar results to IIAS with less student interaction. A subset of these was termed Advanced Automated Systems, which utilized prescribed algorithms and computational power to generate advice based on existing data.

Reference [4] discuss a system which uses database queries on the information stored within each student's transcript in the SIS in order to give students advice as to what courses they should take in the next semester as well as give their graduation status. Another system surrounding PHP, MySQL and Email technology is presented by reference [7] which again uses database queries to group all related student and course information for the purposes of generating a list of suggested courses to be taken in the next semester. A similar system is also shown in reference [13] whereby developers used Wxpython alongside an access database to deploy a desktop solution to facilitate postgraduate students.

4) Intelligent Advanced Automated Systems (IAAS): AAS can be quite effective and produce satisfactory results. However, the database queries used to generate such results can be quite complex and resource intensive. Furthermore, the rules within advising solutions that govern programme structure and university regulations are frequently reviewed, which can then require the need to modify the SQL queries which act as rules governing the system's functionality. A more practical approach, promoting change and easy maintenance would be using expert system technology to manage the rules that govern the advising system. Such systems that provide this functionality, similar to that of AAS are coined as Intelligent Advanced Automated Systems and research showed that this was the most favoured approach when creating advising solutions for institutions.

Such systems were observed to use various reasoning strategies to achieve their appropriate results. One for example showed how Case-Based Reasoning (CBR) was used to develop a system that recommended a suitable major to students based on comparing their student information against similar historical cases [9]. The system was proven to be quite effective when advising students who were reading for general degrees, provided there were sufficient historical cases within its knowledge base (KB). Rule-Based Reasoning (RBR) systems were also designed whereby developers used Forward and Backward Chaining procedures in order to generate appropriate advice. Such capabilities as well as a 'cognitive and emotional filter' were used in a solution proposed by reference [1] in order to provide course advising to students. "IS-Advisor" also followed a rule-based approach alongside its Object Oriented Database [6]. Other systems such as the "Course Advisory Expert System" went a step further by providing both reasoning capabilities in order to facilitate an even higher quality of advising when prescribing course recommendations [8].

With the emergence of ontologies and semantic web technology, we saw that developers harnessed ontology driven methodologies to tackle the dynamic and complex nature of student academic planning and scheduling by creating "E-Advisor", a multi-agent intelligent advising system [3]. Designed for the Master of Science in Information Systems in Athabasca University in Canada, it allowed students the ability to add preferences of specialization to their profile and then recommend courses based on these preferences. Its multi-agent nature also made system maintenance easy, as it allowed the use of other agents while updating others making it a highly available system, and possibly one of the better intelligent web-based advising systems in the world today.

B. AdviseMe as opposed to previously implemented solutions

After much scrutiny of the problem domain as well as the solutions discussed in section II(A), it was clear that a system with an IAAS architecture be the optimal solution for enhancing the level of advising currently experienced in FST, as it would seek not only to minimize the need for student input and generate satisfactory results, but also significantly reduce human advisor workload, among other benefits. This places "E-Advisor" as the top candidate for consideration. However, several issues arose which lent to the ultimate dismissal of such a solution.

While E-Advisor has proved its successful application at Athabasca University, it is noteworthy that the system was tailored to advise students in a single MSc Programme [3] and not a range of programmes that would be required of a system serving FST. Also, the system focuses on course scheduling, but no mention of the system handling other important features required by FST was made in the literature. Determining oral assessment qualifications and graduation status are but only some of the important student issues that human advisors often have to repeatedly solve.

As a result, AdviseMe's design was proposed since not only does it facilitate aforementioned services as well as tracking of student history and other useful features, but also facilitates both undergraduate and postgraduate programmes that can be managed on a department, faculty or even campus based level. Furthermore, AdviseMe's solution requires less staff involvement as it only needs a minimum of 1 administrator to configure the environment while E-Advisor's functionality is dependent on every instructor in the department for the proper working of the system. While this extended staff involvement poses benefits in the context of Athabasca University, it fails to alleviate the staffing issues faced in FST. For these reasons, as well as the fact that no single previously mentioned solution efficiently solves the issues outlined in section I(C), the design and implementation of AdviseMe was born.

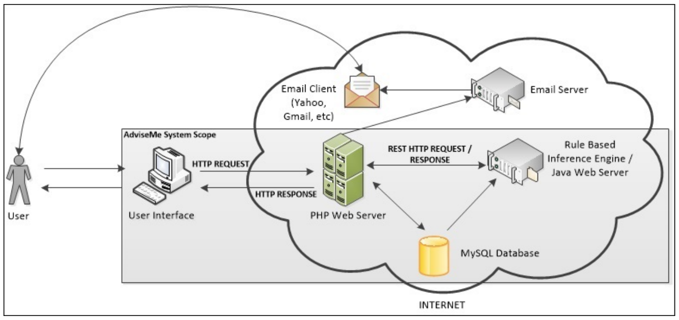


Fig. 2: Flow of data between modules within AdviseMe

III. System Design

A. Overview

AdviseMe typically has 3 main user groups (students, advisors and administrators) all of which interact with the system via a single browser rendered user interface.

When requests are made, they are passed to the PHP Web Server (PWS) which acts as the mediator of all data flow within the system. The system provides both 'intelligent' services (eg. Course Advising) as well as 'non-intelligent' services such as system management and email communication. For all intelligent services, the PWS communicates with the RESTful Java Web Server (JWS), which handles all intelligent processing via its use of ontologies and its rule based inference engine.

The PWS also sends requests to the Email Server for disseminating messages to users based on the services accessed. For example, when an administrator updates a rule in the system, the PWS would send a request to the Email Server in order to broadcast a message to advisors, notifying them of changes made.

All other facilities provided by the PWS are fuelled by information retrieved from or sent to the MySQL database via the UI. The JWS also retrieves information from the MySQL database before processing data and sending results to the PWS. Ultimately, the PWS would manage all processing within the system and render the output in a sleek, intuitive form via the UI.

B. PHP Web Server (PWS)

The PWS was designed and implemented using the CodeIgniter Framework and facilitates the server side implementation of the web application, as well as the interface for communication with the Email Server (Google's Gmail SMTP Server), MySQL database and RESTful JWS (by means of the Curl URL Library). CodeIgniter's extensive documentation and 'Model-View-Controller' architectural style promoted quick and easy implementation of the PHP based application server. Controllers accept data from the models and pass them to views which render the information in an intuitive format for end users. In AdviseMe, the models handle two types of data requests, the first being calls to the MySQL database using CodeIgniter's Database Library and the other being requests being sent to the RESTful JWS. Information from the JWS is retrieved via Uniform Resource Locators (URLs) and the use of the Curl URL Library, for the consumption of REST services. An overview of the PWS architecture is shown in Figure 3.

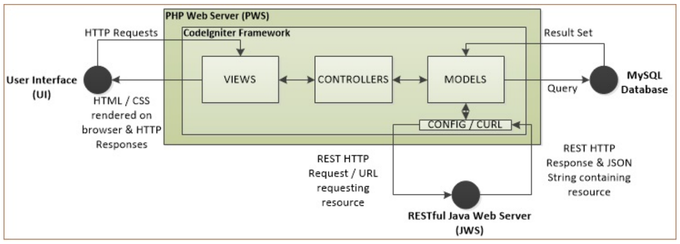


Fig. 3: Overview of the PHP Web Server (PWS) Architectural Design

C. RESTful Java Web Server (JWS) Design

Although not directly accessed by users, the JWS is essentially the driving force behind AdviseMe's core functionality. Using Apache Jena alongside ConnectorJ to create an ontology-based reasoning environment, it proves to be an ideal reasoning engine when seeking to determine non-trivial inferences on a student's academic record. These inferences are then exposed as JSON objects via RESTful web services, made possible by using tools such as Jersey and JSON-Simple.

Jena's API provides a wide collection of classes and interfaces for the management of 'OWL-Based' technologies. OWL, an acronym for 'Web Ontology Language', is typically an extension of the Resource Definition Framework Schema (RDFS) which in turn is an extension of the Resource Definition Framework (RDF). RDF forms the foundation of how resource information should be structured, with RDFS and OWL enhancing the ways in which resources are described.

With RDF being a suitable standard for structuring data and OWL having an extensive vocabulary that can be easily interpreted by machines, Jena's Ontology API was used to create a simple ontology termed 'AdviseMeOnt' to be used within our system context. Comprising of a set of resources / classes immediately surrounding the academic advising environment and their appropriate properties, it was used to model all information within the student's transcript in order to produce meaningful inferences to support the advising process. A listing of some of the concepts within 'AdviseMeOnt' is given in Table 1.

Student Profiles were then created by extracting information from the MySQL database to create and populate an ontology model using the Jena Ontology API, and then exporting the information to a RDF file.

In addition to creating RDF-Based student information, the JWS was also designed to extract information from the database and create a set of user defined rules to be applied to the RDF files. These custom rules included prerequisite and exemption rules, as well as rules containing other user defined variables. The JWS also can generate a set of fixed, system-defined rules by calling the appropriate functions within the server.

Once RDF-Based student profiles and the set of executable rules are existent, the JWS can then generate non-trivial

TABLE I: Listing of some concepts and properties within ‘AdviseMeOnt’

advMe:COURSE
advMe:COURSECODE
advMe:COURSENAME
advMe:SEMESTEROFFERED
advMe:MARKOBTAINED
advMe:NUMBERCREDITS
advMe:PROGRAMME
advMe:PROGRAMMENAME
advMe:LEVELONECREDITS
advMe:CORECREDITS
advMe:ELECTIVECREDITS
advMe:FOUNDATIONCREDITS
advMe:HASCORECOURSE
advMe:HASELECTIVECOURSE
advMe:HASFOUNDATIONCOURSE etc...
advMe:STUDENT
advMe:STUDENTID
advMe:STUDENTNAME
advMe:STUDENTTYPE
advMe:ACADEMICSTANDING
advMe:FUNDINGSTATUS
advMe:CANGRADUATE
advMe:COMPLETEDCOURSE
advMe:ALLOWEDCOURSE
advMe:LOCKEDCOURSE
advMe:FAILEDCOURSE
advMe:POSSIBLEORALCOURSE etc...

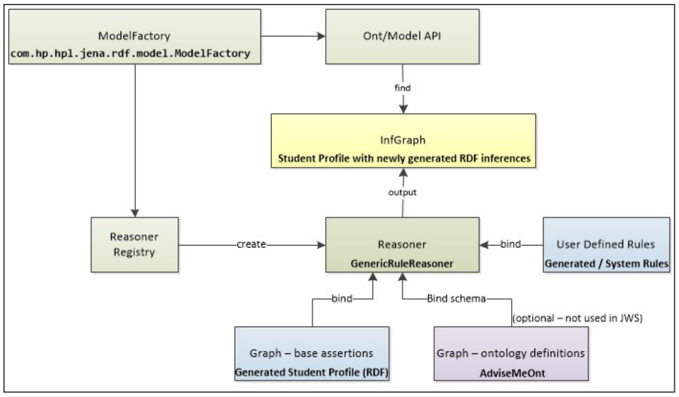


Fig. 4: Overall Structure of Inference Machinery used within Jena

inferences within the student profile using the Jena Inference Subsystem (JIS). This module, accessed via the Jena API, is designed to allow a wide range of inference engines to be integrated with Jena projects for the derivation of additional RDF data, by making inferences on existing RDF assertions. Figure 4 illustrates the overall structure of the inference machinery used within Jena and by extension the JWS.

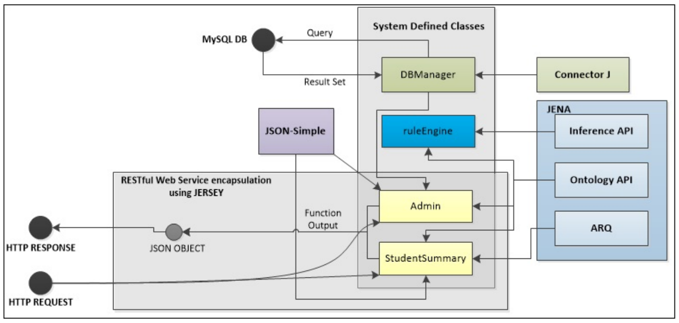


Fig. 5: Overview of the Java Web Server (JWS)Architectural Design

The JWS accesses the JIS using the ModelFactory class to associate the RDF-Based student profiles with the appropriate reasoner. Jena provides a Generic Rule Reasoner which supports the user of user-defined rules as well as forward and backward chaining capabilities, ideal for handling the ruleset within the JWS. The student profile and ruleset are bound to this generic reasoner and after processing, the reasoner outputs an InfGraph object. This is typically a new model of the student profile containing previously asserted as well as newly inferred information. The information is then saved to storage as an RDF file which is later parsed using the Jena Ontology API as well as ARQ (a query engine supported by Jena providing RDF querying capabilities) for extracting results.

In order to expose inference results to users in a presentable format such as via a structured website, information extracted from the InfGraph model was transformed into JSON format before being made available to external users via RESTful web services. This was done using JSON-Simple and Jersey Java packages respectively which allowed information to be passed to the PWS via an HTTP response.

As a result, the JWS communicates seamlessly with the PWS in order to make useful advising information for rendering via the UI. Administrators are also allowed to access the PWS via the UI in order to perform ‘remote’ tasks on the JWS such as re-creating a student’s RDF-Based Profile and refreshing the user-defined rule base. Figure 5 shows an overview of architecture of the JWS.

IV. Methodology

A. Pre-Implementation Phase

Preceding system implementation, research was conducted by acquiring information from both students and advisors currently at UWI. Student’s thoughts on a computerized advising system were collected via an anonymous questionnaire. Participants were selected from first year undergraduates straight up to postgraduate students in order to get a wide range of responses from those who would have been new, as well as quite accustomed to the traditional advisory process. Results from the questionnaire showed that the majority of students were dissatisfied with the traditional advising process, with 90% of them visiting advisors for the handling of general matters such as course scheduling, determination of graduation status and the handling of trivial student issues. Furthermore, students stated that they would expect such a system to offer 24/7 accessibility, course advising capabilities and support for handling other student issues such as holds and academic standing issues. They also suggested that the system possess some level of intelligence so to give possibly better advice than existing advisors who are sometimes unaware of critical changes in the process.

Representing the advisory body was the Deputy Dean and Head of the Advisory unit at FST. In conclusion of her interview she acknowledged that an intelligent computerized system would have immensely benefited advisors by reducing their advisory load, promoting more dedication to special cases and enhancing the overall workflow of the advisory process. She verified the hypothesis that such a system would be able to handle a significant amount of student cases without the need for a human advisor and also highlighted that the ability of students to remotely contact human advisors if necessary could also help to streamline the other aspects of the advisory process. As a result of her confidence in the system, the design and implementation quickly came under way as it posed some value to UWI and more specifically FST.

B. Implementation Phase

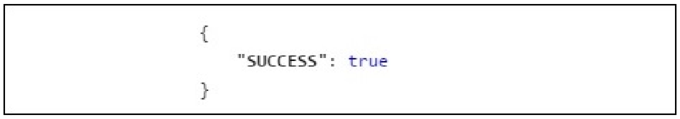
As stated in section III, AdviseMe comprises of two distinct modules, the RESTful JWS which provides all intelligent processing with respect to generating academic advice and the PWS which converts this advice into a presentable format to be rendered via a web browser, among other functions. Section IV(B1) gives a summary of some of the services offered by the JWS while Section IV(B2) summarizes the functionality of the overall system as controlled by the PWS.

1) RESTful JWS Implementation: As stated before, the JWS provides RESTful services accessible via HTTP requests. These services comprise of the following:

- Creating / Refreshing the ‘AdviseMeOnt’ Ontology
- Creating / Refreshing the rule base (comprising of user and system defined rules)
- Creating / Refreshing all student profiles (based on transcript information in SQL database)
- Creating / Refreshing a single student profile(based on transcript information in SQL database)
- Generating advisory information for a given student (based on his student profile)

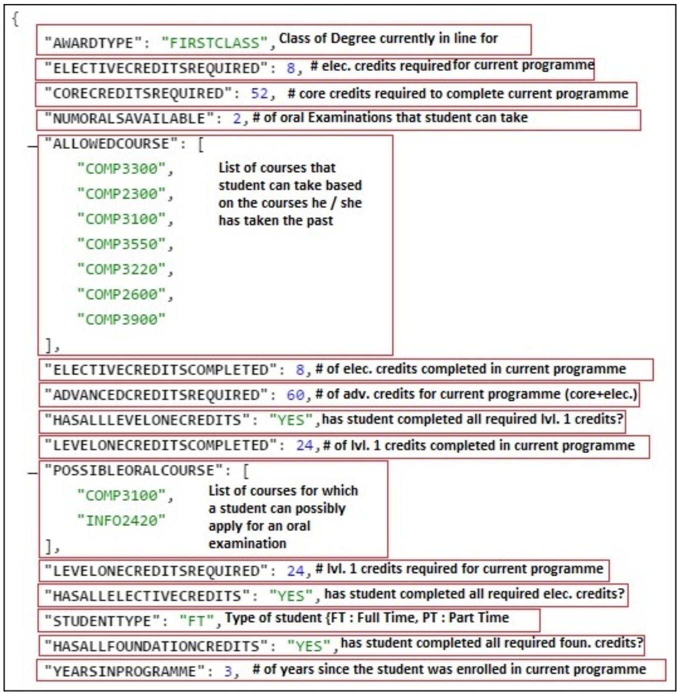
All of these services return a JSON string to the requesting agent after processing. With exception to the last service, all requests return a success value of “TRUE” if the request was successfully processed or “FALSE” otherwise (see Figure 6). In the case of generating the advisory information however, the request,once successful, returns a JSON object containing the following information:

- Student Name, ID and Student Type (Full Time / Part Time)
- Academic Standing and Funding Status (Whether or not the student can be covered for sponsorship payments)
- Number of years in current programme of study
- Number of credits required for each course segment of programme (Eg. Core, Elective, etc.)



```
{
    "SUCCESS": true
}
```

Fig. 6: Sample of service output when genaral request to JWS is made



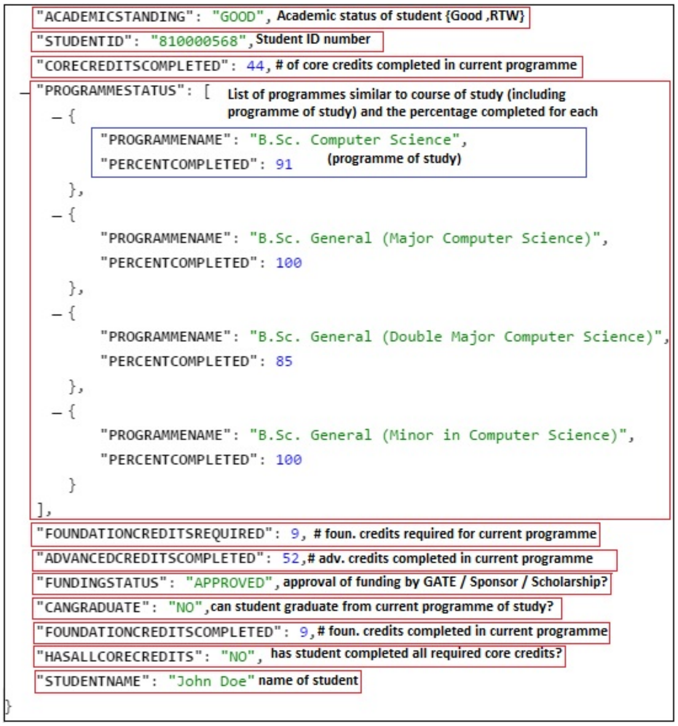
```
{
  "AWARDTYPE": "FIRSTCLASS", Class of Degree currently in line for
  "ELECTIVECREDITSREQUIRED": 8, # elec. credits required for current programme
  "CORECREDITSREQUIRED": 52,  # core credits required to complete current programme
  "NUMORALSAVAILABLE": 2, # of oral Examinations that student can take
- "ALLOWEDCOURSE": [
      "COMP3300",        List of courses that
      "COMP2300",        student can take based
      "COMP3100",        on the courses he / she
      "COMP3550",        has taken the past
      "COMP3220",
      "COMP2600",
      "COMP3900"
  ],
  "ELECTIVECREDITSCOMPLETED": 8, # of elec. credits completed in current programme
  "ADVANCEDCREDITSREQUIRED": 60, # of adv. credits for current programme (core+elec.)
  "HASALLLEVELONECREDITS": "YES",has student completed all required lvl. 1 credits?
  "LEVELONECREDITSCOMPLETED": 24, # of lvl. 1 credits completed in current programme
- "POSSIBLEORALCOURSE": [
      "COMP3100",       List of courses for which
      "INFO2420"        a student can possibly
                        apply for an oral
  ],                    examination
  "LEVELONECREDITSREQUIRED": 24, # lvl. 1 credits required for current programme
  "HASALLELECTIVECREDITS": "YES", has student completed all required elec. credits?
  "STUDENTTYPE": "FT", Type of student {FT : Full Time, PT : Part Time
  "HASALLFOUNDATIONCREDITS": "YES", has student completed all required foun. credits?
  "YEARSINPROGRAMME": 3,  # of years since the student was enrolled in current programme
```

Fig. 7: Sample of service output when request to retrieve student’s advising information is made (Part A)

- Number of credits completed for each course segment of programme (Eg. Core, Elective, etc.)
- List of courses that can be taken by student (based on priority in programme of study)
- Whether or not a student can graduate from current programme of study
- Number of Oral Assessments that the student currently has remaining
- List of all courses which the student can apply for an oral assessment
- Progress in all programmes similar to student’s current programme of study

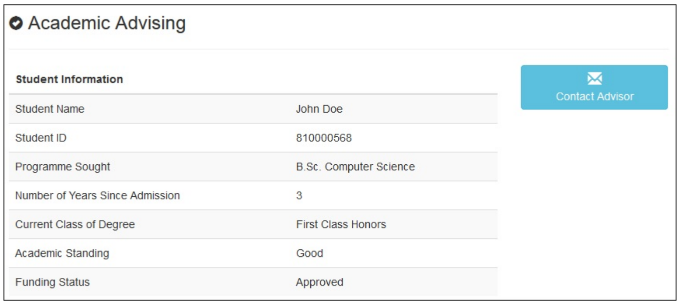
An illustration of this is shown in figures 7 and 8. This information is then consumed by the PWS which then renders it to the user as required.

2) PWS Implementation : The PWS, responsible for controlling all functionality within AdviseMe, provides a wide range of functions and serves three types of users: students, advisors and administrators. A list of the core of AdviseMe is given in the table 2.



```
"ACADEMICSTANDING": "GOOD", Academic status of student {Good ,RTW}
"STUDENTID": "810000568", Student ID number
"CORECREDITSCOMPLETED": 44, # of core credits completed in current programme
"PROGRAMMESTATUS": [   List of programmes similar to course of study (including
                       programme of study) and the percentage completed for each
  {
      "PROGRAMMENAME": "B.Sc. Computer Science",
      "PERCENTCOMPLETED": 91      (programme of study)
  },
  {
      "PROGRAMMENAME": "B.Sc. General (Major Computer Science)",
      "PERCENTCOMPLETED": 100
  },
  {
      "PROGRAMMENAME": "B.Sc. General (Double Major Computer Science)",
      "PERCENTCOMPLETED": 85
  },
  {
      "PROGRAMMENAME": "B.Sc. General (Minor in Computer Science)",
      "PERCENTCOMPLETED": 100
  }
],
"FOUNDATIONCREDITSREQUIRED": 9, # foun. credits required for current programme
"ADVANCEDCREDITSCOMPLETED": 52,# adv. credits completed in current programme
"FUNDINGSTATUS": "APPROVED", approval of funding by GATE / Sponsor / Scholarship?
"CANGRADUATE": "NO",can student graduate from current programme of study?
"FOUNDATIONCREDITSCOMPLETED": 9,# foun. credits completed in current programme
"HASALLCORECREDITS": "NO", has student completed all required core credits?
"STUDENTNAME": "John Doe" name of student
}
```

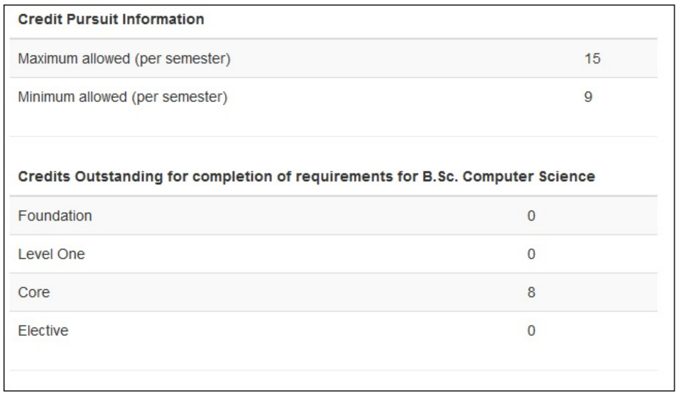
Fig. 8: Sample of service output when request to retrieve student's advising information is made (Part B)



Academic Advising

Student Information	
Student Name	John Doe
Student ID	810000568
Programme Sought	B.Sc. Computer Science
Number of Years Since Admission	3
Current Class of Degree	First Class Honors
Academic Standing	Good
Funding Status	Approved

Fig. 9: Sample of Academic Advising Page (Part 1)



Credit Pursuit Information	
Maximum allowed (per semester)	15
Minimum allowed (per semester)	9

Credits Outstanding for completion of requirements for B.Sc. Computer Science	
Foundation	0
Level One	0
Core	8
Elective	0

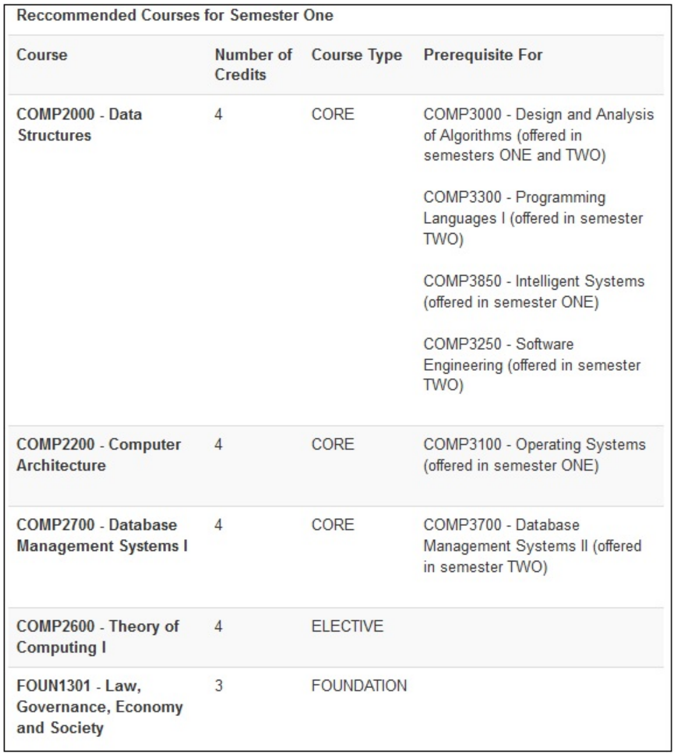
Fig. 10: Sample of Academic Advising Page (Part 2)

TABLE II: Core Features of AdviseMe

Description	Stu	Adv	Adm
View student advising history giving all comments previously placed on student profile by advisors.	YES	-	-
View course advising, by generating a list of courses to be taken per semester as well as show progress in similar programmes in the event of a transfer being considered.	YES	-	-
Request to contact a human advisor in the event that advice given is not sufficient.	YES	-	-
View graduation status of student, showing progress in each of the programme sections.	YES	-	-
View oral examination information, giving the number of orals that a student can still pursue, the number of outstanding credits for completion of programme requirements and the list of courses for which a student can request an oral examination.	YES	-	-
Download a complete summary of all aforementioned information for a student in PDF format.	YES	YES	-
View Information on programme structure as well as common student issues such as requesting overrides and rescindment of RTW status.	YES	YES	-
Place a comment on a students record to be used for future reference by student and advisors.	-	YES	-
Management of all rules within the system inclusive of creation and modification of user defined rules such as prerequisite and exemption rules.	-	-	YES
Management of all courses, programmes, departments, faculties, students' transcripts and users within the context of the university and system.	-	-	YES
Management of the entities associated with the JWS (as mentioned in the previous section).	-	-	YES

As a student, possibly the most useful feature is the "Academic Advising" option, which allows him / her the ability to retrieve information generally sought in academic advising sessions. This initially includes basic student information, his / her academic standing and the class of degree he / she is currently in line for (Figure 9). The proceeding section then gives the maximum and minimum number of credits the student can take per semester as well as the number of credits remaining for each section of the programme (Figure 10). This is then followed by a list of courses (Figure 11) within the current programme of study that can be taken for programme advancement based on semester and also ordered by two levels of priority; the first being how much courses they are prerequisites for and the second being classification by course type (eg. core courses are given higher priority than elective courses). The student is also notified in what semester's possible future courses will be offered so to promote future course planning. Finally, in the last section of the page, the student is able to see his / her progress in similar programmes offered (Figure 12), allowing them to know their stance in other programmes in the event they decide to switch to another programme.

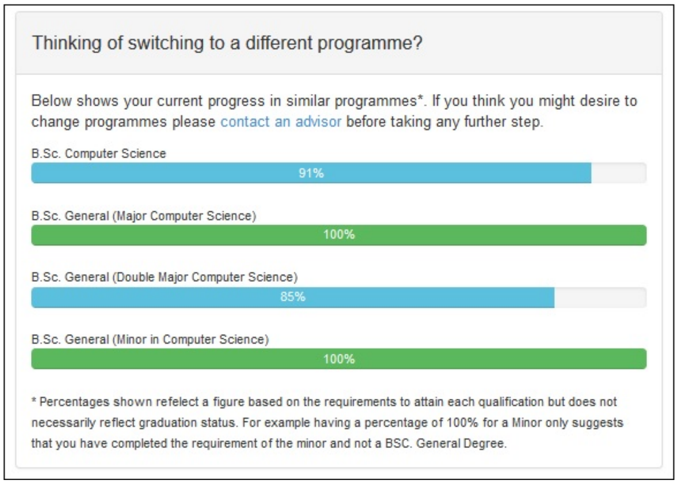
All of these aforementioned sections, in addition to information of graduation status and oral exam assessments give satisfactory advice that answer most common questions faced in academic advising. If however, the student believes that the advice received is insufficient, he / she can then opt to contact a human advisor via email by clicking the "contact advisor" button shown in Figure 9. This then redirects to a new page which allows the student to send an email to the advisor via adviseMe's interface.



Reccommended Courses for Semester One

Course	Number of Credits	Course Type	Prerequisite For
COMP2000 - Data Structures	4	CORE	COMP3000 - Design and Analysis of Algorithms (offered in semesters ONE and TWO) COMP3300 - Programming Languages I (offered in semester TWO) COMP3850 - Intelligent Systems (offered in semester ONE) COMP3250 - Software Engineering (offered in semester TWO)
COMP2200 - Computer Architecture	4	CORE	COMP3100 - Operating Systems (offered in semester ONE)
COMP2700 - Database Management Systems I	4	CORE	COMP3700 - Database Management Systems II (offered in semester TWO)
COMP2600 - Theory of Computing I	4	ELECTIVE	
FOUN1301 - Law, Governance, Economy and Society	3	FOUNDATION	

Fig. 11: Sample of Academic Advising Page (Part 3)



Thinking of switching to a different programme?

Below shows your current progress in similar programmes*. If you think you might desire to change programmes please contact an advisor before taking any further step.

B.Sc. Computer Science
91%

B.Sc. General (Major Computer Science)
100%

B.Sc. General (Double Major Computer Science)
85%

B.Sc. General (Minor in Computer Science)
100%

* Percentages shown refelect a figure based on the requirements to attain each qualification but does not necessarily reflect graduation status. For example having a percentage of 100% for a Minor only suggests that you have completed the requirement of the minor and not a BSC. General Degree.

Fig. 12: Sample of Academic Advising Page (Part 4)

C. Post Implementation Phase

To ensure overall acceptance of the system, students of varying aptitudes within the university were allowed to assess the system by either volunteering their own transcript or using a model of a sample transcript. From the 50 percent of students who produced their transcripts for testing, all received accurate advice in all aspects of the system with regard to student advising. In addition to its functional success, all interviewed students found the system to be exceptionally usable and intuitive, commenting that its simplicity and use of visual charts to illustrate student progress made the system very appealing. They also in particular, appreciated the record of advisory history appended to student profiles and the ability to contact a human advisor if necessary. 20 percent of the sample set however, still held strong to the fact that while they would receive sufficient information from the system, they would still opt to speak to an advisor if required. They did appreciate however that the process of meeting an advisor could now be less time consuming as they can now schedule and appointment beforehand, eliminating the need to wait long periods in the advisory offices.

V. Limitations of the System

The system is designed to give accurate advice only to those pursuing “special degrees” or programmes which follow a clear cut path of courses. While this limits the system to serve only a portion of students within FST and other similar institutions, it would still result in the overall reduction of advising load for human advisors, making their duties less demanding.

Another limitation is that while the student can contact a human advisor via email and successfully have their issues solved, the advisor’s response time is subjective to when he / she chooses to respond to the student as opposed to face to face conversations where solutions are immediately discussed. This can be unsatisfactory to students, especially if advisors take too long to respond to student concerns.

The system is also in the prototype phase and hence requires administrators to input the student transcripts for system processing via the user interface. While UI design makes the uploading process possible in minimal time, it still results in an unnecessary action by a human entity. However, this timeframe is relatively short, which can be deemed quite acceptable as the number of hours saved by both students and advisors using the system would clearly offset the setup period.

VI. Future Work

With respect to the limitations discussed above, further enhancements are proposed to investigate ways in which the system would be able to sufficiently accommodate students pursuing general degrees. Also in terms of data integration, measures to have the system automatically process information from the Banner Student Information System will be explored, in an attempt to reduce administration involvement further.

With respect to added system functionality, measures would be put in place to offer a higher level of student advisor communication, possibly including live chats with advisors via Instant Messaging or Video Conferencing. Also the idea of a blog whereby a community handling frequently asked questions or common issues would be explored. This would possibly increase the level of acceptance of the system by students and solve the issue of delayed responses that can be incurred while waiting for advisors to attend to emails.

Finally, with respect to future long term enhancements of AdviseMe, a programme planning module can be implemented, using collected data from student advising to generate statistics for use by departmental administration to assist in the allocation of teaching resources for upcoming semesters. This would not only add to the value of the system, but increase its scope of alleviating manual processing issues within tertiary education institutions.

VII. Conclusion

Utilizing AdviseMe in order to facilitate academic advising without the possible involvement of human advisors will definitely enhance the efficiency, integrity and transparency of any tertiary based advising process similar to that of FST. The successful prototype discussed in this paper highlights its feasibility and practicality in the context of UWI and their degree programmes. Serving a significant portion of the student body, it would provide sufficient advisory services to the majority of its users thus reducing the student load faced by human advisors at advisory offices. For those students whose issues go beyond the scope of its assistance, it allows an avenue of communication to human advisors that was previously non-existent, by means of email technology. Not only does this create a flexible way of seeking advice, but it can also improve the quality of the qualitative advice received from advisors, since more time can now be dedicated to handling these special cases.

Its current architecture supports the use of a PHP based web application interacting with an intelligent, RESTful Java Web Server, in order to provide expert advice on course scheduling issues via any device that supports internet browsing. However, the fact that the intelligence processing is exposed via web services indicates that the system can be adapted to suit any future front end deployment, provided that access to the JWS API is given. This promotes acceptance of AdviseMe, outside of UWI, as solutions can be tailored to any university context, once their programme structure is the similar to that of UWI. Returning to the context of FST, we see that a significant number of students showed possible acceptance of the system in the infant stages of conceptualization. This figure sought to increase nearing the end of implementation however as all students interviewed for testing made positive remarks about the system's functionality, usability and applicability to the advising context of FST; with the majority of them stating that they would use such a system. Furthermore, when demonstrated to the Head of the Advisory Unit of FST, she was highly pleased with the outcome, to the point of suggesting possible practical implementation within the faculty in the near future.

With such positive feedback, and the fact that the resulting product captured all the requirements that was proposed of such a system, it is clear that the implementation of AdviseMe was indeed a success, lending its services to support new and existing advising processes in order to enhance the overall quality of academic advising received by students today.

References

[1] E. Nwelih and S. Chiemeke, "Framework for a web-based spatial decision supportsystem for academic advising," *African Journal of Computing and ICT*, vol. 5, no. 4, pp. 121–126, 2012.

[2] M. T. Al-Nory, "Simple decision support tool for university academic advising," in *Information Technology in Medicine and Education (ITME), 2012 International Symposium on*, vol. 1. IEEE, 2012, pp. 53–57.

[3] F. Lin, S. Leung, D. Wen, F. Zhang, and M. Kinshuk, "e-advisor: A multi-agent system for academic advising," *International Transactions on Systems Science and Applications*, vol. 4, no. 2, pp. 89–98, 2008.

[4] F. Albalooshi and S. Shatnawi, "Online academic advising support," in *Technological Developments in Networking, Education and Automation*. Springer, 2010, pp. 25–29.

[5] T. Feghali, I. Zbib, and S. Hallal, "A web-based decision support tool for academic advising," *Journal of Educational Technology & Society*, vol. 14, no. 1, pp. 82–94, 2011.

[6] M. A. Al Ahmar, "A prototype student advising expert system supported with an object-oriented database," *International Journal of Advanced Computer Science and Application*, vol. 1, no. 3, pp. 100–105, 2011.

[7] M. Beheshti, T. Trang, K. Kowalski, and J. Han, "Student advising system," in *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, vol. 2006, no. 1, 2006, pp. 2727–2732.

[8] O. Daramola, O. Emebo, I. Afolabi, and C. Ayo, "Implementation of an intelligent course advisory expert system," *International Journal of Advanced Research in Artificial Intelligence*, vol. 3, no. 5, pp. 6–12, 2014.

[9] L. Mostafa, G. Oately, N. Khalifa, and W. Rabie, "A case based reasoning system for academic advising in egyptian educational institutions."

[10] A. N. Nambiar and A. K. Dutta, "Expert system for student advising using jess," in *Educational and Information Technology (ICEIT), 2010 International Conference on*, vol. 1. IEEE, 2010, pp. V1–312–V1–315.

[11] K. Kowalski, J. Goetz, and M. Alam, "Intelligent on-line advising with expert system shell," in *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, vol. 2006, no. 1, 2006, pp. 687–694.

[12] M. J. Martínez-Argüelles, E. Ruiz-Dotras, and E. Rimbau-Gilabert, "The academic advising system in a virtual university," in *Technology Enhanced Learning. Quality of Teaching and Educational Reform*. Springer, 2010, pp. 345–350.

[13] A. Al-Ghamdi, S. Al-Ghuribi, A. Fadel, and F. AL-Ruhaili, "An expert system for advising postgraduate students," *International Journal of Computer Science and Information Technologies*, vol. 3, no. 3, 2012.

Evolutionary Approach to jointly resolve the Power and the Capacity Optimization problems in the multi-user OFDMA Systems

Ndiaye Abdourahmane, Ouya Samuel and Mendy Gervais
Ecole Supérieure Polytechnique
Université Cheikh Anta DIOP
Dakar, Sénégal
BP:5085, Dakar-Fann, Sénégal

***Abstract*—This paper deals with the problem of resources allocation in the downlink of the radio mobile systems. The allocation of resources is established in the context of multi-path effect and Doppler effect. These phenomena cause randomly variations of the channel and make difficult the allocation of the resources. Therefore, the problem of resources allocation is a high non linear optimization problem. Thus, we propose an implementation of genetic algorithms approach to increase the total throughput of users while minimizing the consumption of the total power of the base station. Like the evolutionary algorithms approach, this method is characterized by its robustness, permitting to efficiently solve a non linear optimization problem.**

***Keywords*—Cellular Network; OFDMA; Resources Allocation; Rate Adaptive; Margin Adaptive; Genetic Algorithms**

I. Introduction

Today, the requirements of the users's data rate are becoming more and more increasing; which led to the adoption of efficient resources allocation methods in order to improve the QoS in mobile radio communications. These resources management methods form the basis of the fourth generation of mobile networks systems.

The 4th generation of mobile systems such as LTE-Advanced, will offer high data rate services and large capacity of users. The major change of these systems is the access method OFDMA (Orthogonal Frequency Division Multiplexing Access) which significantly increases the capacity by the improvement of the signal-to-noise ratio obtained with an adaptive resources allocation [13].

The access method OFDMA is base on the transmission technique OFDM(Orthogonal Frequency Division Multiplexing) which consist of the frequency multiplexing of modulated orthogonal sub-carriers. The method OFDM is very suitable for the variations of the channel caused in particular by the multi-path phenomenon and the Doppler effect.

The principle of the modulation OFDM is to split the available bandwidth into many flat sub-channels which can be attributed independently to users [13]. In addition, the quantity of bits transmitted on each sub-channel can be adapted according to the channel gain level in order to obtain a considerable power gain with the required bit error rate [2].

The independence of the OFDM sub-carriers permits an efficient resources allocation. In the context of channel characterized frequency selectivity and time variation, the efficient allocation of resources consists in affecting the sub-carriers to users taking into account the specific channel gain for each users. In preference, the sub-carriers are allocated to users with highest channel gain while satisfying the required data rate for each user.

In this paper, we propose an implementation of genetic algorithms for solving the rate-adaptive and the margin-adaptive problems, subject to the available total power in the base station and the minimal required users's data rate. We consider the downlink of the OFDMA based-systems and we take into account the propagation path-loss so as to guarantee effective throughput to users in reception.

In Section 2, the related work about resources allocation is described. In Section 3, the modeling system and the formulation of the optimization problem are presented. In Section 4, the principle of genetic algorithms is given and in Section 5, the proposed algorithm is described. Finally, in Section 6, we compare the results of the proposed algorithm to others algorithms.

II. Related work

The algorithmic approach for the resources allocation problem is introduced by Ibaraki et al. [7]. Many algorithms are proposed to solve the problem of resources allocation. There are the iterative methods and the heuristic methods. Among the iterative approaches, there are the water-filling algorithm which solves the problem of power allocation in single user case. The work proposed by Tu et al. [16] is based on the water-filling approach. The resolution of the resources allocation problem in multi-users case is proposed in [3][5]. About the heuristic methods, Zhou et al. [17], Ahmadi et al. [1] and Reddy et al. [12] proposed the genetic algorithms approach for the resources allocation problem.

In several works cited above, some aspects are not taken into account for instance the path-loss which is a very determinant parameter in order to have the realist data rate for users at the reception. In addition, the model of channel gain chosen in this work, takes simultaneously into account the multi-path and the Doppler effects. Thus, in this paper we propose an implementation of genetic algorithms for solving the rate-adaptive problem, subject to the available total power in the base station. In this work, we consider the downlink of the OFDMA based-systems and we take into account the propagation path-loss so as to guarantee effective throughput to users in reception.

III. System modeling and Formulation of the problem

A. System modeling

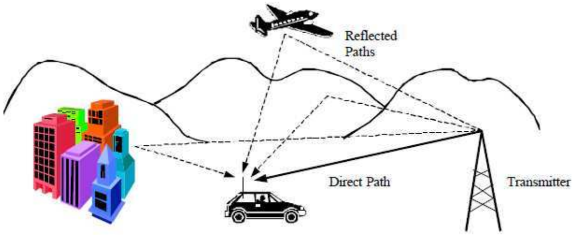


Fig. 1. Multi-path channel with Doppler effect

In this work, we study the allocation of resources in the downlink of mobile network based on the OFDMA access method (Orthogonal Frequency Division Multiplexing Access). We consider a mono-cellular system with multiple access and we suppose that the users are uniformly distributed in the cell. The channel of transmission is assumed to be characterized by the multi-path and the Doppler effects (see figure 1). These phenomena lead respectively to the frequency selectivity and the time-variant of the channel [14]. The time-variant of the radio channel propagation is caused by the movement of the transmitter and/or the receiver. Thus, the channel is non-stationary, and this gain varies in the frequency and in the time domain according to each user.

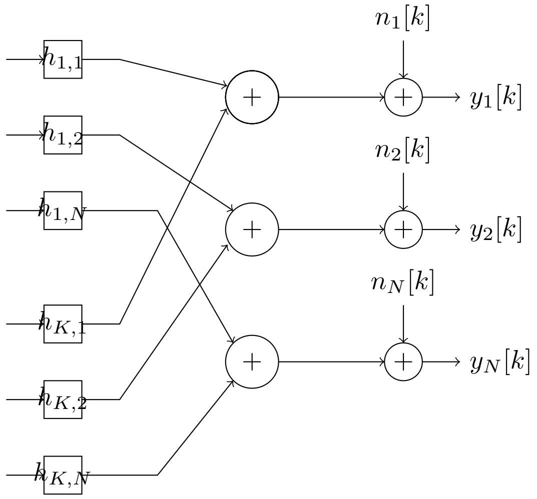


Figure 2: Multi-user OFDM systems

The base station (transmitter) is considered to be fixed and the users are assumed to be in mobility. Therefore, the Doppler effect depends on the speed of users and the frequency used in the transmission. Let's consider v and f_0 respectively the speed of the users and the frequency of the transmitted signal. The maximum Doppler shift is given by:

$$fd_{max} = \frac{v}{c} f_0 \tag{1}$$

where c is the celerity of the light.

For an angle α between the direction of transmission and the movement of user, the Doppler frequency shift is defined by:

$$fd = fd_{max} cos(\alpha) \tag{2}$$

In the case of Doppler effect, the received signal is characterized by spatial interference pattern which causes, when the user is moving in this pattern, many fluctuations of the amplitude of the signal [14]. The time-variant of the channel is characterized by the time of correlation $\tau_{cor} = 1/fd_{max}$. The frequency selectivity corresponds to a variation of channel gain depending on the transmission frequency; therefore for a certain frequency, the attenuation can be very important and entail the loss of the transmitted signal. The coherence bandwidth $B_{coh} = 1/rms$ is the characteristic parameter of the frequency selectivity of the channel, where rms is the maximum root mean square of the path delay.

B. Formulation of the margin adaptive's problem

The problem of margin adaptive deals with the minimization of the total power consumption in the downlink of the cellular system while the request of data rate for users is achieved. The channel transmission is assumed to be characterized by time variation and frequency selectivity. Therefore, the channel presents for users, an specified channel gain which depends on the position et the velocity of the users.

Let's consider $H_{k,n}$ the channel gain of kth user for nth sub-carrier. The required power is given by[11]:

$$p_{k,n} = \frac{f(c_{k,n})}{H_{k,n}^2} \tag{3}$$

where

$$f(c_{k,n}) = \frac{N_0}{3}(2^{c_{k,n}} - 1)[Q^{-1}(\frac{BER_k}{4})]^2 \tag{4}$$

$c_{k,n}$ is the number of bits for kth user on the nth sub-carrier.
N_0 is the power spectral density of noise.
BER_k is the bits error rate of kth user.
$Q(x) = erfc(x)$ is the complementary error function.

We note that when the channel gain $H_{k,n}$ increases the power required decreases. Thus the nth sub-carrier will be allocated in preference to the kth user with least power until the required throughput is reached.
The total power allocated to users is given by:

$$P_T = \sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n}.\rho_{k,n} \tag{5}$$

The total bits for the kth user is given by:

$$r_k = \sum_{n=1}^{N} c_{k,n}.\rho_{k,n} \tag{6}$$

where $\rho_{k,n} = 1$ if nth sub-carrier is allocated to kth user and $\rho_{k,n} = 0$ otherwise.(we consider that nth sub-carrier is allocated only to kth user)
The problem of resource allocation can be written:

$$min(\sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n}.\rho_{k,n}) \tag{7}$$

subject to:

$$\sum_{n=1}^{N} c_{k,n}.\rho_{k,n} > r_0 \tag{8}$$

where r_0 is the minimal data rate for the kth user.

C. Formulation of the rate adaptive's problem

Let's consider N and K respectively the total number of sub-carriers and the total number of the active users in the cell. The access method is based on the OFDMA and each user can have N_k sub-carriers. According to Shannon's theorem [15], the capacity of the sub-carrier n for the kth user is given by:

$$C_{k,n} = \frac{B}{N}\log_2(1 + SNR_{k,n}) \tag{9}$$

where B is the total bandwidth and $SNR_{k,n}$ the signal-to-noise ratio for the sub-carrier n. Let's consider $H_{k,n}$ the gain of the channel for the kth user on the nth sub-carrier. The signal to noise ratio for the kth user with the nth sub-carrier is defined by:

$$SNR_{k,n} = \frac{Pr_{k,n}}{N_0 B_{ch}} \tag{10}$$

where $Pr_{k,n}$ is the power of the received signal, N_0 is the power spectral density of the noise and B_{ch} is the width of one sub-channel. In this work, we assume that all sub-channels are flat during the transmission of the symbol and have the same width. We take into account the conditions of propagation characterized by the path-loss and the shadowing. Therefore, according to the equation of FRIIS[1], the power of the received signal is given by:

$$Pr_{k,n} = \frac{Pe_{k,n}H_{k,n}^2 GeGr}{P_L} \tag{11}$$

where $H_{k,n}$, Ge, Gr, and P_L are respectively the gain of channel for user k on the nth sub-carrier, the gain of the transmitter antenna, the gain of the receiver antenna and the path-loss factor. The path-loss factor is defined by:

$$P_L = P_{L0} + \sigma \tag{12}$$

where P_{L0} is the free space path-loss and σ is the term of shadowing. The path-loss factor is defined in three cases: the indoor radio propagation model, the outdoor to indoor and pedestrian environment and the vehicular environment.

Let's consider R the distance expressed in (km) between the mobile equipment and the base station, n the number of floors in the path and f_n the frequency of the transmission carrier in (MHz). In the case of indoor office radio propagation model, the path-loss is based on the COST 231 model. It's given in typical conditions by [8]:

$$PL_{indoor} = 37 + 30\log_{10}(R) + 18.3n^{(\frac{n+2}{n+1}-0.46)} \tag{13}$$

The path-loss in the outdoor to indoor and pedestrian propagation is the total transmission loss taking into account the reflexion and the diffraction. This path-loss is given by:

$$PL_{outdoor} = 49 + 40\log_{10}(R) + 30\log_{10}(f_n) \tag{14}$$

In the vehicular test environment the path-loss factor is defined by:

$$PL_{vehicular} = [40(1 - 4.10^{-3}\Delta h_b)]\log_{10}(R) \tag{15}$$
$$-18\log_{10}(\Delta h_b) + 21\log_{10}(f_n) + 80 \tag{16}$$

The total throughput can be expressed by:

$$C_T = \sum_{k=1}^{K}\sum_{n=1}^{N} C_{k,n}.\rho_{k,n} \tag{17}$$

where $\rho_{k,n}$ is the indicator of the resources allocation such that $\rho_{k,n} = 1$ if the nth sub-carrier is attributed to the kth user and $\rho_{k,n} = 0$ in otherwise.

By replacing (5) into (3), the total throughput can be written:

$$C_T = \sum_{k=1}^{K}\sum_{n=1}^{N} \frac{B}{N}\log_2(1 + \frac{Pe_{k,n}H_{k,n}^2 GeGr}{P_L N_0 B_{ch}}).\rho_{k,n} \tag{18}$$

Therefore, the main objective of this work is to maximize the total throughput under the constraints of the available power. Indeed, the capacity is proportional with the power of signal $Pe_{k,n}$. Thus, the research of the maximization of the capacity is necessarily obtained with the increase of the power transmitted signal. So, the resolution of this problem is limited by the availability of the total power in the base station.

In this work, we propose to increase the signal-to-noise ratio by applying an efficient resources allocation to users. In fact, the sub-carriers are allocated to users with highest channel gain, in this case the capacity of the system is increased. The power on each sub-carrier is assumed to be constant.

The resources allocation problem can be formulated as:

$$\max(\sum_{k=1}^{K}\sum_{n=1}^{N} \frac{B}{N}\log_2(1 + \frac{Pe_{k,n}H_{k,n}^2 GeGr}{P_L N_0 B_{ch}}).\rho_{k,n}) \tag{19}$$

subject to:

$$\sum_{k=1}^{K}\sum_{n=1}^{N} Pe_{k,n} \leq P_T \tag{20}$$

$$\sum_{k=1}^{K} \rho_{k,n} = 1 \tag{21}$$

where P_T is the total power available in the base station.

The problem described above is highly non linear and there are not an efficient algorithm which gives exact solution.

[1] Harald T. Friis 1893-1976

Therefore, many algorithms are proposed to solve approximately the resources allocation problem. There are the iterative algorithms, and the heuristic approach.

The iterative approach is more simple to implement but its complexity increases highly when the parameters of the problem become very important. In addition, when the problem presents several extrema the iterative algorithms are not appropriated. Indeed, in this situation an iterative approach can converge to a local optimum which is not necessarily the global solution.

In otherwise, the heuristic approach, inspired by natural phenomenon, presents a very high robustness; which permits to efficiently solve the resources allocation problem even if it presents many extrema[6]. The evolutionary algorithm is one part of heuristic approach. The algorithm proposed in this work is based on genetic algorithms inspired by the evolution of species introduced by C. Darwin[2].

IV. Genetic Algorithms

Genetic algorithms are inspired by Darwin's theory of evolution and by Mendel's works about recombination of species[10]. GAs are used to solve many problems of optimization.

Robustness is the main advantage of genetic algorithms relative to traditional resolution methods of optimization problems[6]. In other words, we can see the four major differences between the two methods:

- GAs work with a coding of the set of parameters, while the classical methods use directly the parameters.
- The solution given by GAs is a set of points (chromosomes) and the solution for a classical methods is a single point.
- GAs use the objective function and the standards methods often use derivatives of function or other auxiliary knowledge.
- Gas use probabilistic transition rules when the traditional methods use deterministic rules.

The principle of GAs is based on the evolution of an initial population under the effect of operators such as selection, mutation and crossover. At the end of the GAs's process, the best individual in the population will be the solution of the optimization problem.

The different phases of the GAs are:

- Coding of chromosome: A chromosome of the population represents a resource allocation scheme. The coding of the chromosome is to provide a structure corresponding to the resource allocation problem. In our implementation, chromosome is represented by an array of structure, containing a fixed number of subcarriers, numbered in increasing order, from 0 to $N-1$. In the nth cell, there are the index of the user to which the corresponding subcarrier is allocated. The required power and the quantity of bits are calculated from this allocation for every user. All chromosomes have the same fixed size which corresponds to the total number of subcarriers. The table I shows an example of the structure of the chromosomes.

TABLE I. Structure of Chromosome

Subcarrier	1	2	- - -	N
User	5	3	- - -	10

- Initialization of the population: It consists to set the size of the population and to generate all chromosomes of the population. In this work, the chromosomes of the population are randomly generated and their size is the same and stays constant during the GAs process.
- Evaluation: In this function, the total required power and the total throughput are calculated for each chromosome. The required power corresponds to a fitness (objective function) of the chromosome. The fitness or the objective function represents the criterion of the selection of chromosomes.
- Selection: After evaluating the power of all chromosomes, the selection consists to retain the best chromosome by directly sorting (natural selection) or by organizing tournament between two chromosomes arbitrary selected (selection by tournament). The best chromosome of the population is the one that the requires power is the smallest while respecting the constraint of throughput.
- Mutation: It consists to bring changes to the resources allocation scheme in order to have a better exploration of the search domain. In this work, the number of mutations and the index of the mutated subcarrier are randomly determined for each chromosome. Note that, all chromosomes will not be affected by the mutation.
- Crossover: In this step, a new chromosome is created from two chromosomes in order to take advantages of their best characters. Two points of crossover are randomly chosen and two parts of the first chromosome are concatenated with one part of the second chromosome. So, a new chromosome is created from this concatenation.

V. Descriptions of proposed solutions

The genetic algorithms is among the evolutionary approaches. It's first proposed in engineering by J. Holland[3] and developed by David Goldberg [6] for solving the problem of the control of natural gas pipeline. The principle of genetic algorithms is based on the evolution of the initial population by many factors such as: selection, crossover, recombination and mutation. At the end of the process, a new population is created. These operations are repeated during the cycle of evolution defined by the number of generations.

[2] Charles Robert Darwin, 1809-1882

[3] John Henry Holland, 1929, University of Michigan

Algorithm 1: Genetic Algorithms Approach

```
Data: $N$, $K$, $Size_{pop}$, $N_{gen}$, $H_{canal}$, $P_T$
$Coding\_of\_individual$;
$Init\_population(mat_{pop}, N, Size_{pop})$;
for $i = 0$ to $Size_{pop} - 1$ do
    for $n = 0$ to $N - 1$ do
        $k \leftarrow rand(N)$;
        $Indiv(i, n) \leftarrow k$;
        $P(k, n) \leftarrow P_T/N$;
        $C(k, n) \leftarrow (B/N)\log_2(1 + SNR_{k,n})$;
for $g = 1$ to $N_{gen}$ do
    $selection\_tournament(mat_{pop})$;
    $mutation(mat_{pop})$;
    $crossover(mat_{pop})$;
$solution \leftarrow best\_individual(mat_{pop})$ ;
```

VI. Simulation and Results

A. Parameters of simulation

In figure 3, the model of transmission channel is shown. It's characterized by frequency and time selectivity which take into account the multi-path phenomenon and the Doppler effect. The Jakes' model is used for the simulation of the channel with the parameters given in the table II.

Figure 3 shows that according to the considered frequency , the gain of the transmission channel can vary from 15 dB to -25 dB. This reflects the frequency selectivity of the transmission channel. Note that, this channel gain for a given frequency also varies from one moment to another, reflecting the temporal selectivity of the transmission channel.

TABLE II. Parameters of simulation

Channel bandwidth	20 MHz
PSD of the noise N_0	-174 dBm/Hz
Sub-channel bandwidth	15 kHz
Total power of BS	30 Watts
Radius of cell	2 km
RMS of delay	150 ns
Average speed of users	1 m/s
Numbers of scatters	10
Base Station antenna gain	15 dBi
Mobile Station antenna gain	0 dBi

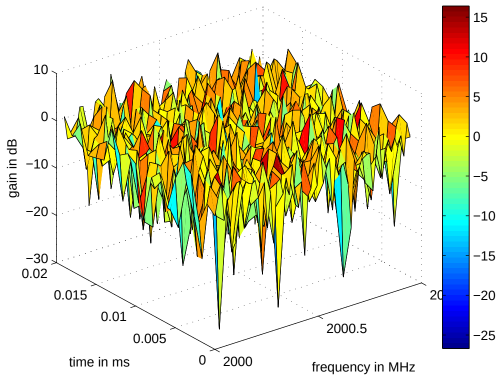


Fig. 3. Variation of channel gain in frequency and time domain

B. Results of simulation

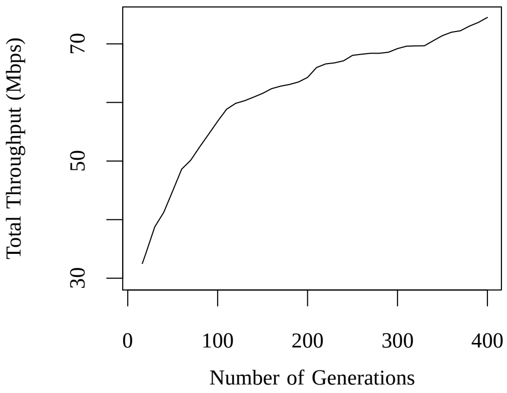


Fig. 4. Total Throughput Vs Nb. generations

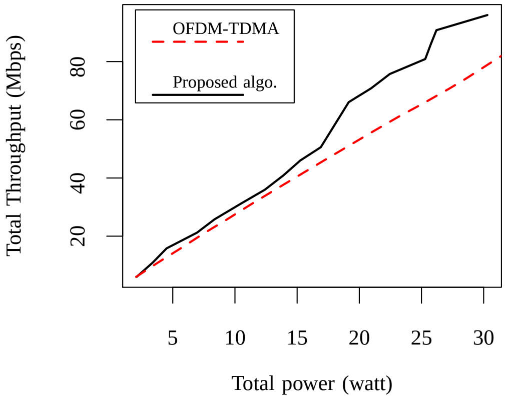


Fig. 5. Total Throughput Vs Total power

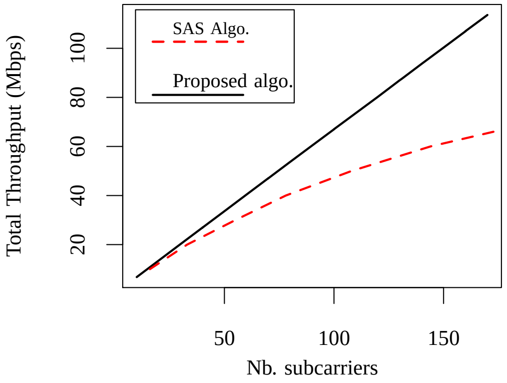


Fig. 6. Total Throughput Vs Nb. subcarriers

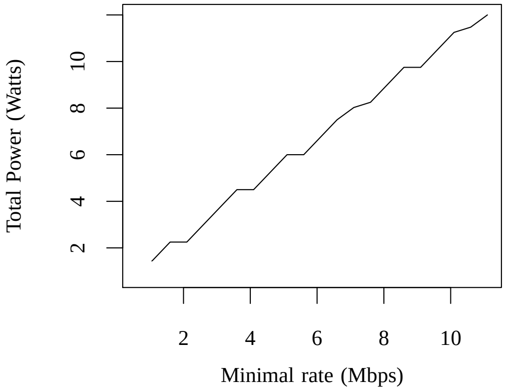


Fig. 7. Total power vs minimal users Throughput

C. Analysis and comments of results

The results obtained with the proposed algorithm are very interesting. Indeed, figure 4 shows the convergence of the proposed method. We note a quick increase of the total throughput when the number of generation grows. However, we note that from the 120th generation, the acceleration of the graph decreases. That corresponds to the neighboring of the optimal solution of the problem given by the proposed method.

In figure 5, we compare the proposed algorithm with the OFDM-TDMA approach. Figure 5 shows that, the proposed algorithm performs the OFDM-TDMA algorithm; and this improvement becomes very important for high data rate. For exmple with power consumption of 25w, the proposed algorithm permits to acheive around 80 Mbps while the OFDM-TDMA gives 60 Mbps.

Also, in Figure 6, we compare the proposed algorithm with the SAS (Simulated Annealing) algorithm proposed in [4]. We note that, for average flows (about less than 50Mbps), the results of our approach are similar to those obtained with SAS algorithm. However, when the required flow rate becomes large(from 60 Mbps), the gain of the proposed solution is more pronounced relative to the SAS algorithm. Sometimes, the proposed approach provides similar flow with more than half of number of sub-carriers used by the SAS method. As an exemple, the proposed algorithm gives 100 Mbps with 150 sub-carriers while the SAS method does not exceed 60 Mbps with more than 150 sub-carriers.

Figure 7 shows the evolution of the users's consumed power based on the minimum required data rate for users. We note that, the consumed power does not evolve linearly with the required users's throughput. Better, figure 7 shows a decrease in the slope of the curve of evolution when the users's throughput increases. This reflects a minimization of the power consumption depending on the data rate.

VII. Conclusion

In this paper, we have proposed the genetic algorithms to solve the problem of resources allocation. The study took place in the context of frequency selectivity and time-variant of the transmission channel. In addition, we have taken into account the propagation loss; which permits to obtain the effective received throughput for users. The proposed implementation of genetic algorithms has led to efficiently solve the problem of resources allocation, and it's better than the OFDM-TDMA approach and the algorithm SAS. The obtained solution is very similar to the standard specifications of OFDMA-based systems in terms of maximum data rate and capacity of users. In this article, the choice of the methods of comparison with our approach is firstly justified by the good performance of some approaches: it's the case of SAS algorithm [4], on the other hand by the accessibility of certain resolution methods found in the literature: it's the case of OFDM-TDMA approach. On the other side, comparison with methods based on evolutionary approach is made difficult by the lack of details on the algorithms presented in most of the articles that we consulted [1][12][17].

However, the proposed method solves the problem of resources allocation in mono-cell case. We hope that this approach will give interesting results when applied to the case of multi-cell system and in radio cognitive systems. In future work, the integration of new metric of QoS measurement of cellular systems such as an effective spectral efficiency area [9], will constitute the basis of a new development in order to achieve a more effective resolution of the resources allocation problem in radio mobile systems.

References

[1] H. Ahmadi and Y. Chew. Adaptive subcarrier-and-bit allocation in multiclass multiuser ofdm systems using genetic algorithm. 2009.

[2] M. Alouini and A. Goldsmith. Adaptive m-qam modulation over nakagami fading channels. *Proc. IEEE. Global Communication Conf.*, pages 218–223, 1997.

[3] Q. An and Y. Yang. Efficient water-filling algorithm for power allocation in ofdm-based cognitive radio systems. *CSE Conference and Workshop Papers*, (196), Janvier 2012.

[4] J. Farah and F. Marx. Greedy algorithms for spectrum management in ofdm cognitive systems - applications to video streaming and wireless sensor networks. *IntechOpen*, November 2008.

[5] S. Pleftschinger G. Münz and J. Speidel. An efficient waterfilling algorirhm for multiple access ofdm. Stuttgart Germany, 2002.

[6] David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. MA Addison Wesley, USA, 1th edition, 1989.

[7] T. Ibaraki and N. Katoh. *Resource Allocation Problems-Algorithmic Approches*. MIT Press, Cambridge MA USA, 1988.

[8] ITU. Guidelines for evaluation of radio transmision technologies for imt-2000. Technical report, International Telecommunication Union, Place des Nations 1211 Geneva 20 Switzerland, 1997.

[9] A. Omri M. O. Hasna and M. Nafie. Effective area spectral efficiency for wireless communication networks with interference management. *EURASIP Journal on Wireless Communications and Networking*, August 2015.

[10] Mélanie Mitchell. *An Introduction to Genetic Algorithms*. MIT Press, USA, 1th edition, 1999.

[11] John G. Proaski. *Digital Communications*. McGraw Hill, New York USA, 4th edition, 2001.

[12] Y. Reddy and N. Gajendar. Evolutionary aproach for efficient resource allocation in multi-user ofdm systems. Las Vegas Nevada USA, 2007.

[13] Hermann Rohling. *OFDM: Concepts for Future Communication Systems*. Springer, Hamburg Germany, 2011.

[14] Henrik Schulze and Christian Lüders. *Theory and Apllication of OFDM and CDMA*. John Wiley & Sons Ltd, Germany, 1th edition, 2005.

[15] Claude E. Shannon. Communication in the presence of noise. volume 86, February 1998.

[16] J. C. Tu and J. M. Cioffi. A loading algorithm for the concatenation of coset codes with multichannel modulation method. *IEEE GLOBECOM*, December 1990.

[17] N. Zhou X. Zhu and Y. Huang. Genetic algorithms based crosslayer resource allocation for wireless ofdm networks with heterogeneous traffic. volume 6, pages 1656–1659, Glasgow Scotland, Aug. 2001.

Image Enhancement Using Homomorphic Filtering and Adaptive Median Filtering for Balinese Papyrus (*Lontar*)

Ida Bagus Ketut Surya Arnawa
Stmik Stikom Bali
Bali, Indonesia

***Abstract*—Balinese papyrus (*Lontar*) is one of the most popular media to write for more than a hundred years in Indonesia. Balinese papyrus are used to document things that are considered important in the past. Most of the balinese papyrus suffered damage caused by weathering, edible fungus and insects making it is difficult to read. One of the efforts made to preserve the existence of balinese papyrus is to perform digitization of it. The problems most often encountered in the process of digitizing the image of the balinese papyrus is less good results as there is noise caused by its conditions that have been damaged and the uneven distribution illumination in this part of the image. In this study the authors propose to combine homomorphic filtering with adaptive median filtering to perform image enhancement. Surve results obtained show the percentage of the average respondents stated that the image enhancement results are good is 83.4%, the percentage of the average respondents stated that the image enhancement results are very good is 4% and the percentage of the average respondents stated that the image enhancement results are enough is 12, 6%.**

Keywords*—*Image Enhancement; Balinese Papyrus; Homomorphic Filtering; Adaptive Median Filtering; Otsu Binarization

I. Introduction

Balinese papyrus (*Lontar*) is one of the most popular media to write for more than a hundred years in Indonesia. Balinese papyrus used to document things that are considered important in the past. The tools used for writing on balinese papyrus is *pengutik / pengrupak*. *Pengutik / Pengrupak* made of steel or iron is a long rectangular, measuring 2 x 15 cm with a thickness between 1.5 mm to 2 mm. The content in balinese papyrus containing spell, as religious teachings, history, stories, songs, knowledge of astronomy and astrology (*Wariga*), philosophy of life and many other useful knowledge for us as a guide in conducting life [1].

Most of the balinese papyrus in Indonesia is currently experiencing damage caused by weathering, edible fungus and insects so it is difficult to read [1]. To preserve the balinese papyrus various attempts have been made either by the government or other institutions. One of the efforts made to preserve the existence of balinese papyrus is to perform scanning or imaging utilizing digital photos to make balinese papyrus as a digital image. In some cases the process of digitizing the balinese papyrus produces digital images that are less than good so it is necessary to perform image enhancement.

A fairly complete review is given by S.R.Yahya et al. on the method of image enhancement of an old manuscript with the background conditions that have been damaged [2]. In outline there are three types of image enhancement method, they are (a) image enhancement and thresholding binarization method, (b) image enhancement with hybridization methods between binarization / thresholding with other methods, and (c) image enhancement without thresholding method. Results of a review from [2] explains that the second method can improve image enhancement. Image Enhancement of the palm leaf manuscript using normalization techniques proposed by Z. Shi et al. [3]. Z. Shi et al. propose a set of transformation methods for processing image enhancement on a palm leaf manuscript. The first model to perform background color approximation with linear and non-linear models. Normalization techniques conducted adaptively in certain local areas. Z. Shi et al. proposed technique generates pretty good image enhancement, but there are still a lot of noise left on the final results.

The problems most often encountered in the process of digitizing the image of the balinese papyrus is there are less good results as there is noise caused by the balinese papyrus conditions that have been damaged and the uneven distribution illumination in this part of the image [4]. Deployment of uneven illumination caused by the balinese papyrus conditions that have been warped or damaged and are not allowed to use the scanner in the process of digitization. To do digitizing on the curved or broken papyrus is to use photographic equipment so that there is a distance between the papyrus with photographic equipment that result in uneven illumination. In this study the authors propose to combine homomorphic filtering with adaptive median filtering to perform image enhancement. Homomorphic filtering is used to normalize the uneven background and adaptive median filtering is used to eliminate noise found on balinese papyrus. Image that has been processed with a combination of homomorphic filtering with adaptive median filtering then binarized with global binarization using otsu binarization method without the image is divided into local images.

II. Literature Review

A. Papyrus (Lontar)

Papyrus is one heritage spiritual wealth of the archipelago which has a very important meaning and strategic [5]. Bali is one of the places in the archipelago known to find papyrus, but some were found in Sulawesi, Java and Lombok. Before the paper is founded, papyrus is one of media that is used for writing and documenting the various things that are considered important in the past [5]. Besides papyrus as the media for writing there was also found other media such as Java used nipa leaf (similar papyrus), perkamen (skin of goat) and dluwang (skin of kind of wood). The tools used for writing on balinese papyrus is *pengutik / pengrupak*. It is made of steel or iron is a long rectangular, measuring 2 x 15 cm with a thickness between 1.5 mm to 2 mm. In the papyrus contained spells, as religious teachings, history, stories, songs, knowledge of astronomy and astrology (*Wariga*), philosophy of life and many other knowledge [1]. Fig. 1 is an example of papyrus.

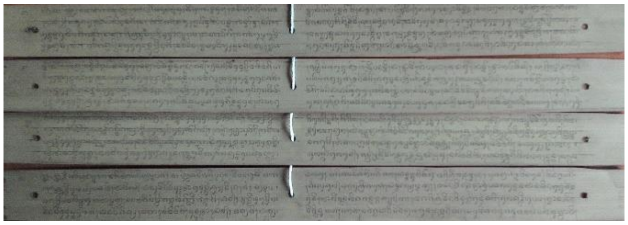


Fig. 1. Balinese Papyrus *Warigasari Kyastapaka*

B. Homomorphic Filtering

In image processing, homomorphic filtering is one method that can be used to compensate for the effects from uneven illumination on the image and enhance the appearance of simultaneous image compression varying intensity and contrast enhancement [6], [7]. According to this model, an image has the following equation:

$$f(x,y) = i(x,y)r(x,y), \tag{1}$$

where $f(x,y)$ is an image that is the result of multiplication (product) from $i(x,y)$ which is a component of illumination with $r(x,y)$ which is a component reflectance. To separate two independent components and facilitate their separate processing, logarithm transform on (1) has been taken, Thus

$$z(x,y) \;= \ln f\,(x,y) \tag{2}$$

$$= \ln i(x,y) + \ln r(x,y), \tag{3}$$

then, the Fourier transform of (3) is calculated:

$$\Im\,\{z(x,y)\} \;= \Im\,\{\ln f\,(x,y)\} \tag{4}$$

$$= \Im\,\{\ln i\,(x,y)\} + \Im\,\{\ln r\,(x,y)\} \tag{5}$$

or

$$Z(u,v) \;= F_i(u,v) + F_r\,(u,v), \tag{6}$$

where $F_i(u,v)$ and $F_r\,(u,v)$ is the Fourier transform from $\ln i(x,y)$ and $\ln r(x,y)$. After being moved in the frequency domain, then the image is processed by using appropriate filters so that the initial goal can be achieved is to weaken the low frequency and high frequency amplify resulting in image enhancement and image sharpening by the formula:

$$S(u,v) \;= \mathrm{H}(u,v)Z(u,v) \tag{7}$$

$$= H(u,v)\,F_i(u,v) + H(u,v)F_r\,(u,v), \tag{8}$$

where $S(u,v)$ is the Fourier transform from image that has been processed. So as to obtain the actual results need to be returned to the spatial domain by the formula:

$$S(x,y) = \Im^{-1}\,\{S(u,v)\} \tag{9}$$

$$= \Im^{-1}\,\{H(u,v)F_i(u,v)\} + \Im^{-1}\,\{H(u,v)F_r(u,v)\} \tag{10}$$

by defining

$$i'(x,y) = \Im^{-1}\,\{H(u,v)F_i(u,v)\} \tag{11}$$

and

$$r'(x,y) = \Im^{-1}\,\{H(u,v)F_r(u,v)\}. \tag{12}$$

Equation (10) can be expressed as follows:

$$s\,(x,y) = i'(x,y) \;+ r'(x,y), \tag{13}$$

the final step is to eliminate logarithms operations conducted at the beginning of the process by performing an exponential operation in order to obtain the desired enhanced image is denoted by $\mathrm{g}(x,y)$ is:

$$\mathrm{g}(x,y) = e^{s(x,y)} \tag{14}$$

$$= e^{i\prime(x,y)}e^{r\prime(x,y)} \tag{15}$$

$$= \; i_0(x,y)r_0(x,y), \tag{16}$$

where $i_0(x,y) = e^{i'(x,y)}$ and $r_0(x,y) = e^{r'(x,y)}$ is the illumination and reflectance components from output each image. The $\mathrm{H}(u,v)$ normally used in this procedure is the Butterworth high pass filter defined as :

$$H(u,v) = (\gamma_H - \gamma_L)\left(\frac{1}{1+(D_0/D(u,v))^{2n}}\right) + \gamma_L\,, \tag{17}$$

where D_0 is the cut off distance measured from the origin, $D(u,v)$ is distance from the origin of centered Fourier transform, and n is the order of the Butterworth filter.

C. Adaptive Median Filtering

The problem faced by the standard median filtering can be resolved with adaptive median filtering. Between the adaptive median filtering and median filtering have a fundamental difference in which adaptive median filtering window surrounding each pixel is variable [8]. In the standard median filter does not take into account variations in image characteristics from one point to another [9]. Flowchart from adaptive median filtering shown in Fig. 2.

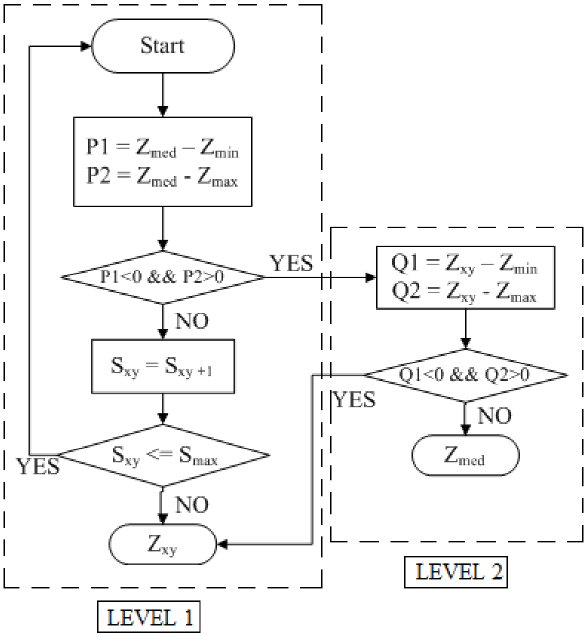


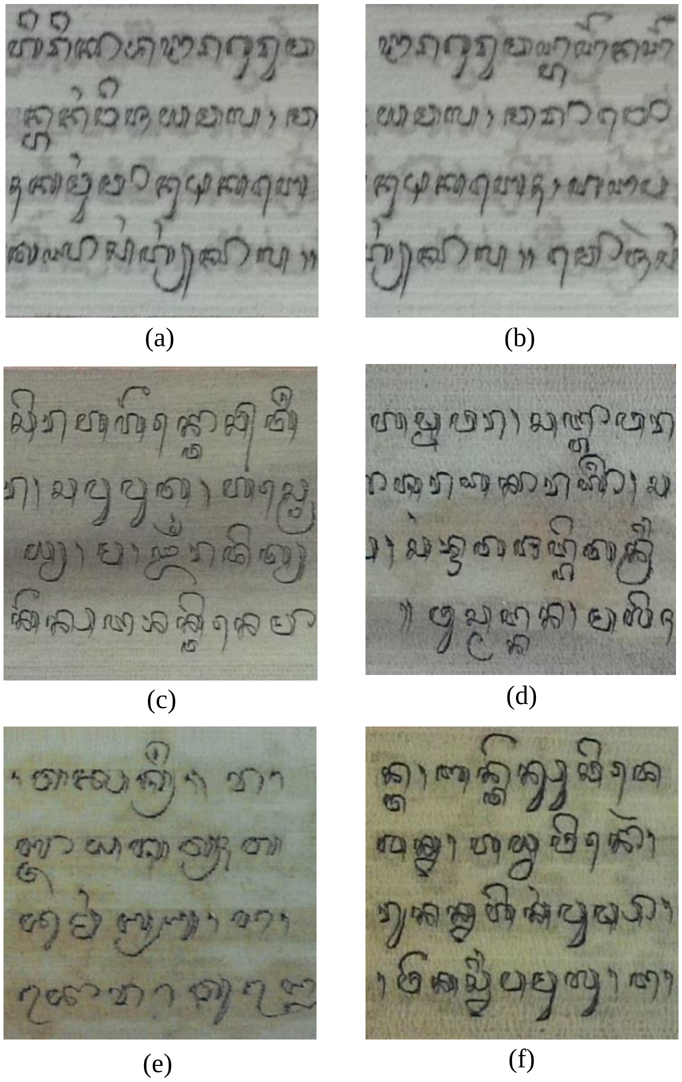
Fig. 2. Flowchart Adaptive Median Filtering

In the Fig. 2 adaptive median filtering works with 2 levels, they are level 1 and level 2. At level 1 serves to determine whether the median filter output Z_{med} is impulse output or not. If the impulse output is not found on the first level then it will proceed to level 2. At level 2, adaptive median filtering would increase the size of the window and repeat the process at the level of 1 to find that the median value is not impulse or the maximum window size has been reached. Adaptive median filtering can reduce computational overhead because every time the value of the output of the adaptive median filtering algorithm, *S*xy window is moved to the next location in the image. Adaptive median filtering algorithm then performs reinitialized and applied to a pixel in a new location. Adaptive median filtering has three main objectives : repair the damaged image that is caused by salt and impulse noise, give smoothing from non-impulsive noise and reduce the disturbance caused by thinning or thickening of object excess limit [9].

III. Methodology

A. Data Acquisition

Data acquisition is the process of acquiring data from analog to digital, from balinese papyrus physically into image files by using the digital camera or scanner. Balinese papyrus were in good condition scanned using the scanner while the balinese papyrus that is in poor condition scanned using a digital camera. Results from scanning balinese papyrus stored in a computer before further processing. Balinese papyrus image used in this study can be seen in Fig. 3 (a) to (f).



(a) (b)

(c) (d)

(e) (f)

Fig. 3. Example of input images

B. Algorithm

Result from the acquisition of the data stored in the computer are divided into local images. Local images is first converted into a grayscale image. Then the result from grayscale images is filtered by using homomorphic filtering and adaptive median filtering. Filtration results of each section then are binarized with otsu binarization and the results are put back together into a complete image [10]. Next binarization results were evaluated by questionnaire to some correspondents to assess the results from image enhancement. Fig. 4 is the steps performed in this study.

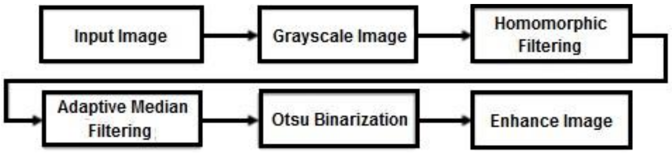


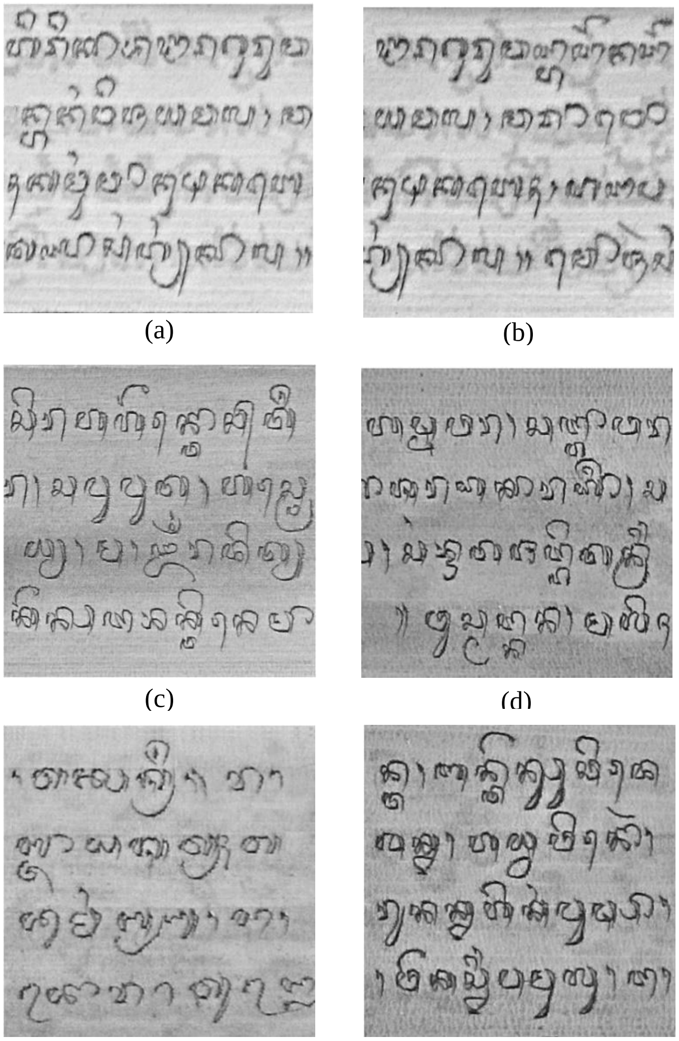
Fig. 4. Algorithm Based Process

IV. RESULT AND DISCUSSION

A. Result

1) Homomorphic Filtering

Homomorphic filtering is applied in this study using a Butterworth filter that is adopted from [11]. Butterworth filter formula is shown in (17), γ_H parameter value, and γ_L respectively are 1.2 and 0.03 and the value of n = 1. The series of homomorphic filtering are implemented in the matlab. Results from homomorphic filtering process shown in Fig. 5. Each image in Fig. 5 (a) to (f) is the result from homomorphic filtering from the image in Fig. 3 (a) to (f).



(a) (b)

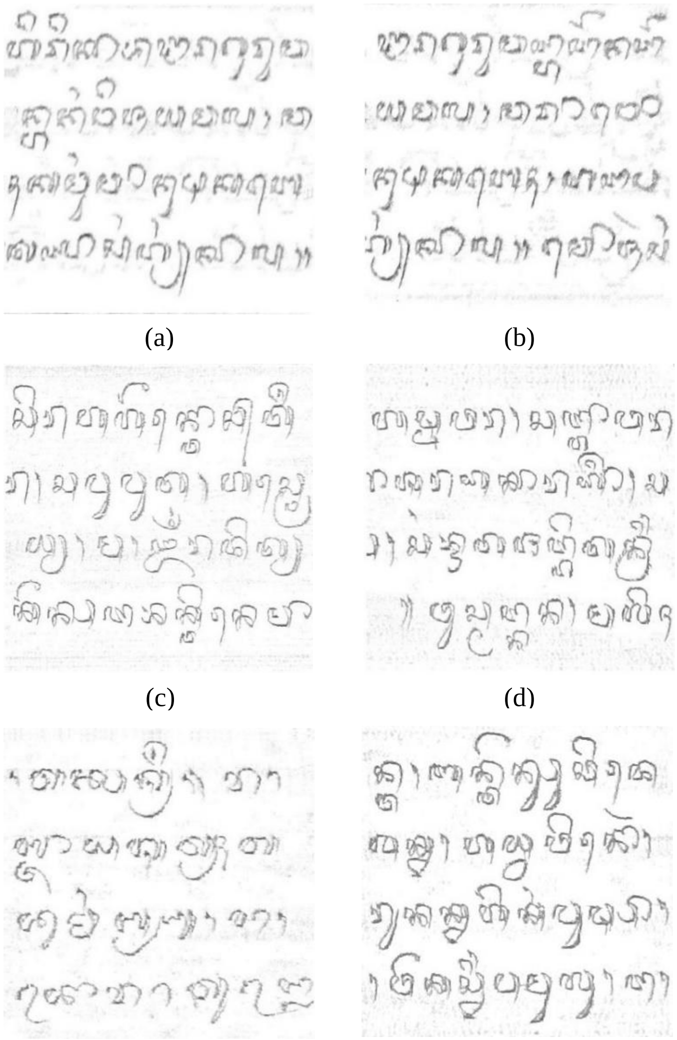
(c) (d)

(e) (f)

Fig. 5. Result Homomorphic Filtering

2) Adaptive Median Filtering

Adaptive median filtering is applied to the image enhancement research using ws parameter values and c respectively are 100 and 0,001. The series of adaptive median filtering are implemented in the matlab. Results from adaptive median filtering process are shown in Fig. 6. Each image in Fig. 6 (a) to (f) is the result from adaptive median filtering from the image in Fig. 5 (a) to (f).



(a) (b)

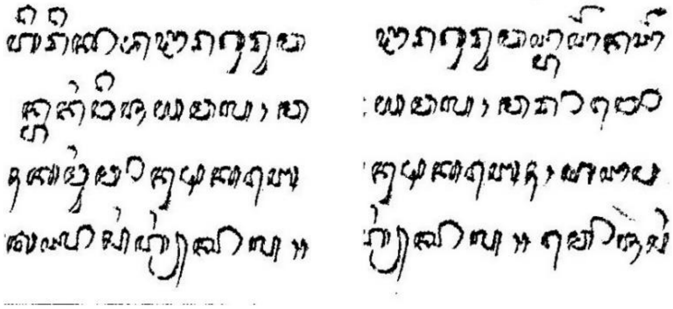
(c) (d)

(e) (f)

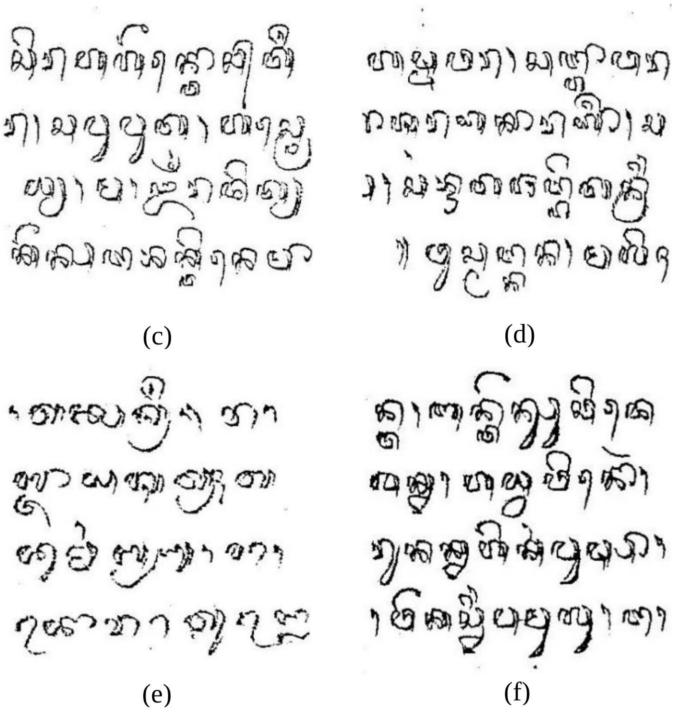
Fig. 6. Result Adaptive Median Filtering

3) Otsu Binarization

After the homomorphic filtering and adaptive median filtering, then it is followed by binarization process using otsu method [11]. Results from the process otsu binarization method shown in Fig. 7 (a) to (f).



(a) (b)



(c) (d)

(e) (f)

Fig. 7. Result Otsu Binarization

4) Field Trial

In the field trial, the authors tested the results from image enhancement using a questionnaire to determine the quality from the resulting image enhancement. In conducting the questionnaire, the authors involved 27 people who come from the balinese language lecturer, teacher and student of balinese language. In the process of filling in the questionnaire, authors included balinese papyrus to compare the original text which is in original balinese papyrus with enhancement of existing results in the form of questionnaires. Diagrams of score percentage of the answers given by the respondents can be described as Fig.8.

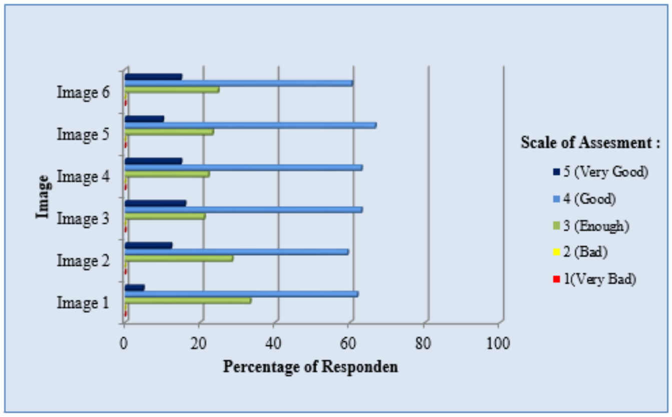


Fig. 8. Percentage Diagram of Respondents Answer Score In Field Trial

Based on the percentage of respondents answer scores diagram in field trial, it can be viewed that the test results from image enhancement is good. This is evidenced by the average percentage of stating the results from image enhancement are good is 83.4% and the average percentage of stating the results from image enhancement are excellent is 4% as well as the average percentage of stating the results from image enhancement are enough is 12.6%.

B. Discussion

From the results obtained can be viewed that the homomorphic filtering and adaptive median filtering has been able to perform image enhancement at balinese papyrus. On the results of image enhancement, there are still some dotted text. Text experiencing this dot caused by the otsu binarization result from parts of the input image is categorized as part of the background and rated the white pixel so that the texts are experiencing dashed in some parts. Moreover, in the image that has a lot of noise, some parts from image are categorized includes the foreground and rated the pixel black, causing there are still some residual noise. Though the results is ideally every part of foreground converted to black and every backgound converted into white. Otsu binarization errors can be overcome by performing filtration processes using homomorphic filtering first and then the results from homomorphic filtering is filtered again using adaptive median filtering before otsu binarization process. Results from homomorphic filtering process shown in Fig. 5 and the results from adaptive median filtering shown in Fig. 6. Filtration process with homomorphic filtering and adaptive median filtering has managed to homogenize backgound and minimize the noise level so that the results from the otsu binarization process can be better in which pixels are categorized into the foreground section and the incoming pixel backgound section as shown in Fig. 7.

Testing of the results of image enhancement is done by giving questionnaires to 27 people consisting from the balinese language lecturer, teacher and student of balinese language. Results from questionnaire that has been done shows that the input image has the 5th lowest percentage is 76.5% said good, 21.5% said enough and 2% said very good. 5th input image is an image with a backgound which has a lot of noise and have low contrast between the foreground with the background. 5th input image shown in Fig. 3 (e). The results are less good at the 5th input image because some parts from foreground categorized as part of backgound and rated as white pixels, resulting in the yield on some text experiencing dashed.

V. Conclusions

Based on the results it can be concluded the combination of homomorphic filtering with adaptive median filtering are good for enhancing old balinese papyrus. Steps taken to perform image enhancement on balinese papyrus is to use a filtration process homomorphic filtering and adaptive median filtering and then binarization is performed by using the otsu binarization method. Results from image enhancement on balinese papyrus using homomorphic filtering and adaptive median filtering shows damaged balinese papyrus image could be improved better. Results of image enhancement was tested by conducting questionnaire to 27 people who come from the balinese language lecturer, teacher and student of balinese language and produces a high percentage for a category of good image enhancement results. In future studies a new method will be added to enhance the results obtained from the combination of homomorphic filtering with adaptive median filtering and will be tested with different case studies.

REFERENCES

[1] M. Sudarma, "Identifying of the Cielab Space Color for the Balinese Papyrus Characters," Telkomnika Indonesian Journal of Electrical Engineering, vol. 13, pp. 321-328, February 2015.

[2] S.R. Yahya, S.N.H.S. Abdullah, K. Omar, M.S. Zakaria, and C.–Y. Liong, "Review on Image Enhancement Method of Old Manuscript with Damaged Background," in International Journal on Electrical Engineering and Informatics, vol. 2, January 2010.

[3] Z. Shi, S. Setlur, and V. Govindaraju, "Digital Enchancement of Palm Leaf Manuscript Images using Normalization Techniques," 5th International Conference On Knowledge Based Computer Systems, pp.19-22, December 2005.

[4] C.H. Chou, W.H. Lin, and F. Chang, "A Binarization Method with Learning-Built Rules for Document Images Produced by Camera" Pattern Recognition, vol.43,pp. 1518-1530, November 2009.

[5] M. Sudarma, and N.P. Sutramiani, "The Thinning Zhang-Suen Application Method in the Image of Balinese Scripts on the Papyrus," International Journal of Computer Applications, vol. 91,April 2014.

[6] R. Bock, J. Meier, L.G. Nyul, J. Hornegger, and G. Michelson, "Glaucoma Risk Index : Automated Glaucoma Detection from Color Fundus Images," Medical Image Analysis, vol.14,pp. 471-481, June 2010.

[7] R. Gonzalez, and R.E. Woods, Digital Image Processing, 3rd ed. NJ, USA: Prentice Hall, 2008.

[8] D. Dhanasekaran, A. Krishnamurthy, and Ramkumar, "High Speed Pipeline Architecture for Adaptive Median Filter," Proceedings of the International Conference on Advances in Computing, Communication and Control, pp. 591-600, 2009.

[9] S. Sarker, S. Chowdhury, S. Laha, and D. Dey, "Use of Non-Local Mean Filter to Denoise Image Corrupted by Salt and Pepper Noise," Signal & Image Processing : An International Journal (SIPIJ), vol. 3,pp. 223-235,April 2012.

[10] N. Otsu, "A Threshold Selection Method from Gray-Level Histogram," Automatica, vol.11, pp. 23-27, 1975.

[11] H. Shahamat, and A.A. Pouyan, "Face Recognition Under Large Illumination Variations Using Homomorphic Filtering in Spatial Domain," Journal of Visual Communication and Image Representation, vol. 25, pp. 970-977, March 2014.

Model Checking Self-Stabilising in Embedded Systems with Linear Temporal Logic

Rim Marah, and Abdelaaziz EL Hibaoui
Faculty of Science
Abdelmalek Essaâdi University,
P.O. Box 2121, Tetuan, Morroco

***Abstract*—Over the past two decades, the use of distributed embedded systems is wide in many applications. One way to guarantee that these systems tolerate transient faults is done by making them self-stabilizing systems, which automatically recover from any transient fault.**
In this paper we present a formalism of self-stabilization concept based on Linear Temporal Logic (LTL), and model checked the self-stabilization in embedded systems. Using a case study inspired by industrial practice, we present in detail a model checking to verify the self stabilization property of our embedded system.

***Keywords*—Distributed Embedded Systems, Linear Temporal Logic, Self-Stabilization, Model Checking, Verification**

I. Introduction

A general-purpose definition of embedded systems is that they are devices used to control, monitor or assist the operation of equipment, machinery or plant. Embedded reflects the fact that they are an integral part of the system. In many cases, their embeddedness may be such that their presence is far from obvious to the casual observe [1]. There are cases in which reliability is the most important requirement of those systems. To guarantee that these systems tolerate transient faults, we should make them self-stabilizing systems. This type of fault-tolerance is desirable in many distributed embedded systems [2], [3]. Verifying the correctness of those systems is a challenging task while testing them is intractable.

To make a rigorous verification of these systems, properties should be described in a precise and unambiguous manner. This is typically done by using properties specification language. There is several variety of different logics, according to the types of properties that they can express. In particular, we will focus on the use of LTL (Linear Temporal Logic) [4] [5], as a property specification language. To verify that a system meets its specification, we use the connection between temporal logic and the automata theoretical approach to model checking [5] [6]. The latter applies the intimate relationship between LTL and automata on infinite words. In [7] it was first proven that the set of infinite words, defined by LTL formula, can be accepted by some automaton on infinite words. Several procedures have been suggested which construct a generalized Bchi automaton that recognizes all models of a LTL formula [8] [9] [10].

Within computer science, the LTL has achieved a significant role in formal specification and verification of concurrent reactive systems. It is a very powerful specification mechanism, since it allows the expression of complex requirements through simple formulas. Actually it is widely used for verification of software systems [11] [12]. It provides a formal specification mechanism allowing the quantitative definition of the desired behaver of a systems. It makes it possible to succinctly express complex objectives due to its similarity to natural language.

The model checking is an automatic technique for verifying correctness properties of reactive systems, the two feathers that made the model checking [7] [12], so popular are that it can easily be automated and that is often able to produce a counterexample when the system can not meets its specifications. However the applicability of model checking is limited by the problem of the state space explosion. But luckily, the model checking on the fly is a remedy to this problem.

The linear temporal logic can be appropriate way to formalize the definition of self stabilisation of the systems. The Self-stabilizing systems witch were first introduced by Dijkstra in 1974 [13] [14] are the systems that can start in any global configuration and achieve behaviour meeting the task specification by them self.

Our main contribution in this paper is a proposition of an LTL formalism of the self stabilization concept, and model checking the embedded systems in order to verify there self-stabilization, based on the different phases of model checking Process:

i. Modelling the system: We will model our embedded system. For this modelling, we use the transition system and KRIPKE stricture that we will define properly later.

ii. Specification of the property to check: In this step, we will present our formalism of the self-stabilization concept. Basing on the LTL language, we will check the specification of our property.

iii. Using an algorithm or method to check whether the property is satisfied by the model: For this step we choose the model checking as a method of verification.

This paper is organized as follows: After the introduction, we define the tools used in modelling our system and we describe our model, in the second section called modelling the system. In the section tree named the specifying of the property, we put on the notation used in this paper and we present our formalism of the self stabilization based on the LTL logic. In the forth section named algorithm of verification, we describe in details the steps of model checking algorithm that we use to verify our formalism. Finally in the conclusion, we conclude and give a future extension of our work.

II. System Modelisation

The semantic framework for algorithmic verification systems is given by transition systems and Kripke structures. These later and automata are used to model the reactive systems. They must then validate the model by determining if it satisfies the required properties of the systems. The system of ownership is expressed in terms of its states, transitions or paths.

In this part we are going to give the formal definitions of transition systems and Kripke stricture used to describe our system and define in details our model.

A. Background

Transition systems Transitions systems define the possible states of a system, its initial states and transitions. They provide a framework for describing the operating semantics for reactive system. To describe the behaviour of systems, we can model them by transition systems with are a digraphs where nodes represent states, and edges model transitions.

A transition system TS is a tuple (S, S_0, R) where

- S is a set of states.
- S_0 is a set of initial states.
- $R \subseteq S \times S$ is a transition relation.

To define Kripke stricture we use the usual definition of transitions systems and extended it by adding a labelling of states with atomic propositions.

Definition AP is a set of atomic propositions. A Kripke structure on AP is $M = (S, S_0, R, L)$ where:

- (S, S_0, R) is a transition system.
- $L : S \longrightarrow 2^{AP}$ is a function that label each state $s \in S$ the set $L(s)$ of atomic propositions true in s.

A path π of a M is a path of transitions system associated.

The trace σ of π is ω-word $L(\pi) = L(s_0)L(s_1)...$ in the alphabet $\sum = 2^{AP}$

1) Our Model: Our model is a parts conveying robot presented with finite state model. It's as a case study of embedded systems. This example is presenting in [15]. The robot has 3 devices: An inlet device parts called Dp, a workpiece transport device, called Td which is an arm provided with a clamp, and finally a workpiece removal device called Rd, where Td transports parts arriving on Dp. At a certain level of abstraction, the system is defined by three principles operations that can be stated as:

1) Td transport device files the piece on the discharge device if De is free.
2) Td transport device can mount just if it is taking up a devise.
3) Td transport device can get off just if it is empty.

For simplicity, we ignore Dp and will only interested in the introduction of the devise in Td and its release on De. We will not take in consideration operations (opening, closing, etc.) of the clamp, we will essentially look at the transport parts. All the developments of the system, in this operation mode is represented by the graph in bellow: The drawn graph is called reachability graph. It represents all possible executions of the system. These are infinite sequences of couples (state transition) represented by a finite graph. As there is no distinction of terminal states, we say that this structure is a transition system. The states are decorated by the values of variables De and Td, where 'free' and 'occupied' values are represented schematically by the presence or absence of a part shown by a gray rectangle. TS with decorated states is called Kripke Structure. It is the The formal model that represents all executions.

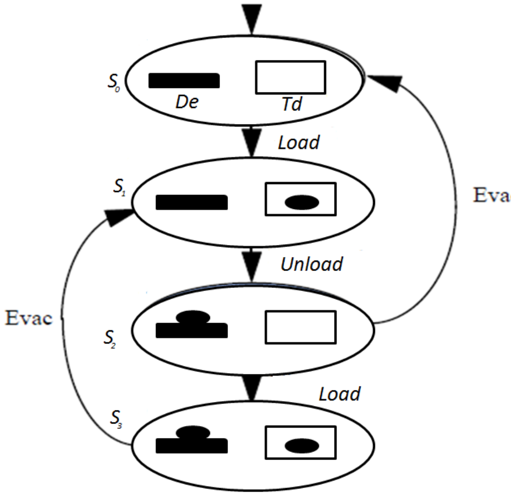


Fig. 1: Kripke Stricture of parts conveying robot

III. Property Specification

Temporal logic was originally developed in order to represent tense in natural language. Temporal logic extends propositional or predicate logic by modalities that permit to referral to the infinite behaviour of a reactive system. They provide a very intuitive and mathematical precise notation for expressing properties about the relation between the state labels in executions. Temporal logic allows for the specification of the relative order of events. The linear temporal logic has achieved its popularity from the number of useful concepts that can formally and concisely be specified by using it.

A. Linear Temporal Logic

The linear temporal logic extends classical logic by temporal modalities. Its formulas are interpreted on infinite sequences of states such as executions of a Kripke structure. Before introducing LTL in more detail, LTL may be used to express the timing for the class of synchronous systems in which all components proceed in a lock-step fashion. In this setting, a transition corresponds to the advance of a single time-unit. The underlying time domain is thus discrete, i.e., the present moment refers to the current state, and the next moment corresponds to the immediate successor state.

This subsection describes the syntactic rules according to which formulae in LTL can be constructed. The basic ingredients of LTL-formulae are atomic propositions. The Boolean connectors like conjunction $\wedge$, and negation $\neg$, and two basic temporal modalities $\bigcirc$ (pronounced next) and $\cup$ (ponounced until). The elementary temporal modalities that are present in most temporal logics include the operators:

- $\Diamond$ eventually (eventually in the future).
- $\Box$ always (now and forever in the future).

1) LTL Syntax: LTL formulae over the set AP of atomic proposition are formed according to the following grammar:

$\varphi ::= true \mid a \mid \varphi_1 \wedge \varphi_2 \mid \neg\varphi \mid \bigcirc\varphi \mid \varphi_1 \cup \varphi_2$

where $a \in AP$.

We mostly abstain from explicitly indicating the set AP of propositions as this follows either from the context or can be defined as the set of atomic propositions occurring in the LTL formula at hand.

The precedence order on the operators is as follows: The unary operators bind stronger than the binary ones. $\neg$ and $\Box$ bind equally strong. The temporal operator $\cup$ takes precedence over $\wedge$, $\vee$ and $\rightarrow$ Parentheses are omitted whenever appropriate.

LTL formulae stand for properties of paths or in fact their trace. A path can either fulfil an LTL-formula or not. To precisely formulate when a path satisfies an LTL formula, we can follow this steps: First, the semantics of LTL formula φ is defined as a language $Words(\varphi)$ that contains all infinite words over the alphabet 2^{AP} satisfy φ. That is, to every LTL formula a single LT property is associated. Then, the semantics is extended to an interpretation over paths and states of a transition system.

2) LTL Semantic: Let φ be an LTL formula over AP. The LT property induced by φ is: $words(\varphi) = \sigma \in (2^{AP})^\omega \mid \sigma \models \varphi$ where the satisfaction relation $\models \subseteq (2^{AP})^\omega \times LTL$ is the smallest relation with the properties.

Here, for $\sigma = A_0 A_1 A_2 \cdots \in (2^{AP})^\omega, \sigma[j \ldots] = A_j A_{j+1} A_{j+2} \ldots$ is the suffix of σ starting in the (j+1)st symbol A_j .

- $\varphi \models true$
- $\varphi \models a$ iff $a \in A_0 (i.e, A_0 \models a)$
- $\varphi \models \varphi_1 \wedge \varphi_2$ iff $\sigma \models \varphi_1$ et $\sigma \models \varphi_2$
- $\varphi \models \neg\varphi$ iff $\sigma \not\models \varphi$
- $\varphi \models \bigcirc\varphi$ iff $\sigma \in [1 \ldots] = A_1 A_2 A_3 \ldots \models \varphi$
- $\varphi \models \varphi_1 \cup \varphi_2$ iff $\exists j \geqslant 0$ such that $\sigma[j \ldots] \models \varphi_2$ and $\sigma[i \ldots] \models \varphi_1$, for all $0 \leqslant i < j$

Essentially, temporal logic extends classical propositional logic with a set of temporal operators that navigate between worlds using this accessibility relation. Typical temporal operators used in LTL are:

- $\bigcirc p$: p is true in the next moment in time
- $\Box p$: p is true in all future moment
- $\Diamond p$: p is true in some future moment
- $p \cup q$: p is true until q is true

B. Self-Stabilization

The idea of self-stabilization in distributed computing was first proposed by Dijkstra in 1974 [13]. The concept of self-stabilization is that, regardless of its initial state, the system is guaranteed to converge to a legitimate state in a bounded amount of time by itself and without any outside intervention.

The self-stabilization principle applies to any system built on a significant number of components which are evolving independently from one another, but which are cooperating or competing to achieve common goals. This applies, in particular, to large distributed systems which tend to result from the integration of many subsystems and components developed separately at earlier times or by different people.

1) Formal Definition: Arora and Gouda [16] introduced a more generalized definition of self-stabilization, called stabilization, which is defined as follows:

The definition of stabilization for a system S with respect to two predicates P and Q, over its set of global states. Predicate Q denotes a restricted start condition. S satisfies $Q \longrightarrow P$(read as Q stabilizes to P) if it satisfies the following two properties:

i. Closure: P is closed under the execution of S. That is, once P is established in S, it cannot be falsified.

ii. Convergence: If S starts from any global state that satisfies Q, then S is guaranteed to reach a global state satisfying P within a finite number of state transitions.

The self-stabilization is a special case of stabilization where Q is always true, that is, if S is self-stabilizing with respect to P, then this may be restated as $TRUE \longrightarrow P$ in S.

2) Self-Stabilization Formalism: Based on the definition above, we propose a formalism of the self-stabilization. The advantage that this definition has among the other versions, is that it uses the predicate Q and P, thing that makes the use of LTL logic easier. To formalize the self stabilization using the LTL logic, we should do it for its two properties: closure and Convergence.

Let P and Q be a predicates of S, and $\sigma = [s_1 s_2 s_3 ....]$ is an execution of S, the $\models$, $\Box$ and $\Diamond$ are defined in section: semantic of LTL.

The semantic of the closure is provided by the definition: Once S is in a legitimate state P, it will stay in that legitimate state. Formally, we interpret it as follow:

- $s \models \Diamond\Box p$

For the convergence, it can be defined as: From any arbitrary state that satisfy Q, S is guaranteed to reach a configuration satisfying P, in a finite number of state transitions. This can be translated to LTL language as follow: $((\sigma, i) \models Q) \Longrightarrow (\Diamond((\sigma, k)_{i \leq k < \propto} \models P))$ witch means that: $\forall i, \exists j$ such that $i \leqslant j < \infty \ (s_i \models Q) \Longrightarrow \Diamond(s_j \models P)$

And we have
$S \models \phi \iff \sigma \models \phi(\forall \sigma \in Exc(S))$ and $\sigma \models \phi \iff s_i \models \phi(\forall s_i \in \sigma)$
Hence we can write
$(S \models Q) \implies \Diamond(S \models P)$
Taking in consideration that $(a \implies b) \iff (\neg a$ ou $b)$ we can formalise the convergence as follow:
$(S \not\models Q) \vee (\Diamond(S \models P))$

IV. Verification Algorithm

In order to test the validity of this both LTL formula, we should follow one of the verification methods. From the variety of methods that exist in literature, we chose the verification of model checking.

The model checking is a verification technique that explores all possible system states in a brute-force manner. It is interested by the determination of whether a property φ is verified by the system M as mention this figure:

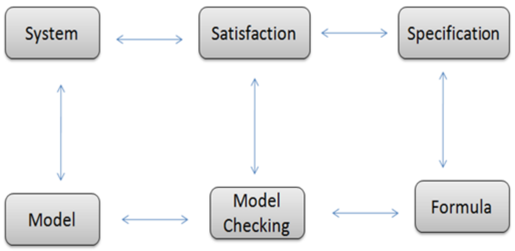


Fig. 2: Modelization with Model Checking

In this way, it can be a good method of verification of our self stabilization formalism, since it refers to the question of whether a formula is true in an interpretation, denoted $M \models \varphi$. Where a Kripke structure M can be a Petri net or a computer system, and the formula φ specifies our formalism witch is a property of system.

A. Model Checking

As principles, the model checking is an automated technique that, given a finite-state model of a system and a formal property, it checks systematically whether this property holds for that model.

A different phases can be distinguished In applying model checking Process:
Modeling phase: We model the system under consideration using the model description language of the model checker. Formalization phase: We formalize the property to be checked using the property specification language. Running phase: We run the model checker to check the validity of the property in the system model.

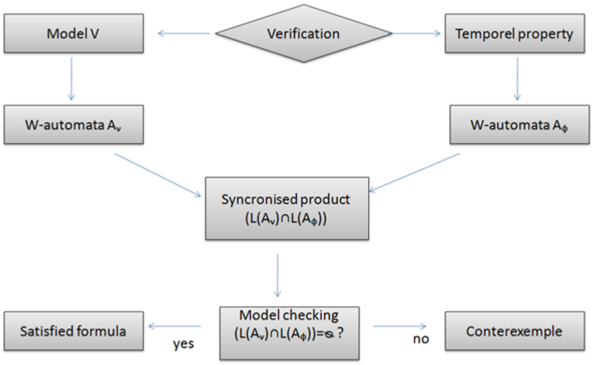


Fig. 3: The principle of model checking

In addition to these steps, the entire verification should be planned, administered, and well organized as shown in this algorithm:

- Input: KRIPKE stricture M, and LTL formula φ.
- Issue: If $M \models \varphi$
- Steps:
  1) Transformation of Kripke stricture M to Buchi automata A_M.
  2) Transformation of the LTL formula φ to Buchi automata A_φ.
  3) Test whether $L(A_M) \subseteq L(A_\varphi) \iff L(A_M) \cap L(A_\varphi)^c$

This can be formally written as: Input: finite transition system TS and LTL formula φ. Output: yes if TS$\models \varphi$; otherwise, no plus a counterexample. Transforming φ to GBA: Construct an NBA $A_{\neg\varphi}$ such that $\iota_\omega(A_{\neg\varphi}) = Words(\neg\varphi)$. Construct the product transition system $TS \otimes A$ if there exists a path π in $TS \otimes A$ satisfying the accepting condition of A then return no and an expressive prefix of π else return yes End if.

We presente the concept of model checking in somewhat in the diagram bellow: Overview of LTL model checking

B. Running Phase

In this subsection we will respect the model checking algorithm, and implement explicitly each step in more details. In the execution of model checking algorithm, the most difficult phase, is the second phase which is the transformation of the LTL formula φ to a Buchi automaton A_φ. For that reason, we will make such a big deal about this step of algorithm.

1) φ to BA Transformation: This stage of model checking algorithm is known of its difficulty. There are several ways to realize it[ref]. We are going to choose the method of Tables. Based on the following equivalences:
$\varphi \cup \varphi_1 \equiv \varphi_1 \vee (\varphi \wedge X(\varphi \cup \varphi_1))$.
$\varphi R \varphi - 1 \equiv (\varphi \wedge \varphi - 1) \vee (\varphi - 1 \wedge X(\varphi \wedge \varphi - 1))$.

And considering a Z-shaped set of negative normal formulas, Z is reduced if:
For all $z \in Z$, z is of the form p, $\neg p$ or $X(z')$.
We obtain the follow reduction of temporal connectors:

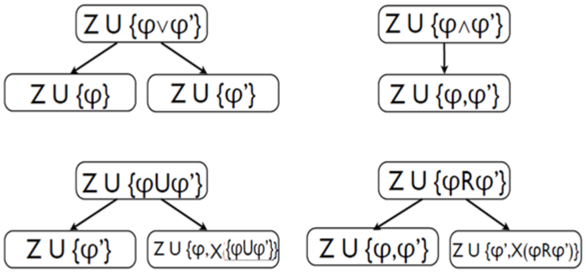


Fig. 4: Reduction of temporal conectors

This method involves, the reduction of the LTL formula φ to the normale negative form, the reduction of the temporal connectors, and finally the transformation to a Buchi automata.

From above, we obtain our Buchi automaton of closure and convergence properties:

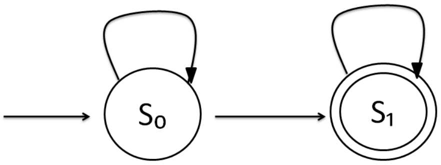


Fig. 5: Buchi automata for closure

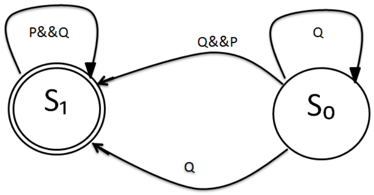


Fig. 6: Buchi automata for convergence

2) Kripk Stricture to BA Transformation : From the Kripke structure of our model, we can is obtain the follow Buchi automata that it correspond:

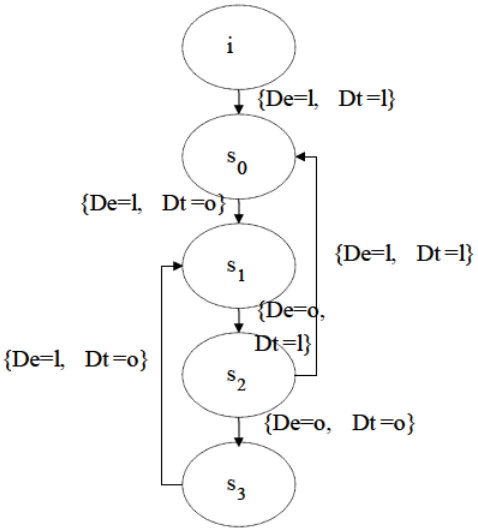


Fig. 7: Buchi automata for convergence

This automaton is obtained by performing three transformations:

1) Introduction of the initial condition i connected by a transition to the initial state of the Kripke structure.
2) Replacing Labels transitions by the decor of their target state.
3) Make all statements as acceptance states.

3) Buchi Product Checking : The model-checking is reduced to a problem in automata theory, since finite-state reactive programs can be represented quite naturally as Buchi automata [17]. A Buchi automaton is a non-deterministic finite-state automaton taking infinite words as input. A word is accepted if the automaton goes through some designated good states infinitely often while reading it. In this level, we are going to make the synchronized product of both Buchi automata, formula's automata and model's automata. The product of model automata and formula's automata is given by : .

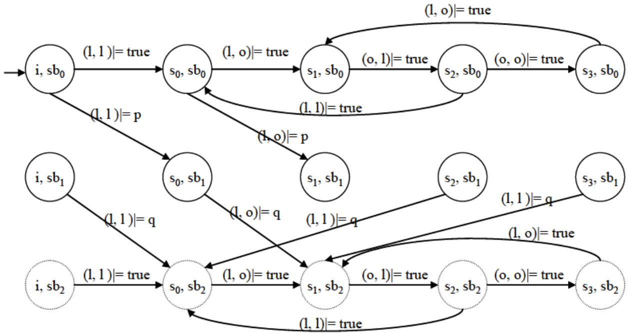


Fig. 8: Product of Buchi automata for closure

With the description of language of infinite strings with an automaton having finitely many states, the infinite string must make a cycle though some of the states of the automaton. For the string to be accepted, it must also be possible to reach this cycle from the start state. We thus process the product automaton by first finding the set of states that are reachable from the start state, and then checking whether any of them is in a cycle. This can be made more efficient by checking for cycles as each state becomes reachable. If a reachable cycle is found, then there is some string in the intersection of the language of the model and the language of the negation of the desired formula. Thus, the desired formula is not valid in the model, and the found string is a counterexample. On the other hand, if there is no reachable state that is part of a cycle, then the language is empty and the original formula is valid in the model.

I our case, the existence of these acceptors cycles indicates that the property of closure is not satisfied, therefore our system is not self stabilized.

V. Conclusion

In summary, we model checked our embedded system to verify LTL self stabilization formalism requiring the following steps:

1) Model: view it as a Buchi automaton, with all states as final states.
2) Formula: Negate the formula of self stabilization and perform the following conversion steps to obtain Buchi automaton $A_{\neg\varphi}$
    - Convert the formula to normal form (essentially, push negations down to atoms).
    - Convert the normal form to a graph.
    - Convert the graph to a generalized Buchi automaton (essentially identify the final states and appropriate state labelling).
    - Convert the generalized Buchi automaton to a Buchi automaton (essentially convert the set of sets of final states to a single set of final states, by duplicating the automaton)
3) Compute the product of the two automata, which accepts the intersection of the two languages.
4) Determine whether the product automaton accepts any strings. we noticed that closure formulas is not satisfied in our system, then our robot is not self-stabilizing system.

References

[1] D. Estrin, R. Govindan, and J. Heideman, “Embedding the internet.” in *Communications of the ACM*, May 2000.

[2] Whittlesey-Harris, R. S., and M. Nesterenko, “Fault-tolerance verification of the fluids and combustion facility of the international space station,” in *Distributed Computing Systems*, 2006.

[3] M. Arumugam, , and Kulkarni, “Self-stabilizing deterministic tdma for sensor networks,” in *Distributed Computing and Internet Technology.*, 2005.

[4] D.Gabbay, A.Pnneli, S.Shelah, and J.Stavi, “The temporal analysys of fairness,” in *Princ. of Prog. Lang.*, 1980, pp. 163– 173.

[5] A. Pnuelii, “The temporal logic of programs,” in *The temporal logic .*, 1977, pp. 46–57.

[6] E. Clarke, O. Grumberg, and D. Peled, “Model checking,” *MIT Press*, 1999.

[7] R. Gerth, D. Peled, M. Vardi, , and P. Wolper, “Simple on-the-fly automatic verification of linear temporal logic,” in *Specification, Testing, and Verification.*, June 1995.

[8] F. Somenzio and R. Bloem, “Efficient bchi automata from ltl formula,” *Springer-Verlag*, 2000.

[9] R. Sherman, A. Pnueli, and D. Harel, “the interesting part of process logic,” *SIAM Journal on Computing*, 1984.

[10] J.-M. Couvreur, “On-the-fly verification of linear temporal logic,” in *Formal Methods in the Development of Computing Systems.*, 1999.

[11] Z. Manna and A. Pnueli, “The temporal logic of reactive and concurrent systems,” *Springer-Verlag*, 1992.

[12] E. M. M. Clarke, D. Peled, and O. Grumberg, “Model checking,” *MIT Press*, 1999.

[13] E. Dijkstra, “Self-stabilizing systems in spite of distributed control,” 1974.

[14] M. J. Fischer and H. Jiang, “Self-stabilizing leader election in networks of finite-state anonymous agents,” in *Principles of Distributed Systems.*, 2006.

[15] J. ABRIAL, “Formal approch for software devloppement (approches formelles pour l’aide au dveloppement de logiciels),” 1997.

[16] E. Cohen and S. Shenker, “eplication strategies in unstructured peer-to-peer networks,” *ACM SIGCOMM*, 2002.

[17] J. Buchi, “On a decision method in restricted second order arithmetic,” in *Logik Grundlag. Math*, 1960, p. 6692.

Modeling and simulation of the effects of landslide on circulation of transports on the mountain roads

Manh Hung Nguyen[1,2]
[1]Posts and Telecommunications Institute of Technology (PTIT)
[2]UMI UMMISCO 209 (IRD/UPMC), Hanoi, Vietnam

Tuong Vinh Ho[2,3]
[3]IFI, Vietnam National University in Hanoi
[2]UMI UMMISCO 209 (IRD/UPMC), Hanoi, Vietnam

Trong Khanh Nguyen[1]
[1]Posts and Telecommunications Institute of Technology (PTIT)
Hanoi, Vietnam

Minh Duc Do[4]
[4]Department of Geotechnics, Faculty of Geology
VNU University of Science,
Vietnam National University in Hanoi, Vietnam

***Abstract*—Landslides, as one of the major natural hazards, account each year for enormous property damage in terms of both direct and indirect costs. Mountain roads where probability of land sliding is the highest, causes hurdles not only in the traffic flow but generate various traffic problems in the form of congestion, high accidents rate and waste of time. This paper introduces an agent-based model for modeling and simulation of the effects of landslide on the circulation of transports on mountain roads. This model is applied to the National Road N°6 of Vietnam to visualize and analyze the effects of landslide on the road when it occurs. This model could help us to improve and optimally organize of landslide warning and rescue system on the road.**

***Keywords*—Simulation model, Traffic network simulation, Landslide effect, Multiagent system**

I. Introduction

Landslides, as one of the major natural hazards, account each year for enormous property damage in terms of both direct and indirect costs. Mountain roads where probability of land sliding is the highest, causes hurdles not only in the traffic flow but generate various traffic problems in the form of congestion, high accidents rate and waste of time. As the others extreme conditions, landslides are rare and high-impact events and, thus, are difficult to manage. The difficulties arise from the following:

- They are rare and, as a result, there is a lack of data to help analyzing the mechanisms of the emergencies and the responses of the real world systems.
- They are non-recurrent, so management strategies developed in response to these extreme conditions may hardly be used in practice, and thus it is difficult to evaluate their effectiveness.
- They are disruptive and destructive, so it might be prohibitive to physically replicate these conditions for learning and training purposes.

As a low-cost, safe, non-disruptive, reproducible, and testable means of problem solving, modeling and simulation is particularly attractive in addressing problems under extreme conditions like landslides. In this report, we only concentrate on the landslide impacts after they happen, but not how do they happen. Modeling and simulation of how the landslide occurs based on the underground structure, geology, rainfall and storms, etc. in the mountain region is thus out of the scope of this report.

Concretely, we are interesting in landslides impacts on circulation on mountain roads. We consider the modelling and simulation of the effects of landslide on circulation of transports on mountain roads as a kind of transportation network simulation problem. In which, the landslide is considered as a cause of obstacles on road that the transports have to avoid during their circulation. And the possibility of occurrence of landslide at a point on a road is considered as an input data for the model we propose. We concentrate on the effects of landslide after it happens, but not how does it happen. The modelling and simulation of how the landslide occurs based on the underground structure, geology, rainfall and storms, etc. in the mountain region is thus out of the scale of this paper.

Recently, there have been many researches interested in the field of transportation network simulation. Therefore, there have been many models and tools proposed. Most of them are agent-based models. In which, intelligent agent and multiagent system seem to be suitable for simulate transportation network at the micro level. Each transport is thus modelled as an intelligent agent. It could observe other transports and obstacles to change its own speed as well as direction to go to its destination as fast as possible. The transportation network therefore could be modelled as a multiagent system whose each agent has a personnel goal (its destination to go) and they have to coordinate and/or interact together in order to prevent accidents from happening. There are many models proposed in this tendency. For instance, AgentPolis [12] is a fully agent-based platform for modeling multi-modal transportation systems. It comprises a high-performance discrete-event simulation core, a cohesive set of high-level abstractions for building extensible agent-based models and a library of predefined components frequently used in transportation and mobility models. MATSim development team [6] is developing a framework and platform for a transportation network simulation, called MATSim. MATSim provides a toolbox to implement large-scale agent-based transport simulations. The toolbox consists of several modules which can be combined or used stand-alone: toolbox for demand-modeling, agent-based mobility-simulation (traffic flow simulation), re-planning, a controller to iteratively run simulations as well as methods to analyze the output generated by the modules. SUMO (Simulation of Urban MObility) [4], [13] is a highly portable, microscopic road traffic simulation package designed to handle large road networks. It is mainly developed by employees of the Institute of Transportation Systems at the German Aerospace Center. SUMO allows to simulate how a given traffic demand which consists of single vehicles moves through a given road network. The simulation allows to address a large set of traffic management topics. It is purely microscopic: each vehicle is modelled explicitly, has an own route, and moves individually through the network. The V2X simulation runtime infrastructure - VSimRTI [25] - enables the preparation and execution of V2X simulations. It is a flexible system which simulates traffic flow dynamically. VSimRTI couples different simulators, thus, allowing the simulation of the various aspects of future intelligent transportation systems. Al-Dmour [1] developed TarffSim, a Multiagent Traffic Simulation for micro-simulation and macro-simulation of traffic. TraffSim was implemented by using NetLogo. Balakrishna et al. [2] presented a simulation-based framework for the modeling of transportation network performance under emergency conditions. The system extends the well-established dynamic traffic assignment (DTA) framework. Barcel and Casas [3] discussed some of the most critical aspects of the dynamic simulation of road networks, namely the heuristic dynamic assignment, the implied route choice models, and the validation methodology, a key issue to determine the degree of validity and significance of the simulation results. Caris et al. [5] proposed a discrete event simulation methodology to understand the network dynamics and analyze policy measures with the intention of stimulating inter-modal barge transport. The simulation model allows to quantify a number of network properties resulting from the interaction of freight Frick [9] dealt with the flexible implementation of transportation networks into a simulation model and realizing scenarios. Gokulan and Srinivasan [10] have been implemented two different types of multi-agent architectures on a simulated complex urban traffic network in Singapore for adaptive intelligent signal control. Holmgren et al. [11] presented the Transportation And Production Agent-based Simulator (TAPAS), which is an agent-based model for simulation of transport chains.Lotzmann [14] presented an agent-based traffic simulation approach which sees agents as individual traffic participants moving in an artificial environment. Meignan et al. [16] presented a bus-network simulation tools which include these specificities and allows to analyze and evaluate a bus-network at diverse space and time scales. They adopted a multiagent approach to describe the global system operation as behaviors of numerous autonomous entities such as buses and travelers. Mengistu et al. [17] provided a framework for development and execution of parallel applications such as multi-agent based simulation (MABS) in large scale. In the work of Manning et al. [15], a mathematical model is given for the system failure and a statistical model is formulated for the joint distribution of rainfall at different points along the railway line. Perumalla [22] designed a simulator with memory and speed efficiency as the goals from the outset, and, specifically, scalability via parallel execution. The design makes use of discrete event modeling techniques as well as parallel simulation methods. Piorkowski et al. [23] are developing TraNS, an open-source simulation environment, as a necessary tool for proper evaluation of newly developed protocols for Vehicular Ad Hoc Networks (VANETs). Ramos et al. [24] described the main features of contemporary ITS, emphasizing the Portuguese case, and describes the fundamental modeling & simulation tools that are considered critical to support the daily operation of the urban transportation system. Xu and Tan [26] introduced a hyper graph-based offline road network partitioning solution, which is suitable for future distributed transportation simulations with ITS applications. Zacharewicz et al. [27] dealt with the development of this simulation platform, based on Generalized Discrete Event Specification (G-DEVS) models and HLA (High Level Architecture) standard. Zhang et al. [28] presented an agent-based discrete-event simulation (AB-DES) modeling approach for transportation evacuation simulation based on a hybrid continuous and cell space. Zhang and Lv [29] presented a multi-agent competition model to describe the competition relationship in integrated transportation system.

In Vietnam, there are also some researches on modelling and simulation of urban transportation network. For instances, Nguyen and Ho [18] proposed an agent-based model for modelling and simulation the Vietnamese behavior in circulation in the city of Hanoi. Nguyen et al. [19], [20], [21] also proposed an agent-based model for simulation of urban transportation network.

Our objective thus is to develop a simulation model and tool to visualize the effects of landslide on circulation of transports on the mountain road. The model could be applied not only for the National Road N°6 of Vietnam but also be applied for any mountain road in Vietnam. This tool also aims to help the mountain road management office to improve and optimise the organisation of landslide rescue center along the roads to reduce the waiting time of transport as well as the resources needed (rescue machine, communication, energy) to clean and repair the roads once landslide occurs.

This paper is organised as follows: Section II presents our agent-based model of the system. Section III presents a case study in which we apply the proposed model to simulate the effects of landslide on the National Road N°6. Finally, section IV discusses the presented work and draws some perspectives for future work.

II. Modelling of the system

This section presents the modelling of the system. The system is based on multi-agent system. We thus model find kinds of agents (Fig. 1):

- Road agent: The agent represents the road
- Landslide point agent: The agent represents the points where the landslide occurs.
- Transport agent: The agent represents the transport on the road
- Rescue center agent: The agent represents the emergency landslide rescue center.

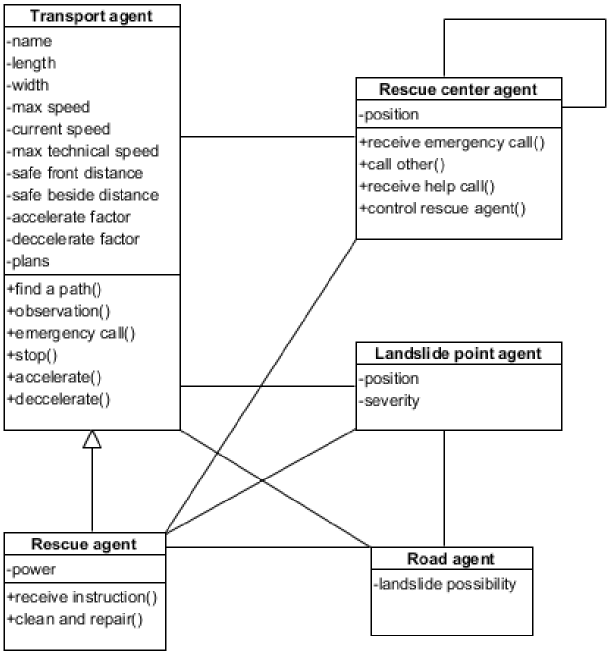


Fig. 1: Modelling of agents in the system

- Rescue agent: The agent represents the emergency rescue machine to clean and repair the road where the landslide occurs.

These agents will be described in detail in the following sections.

Note that in this model, we do not model an object in the transportation network, that is *traffic light*. The reason is that the traffic light has important effects in the circulation of transport in urban city circulation simulation scale: the average circulation time of transport, the average waiting time at intersections of transport could be affected by the change of traffic light operation policy when we consider and simulate the system in a small scale as at an intersection (see Nguyen et al. [20]), or in a city (see Nguyen et al. [18], [19]). However, when we consider the system in a large scale as a whole along a road on the mountain, the length of road could reach hundreds kilometres, the waiting time at intersections caused by traffic light is so small, in regarding the whole time to pass the road, that we do not need to model the traffic light in this model.

A. Road agent

Road agent represents the real road. The real data is got from GIS file (Geographical Information System). So this kind of agent has all geometry attributes: position, length, etc. Each agent represents a short segment of the road.

This kind of agent has no behavior.

The most important attribute of this agent is its position because this attribute is related to the potential landslide at that position. So we need to model an additional attribute for this agent, that is:

- *landslide possibility*: the possibility to occur the landslide at that segment of road.

B. Landslide point agent

landslide point agent represents the points at that the landslide occurs. This agent has following attributes:

- *position*: the position where the landslide occurs.
- *severity*: the level of severity of landslide, or the amount of stone slides on the road. The higher the severity is, the longer time the rescue agent need to rescue the road to get normal circulation status.

This agent has no behavior. It will be died when the rescue agents finish their work at that point.

C. Transport agent

A transport could be a trunk, a bus, a car (including taxi), a motor. It has some attributes, behaviors and ability to move. So we need to model it as an agent in the system. Because a driver is assigned to his transport so we consider the whole of a driver and his transport as an unique transport agent.

1) Attributes: A transport has these attributes:

- *name*: name of a transport.
- *length (denoted as l)*: real length of a transport.
- *width (denoted as d)*: real width of the transport.
- *max speed (denoted as v_{max})*: the maximal allowed speed, for the transport, by law
- *current speed (denoted as v)*: the current speed of the transport
- *max technical speed (denoted as v_{tech})*: the maximal technical speed of a transport, limited by the engine of the transport.
- *safe front distance (denoted as d_f)*: the minimum distance to the nearest transport in front that keep safe for circulation. This distance is estimated by following formula:
  $$d_f = \frac{l * v}{v_{max}} + \theta \tag{1}$$
  where θ is the minimum distance allowed among transports when stopping.
- *safe beside distance (denoted as d_b)*: the minimum distance to the nearest beside transport that keep safe for circulation. This distance is estimated by following formula:
  $$d_b = \frac{d * v}{v_{max}} + \theta \tag{2}$$
- *accelerate factor (denoted as α)*: the ability to speed up in an unit of time
- *decelerate factor (denoted as β)*: the ability to slow down in an unit of time
- A set of *circulation plans*. A plan contains following information:

  - start time: the start time to circulate
  - department: the start position of the circulation. A position is represented by its real (longitude (x), latitude (y)).
  - destination: the destination(s) to get to of the individual
  - max speed: the maximal speed for the individual. This must not be higher than the maximal permitted speed by law.
  - type of vehicle: this attribute is reserved to determine the size of vehicle on the road.

2) Behavior: Behavior of transport agents:

- *find a path*: this will find a path to go. A transport agent finds a path when: either it starts a new plan; or it want to change the path when it is blocked somewhere on the way. A path is simply determined by the Dijkstra's algorithm (Dijkstra [7]) on a graph constructed from the road as follow:
  - Each intersection forms a node of the graph.
  - Each road forms an arc with the same direction. If a long road has many segment points to change the direction, then each segment of road (between two consecutive segment points) will form an arc with the same direction.
  - The weight of each arc is proportional to the length of the corresponding segment of road.
- *observation*: this is an action of the driver of a transport, including of observation of the landslide, observation of obstacles.
- *emergency calling*: the *transport agent* faces a landslide, it will call to one of *rescue center agents* to alert the occurrence of landslide.
- *stop*: a transport agent stops at a blocked point caused by landslide.
- *accelerate*: a transport agent will accelerate when: (1) there is no obstacles in *safe front distance* and *safe beside distance* of it; and (2) its speed does not reach the *max speed* yet. The new speed will accelerate to:

$$v_{t+1} = min\{v_t + \alpha, v_{max}\} \tag{3}$$

  where v_t, v_{t+1} are the speed of the transport agent at the simulation step $t, t+1$, respectively.
- *decelerate*: a transport agent has to decelerate when: (1) there is some obstacles in *safe front distance* or *safe beside distance* of it; or (2) it faces a *rescue agent*; or (3) it intends to stop. The new speed will decelerate to:

$$v_{t+1} = max\{v_t - \beta, 0\} \tag{4}$$

The transition among behaviors of a transport agent is depicted in the Figure 2: A transport agent starts his movement by starting a plan on time. Firstly, it finds a path to go. During moving, it observes three kinds of objects: the landslide, the destination, and the obstacle. In observing the landslide: if there is some landslide occurs, it will *stop*, take an emergency call to one of *rescue center agent*, and may be re-find the path; otherwise, it will continue to move. In observing the obstacle, it will *accelerate* if there is no obstacle and its current speed v is still lower than the allowed speed v_{max}. It will *decelerate* if there are some obstacles or its speed v is already higher than the allowed speed v_{max} (for the case of obey transport). Otherwise, it continues to move with the current speed. Note that, the circulation priority of *transport agent* is lower than that of *rescue agent*, so it have to *decelerate* and release the road for *rescue agent*. In observing the destination, if it is at the destination, it finishes the circulation. Otherwise, it continues to move.

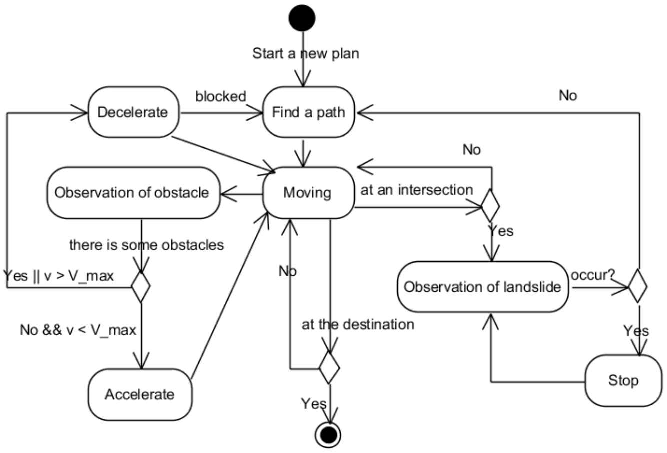


Fig. 2: Behaviors of transport agent

D. Rescue center agent

Rescue center agent represents the emergency landslide rescue center. This agent plays the role of managing a set of *rescue agents*. They are also able to communicate with each others to optimally control *rescue agents* if it is necessary.

1) Attribute:

- *position*: the position where it is situated.

2) Behavior:

- *receiving of emergency calls*: this agent is responsible for receiving the emergency call from *transport agents* to be aware the occurrence and the position of landslide.
- *call other*: In the case of there is not enough *rescue agents*, or it is not optimal if using its own *rescue agents*, then the *rescue center agent* could call other *rescue center agents* to ask for help.
- *receiving of helping calls*: this agent is also able to receive the helping call from other *rescue center agents*.
- *control rescue agent*: after receiving emergency call from *transport agent* or helping call from other *rescue center agents*, it could control some of its *rescue agents* to go to the point where the landslide occurs.

The transition among behaviors of a rescue center agent is depicted in the Figure 3: It is always in the status of waiting. When it receives a call (emergency call from *transport agent*,

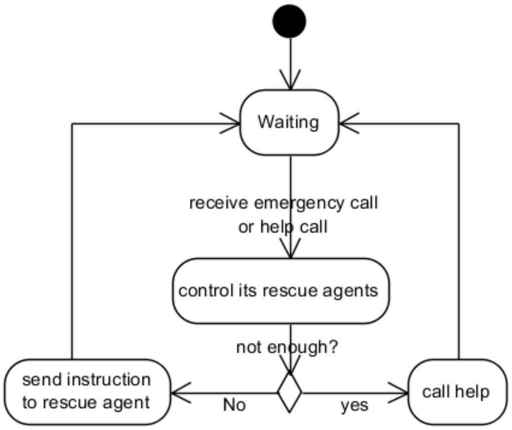


Fig. 3: Behaviors of rescue center agent

or helping call from other *rescue center agent*), it will start the control process by considering the number of *rescue agent* that it owns. If the number is not enough for rescue propose, it will *helping call* to other for requirement of some help. Otherwise, it sends instruction to selected *rescue agents* to put them into mission. After all, it returns to the permanent status of waiting for the next call.

E. Rescue agent

Rescue agent represents the emergency rescue machine when the center of landslide rescue receives the landslide occurrence news. This agent is basically the same with transport agents on the attributes and the behaviors to move. Moreover, this has some more attributes and behaviors.

1) Additional attribute:

- *power*: the ability to repair and clean the road after landslide. The more powerful it is, the faster the road is rescued.

2) Additional behavior:

- *instruction receive*: rescue agent is under the control of the rescue center agent. So it is able to receive and treat instructions from rescue center agent. The instruction may be one of these kinds: start to move toward the landslide position, start to repair and clean, pause (repair and clean), return to a rescue center.
- *repair and clean*: when arriving the point of landslide, it cleans the rolled stones and repairs the road. When it finishes its work, the road is through and the rescue agent returns to the center of rescue.

The transition among behaviors of a rescue agent is depicted in the Figure 4: The *rescue agent* is always in the status of waiting at one of *rescue center agents*. When it receives an instruction from the *rescue center agent*, it will start to move to the reported *landslide point agent*. During this movement, it acts as any kind of *transport agent* with the highest priority of circulation on the road. When it arrives at the *landslide point agent*, it starts to clean and repair the effected *road agent*. The time to clean and repair depends on the relation between the *severity* of *landslide point agent* and the total of *power* of all *rescue agents* in the mission. After finishing the cleaning and repairing the effected *road agent*, the *rescue agent* returns to its original *rescue center agent*. During this movement, it also acts as any kind of *transport agent* with normal priority of circulation on the road. When it arrives at the *rescue center agent*, it returns to the permanent status of waiting for the next instruction.

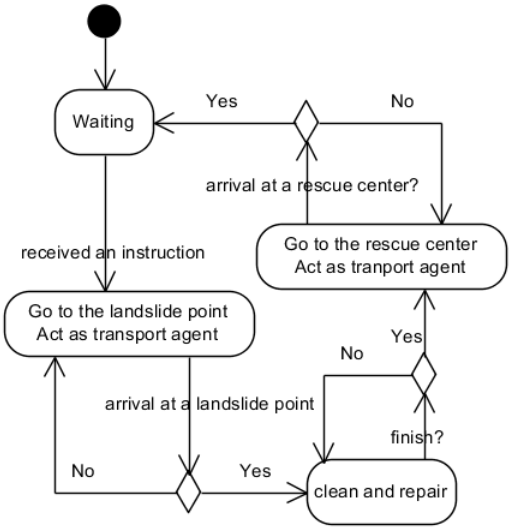


Fig. 4: Behaviors of rescue agent

F. A standard scenario

This is a scenario of the simulation:

1 All *transport agents* are normally moving on the road to their target.

2 There is some *landslide point agent* occurs at some points on the road

3 The *road agent* is blocked at that point.

4 The first *transport agent* faces the *landslide point agent* calls one of *rescue center agents*. It is also blocked at that point.

5 The called *rescue center agent* calls some *rescue agent* to go to the blocked point. In some variant case, the *rescue center agent* could call other *rescue center agents* to have more *rescue agents* if it is not enough or for the reason of optimisation of waiting time.

6 The selected *rescue agent* goes to the blocked point after *receiving the instruction* from *rescue center agent*.

7 When the *rescue agent* arrives at the blocked point, it stats to clean the stones from the road and then repair the road. The time taken for this operation depends on the *severity* of *landslide agent*, the *power* and the number of *rescue agent*.

8 During the time of cleaning and repairing of *rescue agent*, all *transports agents* passing the blocked point

are still blocked, therefore the *waiting time* of each one continuously increases.

9 When the blocked point is cleaned, the *landslide agent* dies, and the blocked *transport agents* continue to go to their target (their *waiting time* stops counting)

10 The *rescue agent* returns to its original *rescue center agent*.

We would like to display this scenario on the visualization level of the simulation.

III. A case study: Simulation of the landslide's effect on circulation of the National Road N°6

This section applies the proposed model to simulate the effects of landslide on the circulation on the National Road N°6 of Vietnam (in the North-West of Vietnam). Section III-A presents the modelling and simulation of the system. Section III-B presents the obtained simulation results.

A. Simulation setup

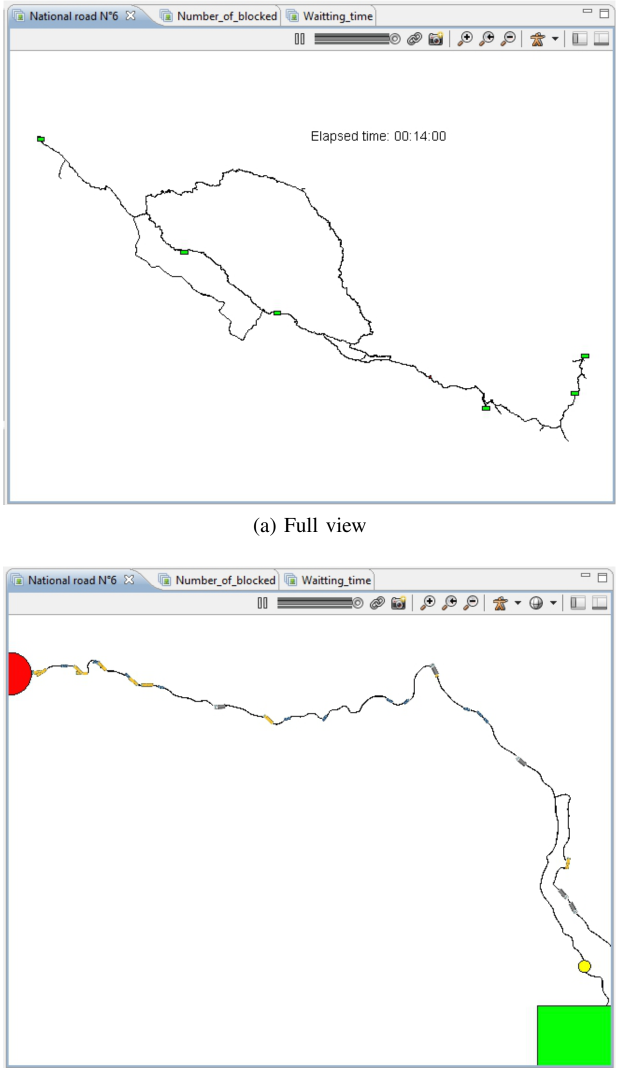
1) Input data: The input data for the simulation is set as follow:

- *Roads map*: The National Road N°6 is situated in the North-West of Vietnam. We extracted its data as a GIS file (Fig.5.a,b). It is considered from Hoa Binh city (attitude: 20.789N, longitude: 105.342E) to Son La city (attitude: 21.325N, longitude: 103.915E).
- *Landslide possibility along the road*: This data is provided by The Department of Geology, The Hanoi National University. In which, there are 41 potential point of landslide occurring, with different level, along the National Road N°6 (Fig.5.c). We normalise the occurring potential of each point by a number in the interval $[0, 1]$: the more the value is high, the more the possibility of landslide occurs at that point is high.
- *Transport distribution and plans*: We refer this data from the statistic data of the National Center of Statistic taken in 2009. On basing on this data, we estimate the number of permanent transport is about 500-1000 on the whole road at any daily moment, the average speed is about 25-50km/h.
- *Landslide rescue centers*: We estimate that each big town of each city along the whole National Road N°6 has at least a rescue center, and each center has at least a rescue machine (Fig.5.d).

2) Analysis and evaluation criteria: At the output of simulation, we need to calculate the following parameters:

- *Number of transports blocked*: The number, in total, of transports which are blocked at the point where the landslide occurs.
- *Total waiting time of each blocked transport*: The total waiting time that each transport wasted in being blocked at the point where the landslide occurs.

These parameters help us to compare the effect of different rescue solutions when the landslide occurs. By reducing the value of these parameters, we could improve and optimise the organisation of landslide rescue system.



(a) Full view

(b) A zoomed view

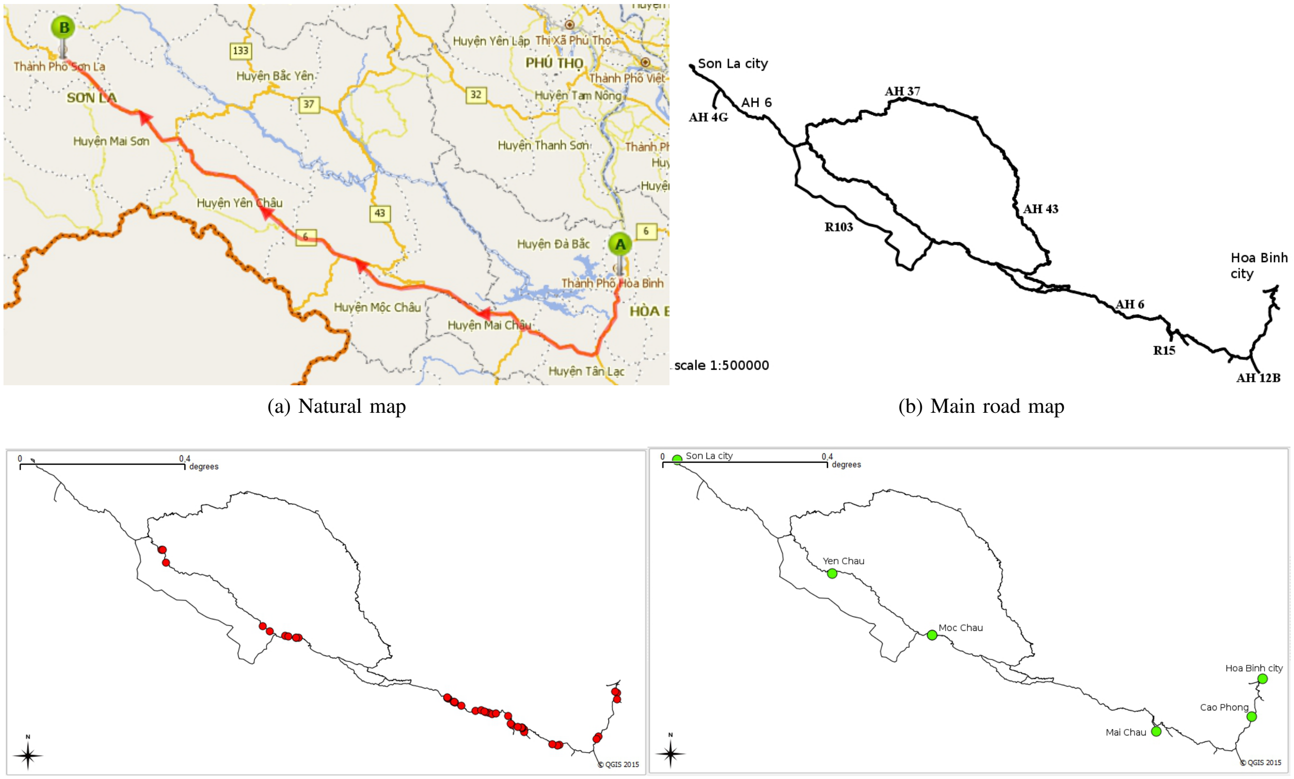
Fig. 6: Visual results of the simulation

3) Simulation platform: Our simulation of the effects of landslide on circulation on the National Road N°6 of Vietnam is implemented in the simulation platform GAMA [8]. GAMA is integrated and generic tools to support the representation of features usually associated with real complex systems, namely rich, dynamic and realistic environments or multiple levels of agency. It allows modelers, thanks to the use of a high-level modeling language, to build, couple and reuse complex models combining various agent architectures, environment representations and levels of abstraction.

B. Results

In this section, we presents the simulation results at two levels: visualisation level, and statistic analysis level.

1) Visualisation of the system: At the visualisation level, the global view of the system is depicted in Fig.6.a. Because



(a) Natural map

(b) Main road map

(c) Landslide points map

(d) Landslide rescue centers map

Fig. 5: Input GIS data for simulation

the big scale of the system, it is not easy to see all kind of agent in the system except we have to zoom it in detail. Fig.6.b shows in detail all kinds of agent in the system.

2) Statistical analysis: At the statistical analysis level, the simulation also displays the variance of output parameters by time: number of blocked transport, and total waiting time of blocked transport.

Fig.7.a depicts the number of blocked transport during simulation time. As we start to simulate the system at the moment when the landslide occurs, the number of blocked transport is regularly increased along the simulation time until the moment when the *Rescue agent* finishes its work and the road is cleaned and repaired. At that moment, the number of blocked transport is immediately down to zero.

In the same principle for the total waiting time of blocked transport, as depicted in Fig.7.b: The value of this output parameter is also increased along the simulation time until the moment when the *Rescue agent* finishes its work and the road is cleaned and repaired. From that moment, the total waiting time of blocked transport is no change because there is no more any blocked transport in the system.

By analysing the output parameters, we could compare the effect of different rescue solutions when the landslide occurs. The more the solution brings a lower value of these parameters, the more the solution is better.

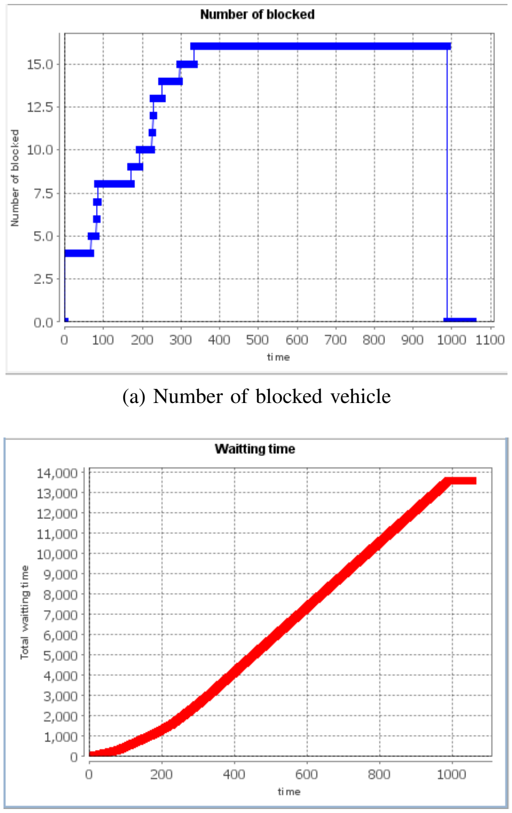
3) Experiment 1: The effects of rescue agent power: The object of this experiment is to evaluate the importance of power of rescue agent on the number of blocked point and waiting time. We suppose that the higher the power of rescue agent, the shorter waiting time and the fewer number of blocked points we have. By default, the minimum portion of the power of a rescue agent in regard with the severity of landslide is 0.125% (The power is 1.5, the severity is 1200). In order to evaluate this hypothesis, five scenarios have been performed, in which the severity is fixed while the power is changed from 1.5 to 3.5.

TABLE I: The effects of rescue agent power on output parameters

Scenario	**S1**	**S2**	**S3**	**S4**	**S5**
Avg. Power	1.5	2.0	2.5	3.0	3.5
Number of rescue agent	7/7	7/7	7/7	7/7	7/7
Blocked vehicle	42	32	26	21	15
Total blocked time (m)	2687	1830	1418	995	886
Avg. blocked time (m/vehicle)	64	57	55	47	59

The results of this experiment are depicted as in Tab.I. The number of blocked transport and also the waiting time decrease regularly when the power of rescue agent increases.

4) Experiment 2: The effects of vehicle kinds: In a transportation problem, the distribution of different kind of vehicle cans influent to the status of overall system. In this experiment, the ratio of distribution of car/bus/truck has been changed to evaluate how it affects to the number of blocked transports and waiting time. Three scenarios have been performed, in which the ration is changed.



(a) Number of blocked vehicle

(b) Total blocked time

Fig. 7: Statistic results of the simulation

TABLE II: The effects of vehicle kind ratio on output parameters

Scenario	S1	S2	S3
Car/bus/truck ratio (%)	50/30/20	30/20/50	20/50/30
Blocked vehicle	28	30	38
Total blocked time (m)	1557	3000	2865
Avg. blocked time (m/vehicle)	56	100	75

The Tab.II illustrates the results of this experiment about the affect of different kind of vehicles. Three scenarios have been hold. For the first one, the number of blocked transport is less than the two other cases, while the waiting time is less in compare with the rest. This is explained by the size of each vehicle; in our experiment, the size of bus and truck is equal, and greater 60% with the size of car. Thus, the more number of bus and truck is, the higher waiting time is.

5) Experiment 3: The effects of the number of landslide points: In these experiments, we change only the number of landslides, and their position, the other parameter is fixed. The number of landslide varies from 1 to 6.

TABLE III: The effects of the number of landslide points on output parameters

Scenario	S1	S2	S3	S4	S5	S6
Number of landslide	1	2	3	4	5	6
Blocked vehicle	8	18	30	60	62	73
Total blocked time (m)	697	1250	2732	5033	9102	12718
Avg. blocked time (m/vehicle)	87	69	91	84	147	174

TABLE IV: The effects of the number of landslide position on output parameters

Scenario	S1	S2	S3	S4	S5	S6	S7
Landslide in group	1	2	3	4	5	6	7
Blocked vehicle	14	28	30	69	19	33	20
Total blocked time (m)	1437	1901	3300	12499	719	2692	1522
Avg. blocked time (m/vehicle)	103	68	110	181	38	82	76

The results of experiment are shown in the Tab.III. As we start to simulate the system at the moment when the landslide occurs, the number of blocked transport has is increased when the number of landslide increases. The increasing amount between two consecutive experiments is quite significant.

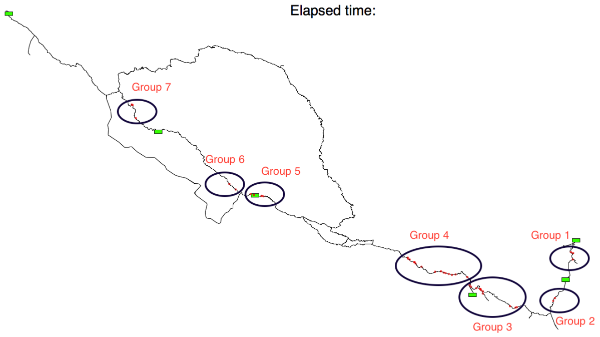


Fig. 8: Regrouping landslide points into seven groups

6) Experiment 4: The effects of landslide position: Actually, we have 41 potential landslides along the National Road N°6. With different landslide positions, are we have different affect on the waiting time and number of blocked transport? In this experiment we will examine it. In order to separate the position of landslide, we regroup 41 landslide points into seven groups as depicted in Fig.8. And then, we examine seven scenarios: each scenario has a landslide in each group.

The results of this experiment are depicted in Tab.IV. Based on it, we can divide 4 groups. The first one contains the position of group 1, 5 which have the lowest waiting time and number of blocked transports; Next, it is the group 6, 7 in which the number of blocked transports and waiting time is greater that the first one; The third one is at the position of group 2,3; The last one is the position at group 4 where the number of blocked points and the waiting time are the worst.

At the position of group 4, there is only a route; no detour to avoid this one, so that when landslide occurs here, the number of blocked transports and the waiting time are the highest. The landslides at the group 1 are situated between two rescue centers; while the group 5 is very closed to a rescue center. Moreover, the position at the group 1 is near by the end point of the National Road N°6, thus the quantity of vehicles is small. So that, when the landslide occur, the rescue agent can resolve it fast.

IV. Conclusion

This paper proposed an agent-based model to model and simulate the effects of landslide on the circulation of transports on the mountain roads. In which we modeled five kinds of agent: *Road agent, Landslide point agent, Transport agent, Rescue center agent, Rescue agent*, and then simulated their behavior. The model is applied to simulate the effects of landslide on the circulation of transports on the National Road N°6 of Vietnam. This is an useful tool for visualization the effects of landslide on the road.

Simulating to optimize some routing strategies when the landslide occurs or optimise the organisation of landslide rescue center network on the road are our works in the near future.

Acknowledgment

This work is funded by the SRV-10/0026 Programme: *Capacity Building and Technology Transfer for Mitigation of Geo-hazards in Vietnam in the Context of Climate Change*.

References

[1] Nidal Abid Al-Hamid Al-Dmour. TarffSim: Multiagent traffic simulation. *European Journal of Scientific Research*, 53(4):570–575, 2011.

[2] Ramachandran Balakrishna, Yang Wen, Moshe Ben-Akiva, and Constantinos Antoniou. Simulation-based framework for transportation network management in emergencies. *Transportation Research Record*, (2041):80–88, 2008.

[3] Jaime Barceló and Jordi Casas. Dynamic Network Simulation with AIMSUN. In Ryuichi Kitamura and Maso Kuwahara, editors, *Simulation Approaches in Transportation Analysis*, volume 31 of *Operations Research/Computer Science Interfaces Series*, pages 57–98. Springer US, 2005.

[4] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz. SUMO - simulation of urban mobility: An overview. In *SIMUL 2011, The Third International Conference on Advances in System Simulation*, pages 63–68, Barcelona, Spain, October 2011.

[5] A. Caris, G. K. Janssens, and C. Macharis. *Modelling complex Intermodal Freight flows*, pages 291–300. Springer, Berlin/Heidelberg, 2009.

[6] MATSim development team (ed.). MATSIM-T: Aims,approach and implementation. Technical report, IVT, ETH Zrich, Zrich, 2007.

[7] Edsger. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[8] Alexis Drogoul, Edouard Amouroux, Philippe Caillou, Benoit Gaudou, Arnaud Grignard, Nicolas Marilleau, Patrick Taillandier, Maroussia Vavasseur, Duc An Vo, and Jean-Daniel Zucker. GAMA: A Spatially Explicit, Multi-level, Agent-Based Modeling and Simulation Platform. In Yves Demazeau, Toru Ishida, Juan M. Corchado, and Javier Bajo, editors, *International Conference on Practical Applications of Agents and Multiagent Systems (PAAMS), Salamanca, Spain, 22/05/2013-24/05/2013*, volume 7879 of *Lecture Notes in Computer Science*, pages 271–274, http://www.springerlink.com, 2013. Springer.

[9] Rainer Frick. Simulation of transportation networks. In *Proceedings of the 2011 Summer Computer Simulation Conference*, SCSC '11, pages 188–193, Vista, CA, 2011. Society for Modeling & Simulation International.

[10] Balaji Parasumanna Gokulan and Dipti Srinivasan. Multi-agent system in urban traffic signal control. *IEEE Comp. Int. Mag.*, 5(4):43–51, 2010.

[11] Johan Holmgren, Paul Davidsson, Jan A. Persson, and Linda Ramstedt. Tapas: A multi-agent-based model for simulation of transport chains. *Simulation Modelling Practice and Theory*, 23:1–18, 2012.

[12] Michal Jakob, Zbyněk Moler, Antonín Komenda, Zhengyu Yin, Albert Xin Jiang, Matthew P. Johnson, Michal Pěchouček, and Milind Tambe. Agentpolis: towards a platform for fully agent-based modeling of multi-modal transportation (demonstration). In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, AAMAS '12, pages 1501–1502, Richland, SC, 2012. International Foundation for Autonomous Agents and Multiagent Systems.

[13] Daniel Krajzewicz. Traffic Simulation with SUMO Simulation of Urban Mobility Fundamentals of Traffic Simulation. volume 145 of *International Series in Operations Research & Management Science*, chapter 7, pages 269–293. Springer New York, New York, NY, 2010.

[14] Ulf Lotzmann. *TRASS: A Multi-Purpose Agent-Based Simulation Framework for Complex Traffic Simulation Applications*, pages 79–107. 2009.

[15] L. J. Manning, J. W. Hall, C. G. Kilsby, S. Glendinning, and M. G. Anderson. Spatial analysis of the reliability of transport networks subject to rainfall-induced landslides. *Hydrological Processes*, 22(17):3349–3360, 2008.

[16] David Meignan, Olivier Simonin, and Abderrafiaa Koukam. Simulation and evaluation of urban bus-networks using a multiagent approach. *Simulation Modelling Practice and Theory*, 15(6):659–671, July 2007.

[17] Dawit Mengistu, Peter Tröger, Lars Lundberg, and Paul Davidsson. Scalability in distributed multi-agent based simulations: The jade case. In *Proceedings of the 2008 Second International Conference on Future Generation Communication and Networking Symposia - Volume 05*, FGCNS '08, pages 93–99, Washington, DC, USA, 2008. IEEE Computer Society.

[18] Manh Hung Nguyen and Tuong-Vinh Ho. Modelling circulation behaviour of vietnamese: Applying for simulation of hanoi traffic network. In Phan Cong Vinh, Vangalur S. Alagar, Emil Vassev, and Ashish Khare, editors, *ICCASA*, volume 128 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 24–34. Springer, 2014.

[19] Manh Hung Nguyen, Tuong-Vinh Ho, Manh Son Nguyen, Thi Hoai Phuong Phan, Thi Ha Phan, and Van Anh Trinh. An agent-based model for simulation of traffic network status. In Lam Thu Bui, Yew-Soon Ong, Nguyen Xuan Hoai, Hisao Ishibuchi, and Ponnuthurai Nagaratnam Suganthan, editors, *SEAL*, volume 7673 of *Lecture Notes in Computer Science*, pages 218–227. Springer, 2012.

[20] Manh Hung Nguyen, Tuong Vinh Ho, and Tan Hiep Nguyen. On the dynamic optimization of traffic lights. *Asian Simulation and Modeling, Mahidol University*, pages 35–43, 2013.

[21] Manh Hung Nguyen, Manh Son Nguyen, Thi Ha Phan, and Van Anh Trinh. Dynamic path optimization in traffic routing. *Asian Simulation and Modeling, Mahidol University*, pages 43–51, 2013.

[22] Kalyan S. Perumalla. A systems approach to scalable transportation network modeling. In *Proceedings of the 38th conference on Winter simulation*, WSC '06, pages 1500–1507. Winter Simulation Conference, 2006.

[23] M. Piórkowski, M. Raya, A. Lezama Lugo, P. Papadimitratos, M. Grossglauser, and J.-P. Hubaux. TraNS: realistic joint traffic and network simulator for VANETs. *SIGMOBILE Mob. Comput. Commun. Rev.*, 12(1):31–33, January 2008.

[24] Ana Lusa Ramos, Jos Vasconcelos Ferreira, and Jaume Barcel. Modeling & simulation for intelligent transportation systems. *International Journal of Modeling and Optimization*, 2(3):274–279.

[25] Björn Schünemann. V2X simulation runtime infrastructure VSimRTI: An assessment tool to design smart traffic management systems. *Comput. Netw.*, 55(14):3189–3198, October 2011.

[26] Yan Xu and Gary Tan. An offline road network partitioning solution in distributed transportation simulation. In *Proceedings of the 2012 IEEE/ACM 16th International Symposium on Distributed Simulation and Real Time Applications*, DS-RT '12, pages 210–217, Washington, DC, USA, 2012. IEEE Computer Society.

[27] Gregory Zacharewicz, Jean-Christophe Deschamps, and Julien Francois. Distributed simulation platform to design advanced rfid based freight transportation systems. *Comput. Ind.*, 62(6):597–612, August 2011.

[28] Bo Zhang, Wai Kin (Victor) Chan, and Satish V. Ukkusuri. Agent-based discrete-event hybrid space modeling approach for transportation evacuation simulation. In *Proceedings of the Winter Simulation Conference*, WSC '11, pages 199–209. Winter Simulation Conference, 2011.

[29] Jiashun Zhang and Rongjie Lv. Multiagent competition simulation of integrated transportation system. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 11(1):345–350, 2013.

SmartFit: A Step Count Based Mobile Application for Engagement in Physical Activities

Atifa Sarwar*, Hamid Mukhtar*, Maajid Maqbool* and Djamel Belaid†
*National University of Sciences and Technology (NUST),
Islamabad, 44000, Pakistan
†Institut Mines-Telecom; Telecom SudParis;
CNRS UMR SAMOVAR 9 rue Charles Fourier,
91011 Evry, France

***Abstract*—Research has found that relatively few people engage in regular exercise or other physical activities. Despite the availability of numerous mobile applications and specialized devices for self-tracking, people mostly lack the motivation for performing physical activities. In this article we present SmartFit, a mobile application that uses step count for promoting the physical activities in adults. This article points out that while considering walk, activity duration is not sufficient for determining users activeness state. Step count is another factor that should be taken into account. For this we propose an approach for converting the steps into duration for which activity has been performed. This duration is then used in Smartfit for categorizing user into different activeness levels. Gamification techniques have been incorporated in SmartFit as they are found to serve the purpose of motivating and encouraging the user. Gamification is used for awarding/deducting points to user in order to keep them engaged for longer period. Furthermore feedback is also provided to users depending upon their goal and achieved progress. The objective is to facilitate and motivate the user and then keep them engaged in carrying out the recommended level of physical activities.**

***Keywords*—Feedback, Gamification, Physical Activities, Step Count**

I. Introduction

As much as the human body is complex, so is the issue of human health. Currently obesity due to sedentary life style is one of the major concerns of developing nations. Sedentary life style has negative impacts on health resulting in chronic diseases such as increased risk of weight gain, metabolic disruption and premature mortality [1]. One of the growing body of research suggests that taking short breaks during sedentary behavior and performing some physical activity has positive effects on health [2]. On the other hand, there have been numerous evidences that exercise is such a factor that contributes heavily to a better health and consequently to the quality of life as well.

In general, participating in physical activity is beneficial to people of all ages. Physical activity contributes to fitness, a state in which people's health characteristics and behaviors enhance the quality of their lives. When it comes to mental health, the benefits of exercise cannot be overemphasized. A clinical study conducted by Babyak et al. [3] has revealed that treating depression with exercise was just as effective as medication, and vice versa. According to center for Disease Control and Prevention (CDC)[1], regular physical activity helps improve the overall health and fitness, and reduces the risk for many chronic diseases [4].

In the current age of technology, smart phones due to their ubiquitous nature are commonly used as a medium for promoting physical activities. But people still rely less on gadgets, mobiles phones, and other specialized devices to track a health indicator. This is in contrast to the fact that recently there has been an increase in the healthcare-related research and unprecedented growth in the systems, devices, and applications for healthcare management, be it on a personal or a larger scale. So what is the reason that despite the availability of plethora of mobile phone applications and devices, half of the people still track their health status in their head? This is due to the fact that people mostly lack the motivation for performing the physical activities. Users use these applications and ultimately lose their interest after using it for a few times only. Thus creating the eagerness of living a healthy life by engaging users, a goal that is more challenging and demands more effort.

Having outlined the problem and our motivation for solving the problem, we now describe our goals as following. We aim to use gamification in the health systems and study its impact on the user's motivation and long term engagement. For this, we propose a mobile application, SmartFit that promotes physical activities by counting user's steps and converting these steps into activity duration. To make Smartfit engaging and a source of motivation for users, gamification is incorporated in it. Gamification is defined as use of gaming elements in non-gaming context [5]. McGonigal [6] in her book "Reality is Broken" writes that games have the power to satisfy human needs as well as solve real world problems. Franck in [7] suggests that gamification is one of the feature that could possibly change user's behavior and helps to develop habit as it serves the purpose of motivating the users and keep them engaged. Futhermore [8] states that games are not only used for entertainment but it also serves as a source of inspiration, persuasion and engagement. Thus in SmartFit, we have defined a scoring system that awards/deducts user's points depending on his/her progress. Points are awarded to users if their performance is up to the mark while points are deducted if they are lagging behind. So the user enjoys using the system rather then being compelled to use it unwillingly.

Unlike existing systems, which are rigid in setting the goals

[1] http://www.cdc.gov/

and fails to provide continuous motivation to the users, our proposed system is flexible and with the use of gamification techniques it provides continuous motivation to users. This help users not only in achieving the activeness goal once but also maintaining activeness afterwards.

The rest of this article is organized as following. First we briefly review existing work in Section II. Section III presents our proposed solution for converting the step count into activity duration. We describe our application, SmartFit, and present some of its key features in section IV. In section V we discuss the scoring system developed in SmartFit for user motivation. After that we show results of our proposed approach in section VI and finally conclude this article in section VII.

II. Related Work

Recently, there has been much focus on auto-analytics for self-tracking, particularly health tracking. Several Personal Informatics systems have been developed that can help the user by facilitating the collection of personal information and the reflection on that information at any time [9][10]. However, the challenging part is to get access to user's data without overburdening the user, i.e., such information must be obtained in a non-intrusive manner. Fortunately, with the advances in mobile phone technologies, many of the recent mobile phones comes with a large number of sensors that can be used for automatic tracking of user's data without user's involvement.

SitCoach [11] was designed for creating sedentary awareness among users. Using the built-in accelerometer in the smart phone, the user's activity is classified in an active or inactive state. Whenever a certain number of sedentary minutes is recorded, SitCoach reminds the user to take a break from sitting. Reminders are given in the form of audio and tactile messages. Similarly FitBit+ [12] aims to decrease sedentary behavior by detecting when people have been inactive for longer period and then generate messages for taking short breaks and perform some physical activity. User wears device known as Fitbit that transmits data related to steps taken, intensity of physical activity, duration of activity, distance etc. to their workstation. This data is analyzed and if user is found inactive for longer period, the alerts are generated prompting them to perform some physical activity. However, as we have identified, activeness is not just the prevention of the sedentary behavior, but inculcating the eagerness of living an active life; a goal that is more challenging and requires more treatment.

The Motivate [13] project is about the design of a smartphone application that promotes the adoption of a healthy and active lifestyle. It provides personalized and contextualized advice based on geo information, weather, user location and agenda. However, the context parameters used have nothing to do with user's overall goal of activeness. In fact, the user has no fitness goal whatsoever, and hence may not find the application useful at all after giving it a try a few times only.

StepUp [14] is a mobile application that uses sensor enabled mobile phones to count the number of steps taken by a user. The aim of StepUp is to provide users a quantitative measure of their daily activities and facilitate them to integrate regular exercise into their daily life. StepUp built its own algorithm for counting the number of steps taken by user. However, StepUp only provide users with the information of the activities performed by them and does not include any feature that keep users engaged towards the application.

Players are usually addicted to video games and they prefer to play video games rather than playing outdoor games. This can cause obesity or other chronic diseases. Hirzallah [15] aims to incorporate exercise in the video games so to make games a source of healthy life. The paper targets games where players virtually live inside some city, forest or a war zone. A treadmill design has been proposed that act as a device to capture player's commands for movements. These movements include running, walking, stopping and turning. In this way the players enjoy exercising as well as playing the game. However, sooner or later, the players get exhausted due to exercise and force them to exit the game. That way, such a proposal would keep players healthy. But this design requires a separate device for playing the games as well as extra space is required for the proposed treadmill. With the emergence of more friendly consoles like Microsoft Xbox, this problem will be a serious concern.

Sedentaware [16] is a mobile application that relies on activity recognition to identify sedentary users. Users can set their goals and recommendations are given to them for completing goals. User's activity is monitored using Android API and feedback is provided in the form of alerts to prompt user for performing some physical activity. However Sedentaware lacks in producing a continuous source of motivation to use it for longer period. Users can set their goals but no approach has been defined for awarding or penalizing the users if they fail to complete their goals. This will ultimately lose user's interest and they may not find it helpful in developing a habit of living an active life.

Step Up Life [17] is a smart phone application that encourages users to have a healthy life style by performing contextually suitable physical exercises. Step Up Life provides physical activity reminders to the user after detecting a prolonged interval of sitting. Reminders are provided by considering the user's context information like user location, personal preferences, time of the day and the weather. This makes Step Up Life smart enough to provide timely and relevant feedback. However an approach can be defined for awarding the users on performing exercise when reminder appears. This will create a feeling of accomplishment in the user which in turn results in high motivation.

Authors in [18] have developed Phone Row, a smartphone boat racing game that requires users to make rowing movements and therefore engage in moderate-intensity physical activity. With these rowing movements, users can control the movement of a virtual boat across a virtual track on an external screen. However, the authors reported that due to sub-optimal implementation, users would not want to replay the game and thus not developing a habit of performing a physical activity.

Move2Play [19] uses sensor enabled smart phones to measure, assess and recommends physical activity. The focus of Move2Play is on the number of steps taken by the user as well as on walking distance. For this purpose a pedometer has been built in the application. Goals are set at the start but can be customized lately according to the needs. Users are motivated by showing progress, providing rewards, publishing scores to the social network and by associating emotions to the avatar. Users are awarded with points when their performance is up to mark but no points deduction is considered when they lag behind in completing their goal. This will ultimately lose user's interest and fail to engage users for longer period.

As identified above, the existing systems do not provide continuous motivation to the users and thus fail to keep them engaged for longer period. Due to lack of motivation, users cannot develop a habit to make exercise, a vital component of their life. User uses these systems and will ultimately give up after giving it a few tries. SmartFit engages users by not only providing feedback but also provides the flexibility to change their goals whenever required. Use of gamification techniques for awarding points on performing activities and deducting points on not performing any activity will further boost user's emotions to use the system as a part of their daily life activities. Furthermore as per our knowledge no such system exists that make use of step count to calculate the duration for which the activity has performed. This feature makes our proposed solution distinct from the existing systems and opens new doors of thoughts towards health systems.

III. From Step Count to Activity Duration

We consider a physical activity as any activity performed by the user that involves significant movement of the various parts of the body. According to the WHO physical activity recommendations [20], one needs to do two types of physical activity each week to improve health - aerobic and muscle-strengthening. An aerobic activity is that which gets ones heart beat faster and breathe harder than when one is not performing it. It includes activities from walking, running, pushing a lawn mower, to biking to the store etc. Muscle-strengthening activities are those which work all the major muscle groups of the body (legs, hips, back, chest, abdomen, shoulders, and arms).

Each physical activity is defined by its duration and intensity. Duration is the amount of time spent participating in a physical activity session. Intensity is the rate of energy expenditure while performing physical activity. Intensity can be considered as either light, moderate or vigorous.

Walking is the most common and inexpensive form of the physical activity and has been shown to be an integral component of the physical activities performed by the adults's population[21]. According to CDC guidelines [4], it is recommended that adults should perform at least 150 minutes of moderate-intensity aerobic activity i.e. brisk walking per week or 30 minutes of walk five times a week to remain healthy and physically fit. Health benefits can also be achieved when this goal is completed as a series of shorter bouts i.e. 10 minutes of activity three time a day. However if we consider walk as the performed activity, then user cannot be categorized into any activeness level, only on the basis of activity duration. Every person has its own walking pace depending upon weight, age, height etc. Walking for 10 minutes is different for every person depending on his/her walking pace. Thus a person walking with the fast pace will cover more distance and perform more activity as compared to the one walking with slow pace. Let's take a person who performs 10 minutes of walk and take about 1000 steps while another person also performs 10 minutes of walk but take 800 steps. Both have performed same activity for same duration but the first one has performed more activity as well as covered more distance as compared to second one. This depicts that while walking step count should be considered for categorizing user into any activeness level.

Keeping this as our motivation we propose a framework that take walk as the performed activity and consider user's step count rather then activity duration in order to differentiate between activeness levels. These steps are then converted to the duration for which activity has been performed. Marshall et al. [22] performed a research for translating current physical activity recommendation to pedometer based guidelines. A sample of 97 adults consisting of both males and females with different ages, weights (Normal weight, overweight,obese) and height was taken. They proposed that moderate intensity walking appears approximately equal to at least 100 steps per minute for adults. To meet the goal of 30 minutes moderate intensity walk in a day, individuals have to walk at the speed of 100 steps per minute thus covering 3000 steps in 30 minutes and 15000 steps per week. We consider 100 steps per minutes as a threshold for converting user's steps into activity duration. Considering this, we derive a formula to convert user's step into duration of activity as:

TABLE I: Rules for Determining Activeness Level

Activity State	Moderate
Sedentary	NIL
Lightly Active	Light exercise 1-3 days
Mod. Active	total >150 min
Active	total > 200 min
Energetic	total > 300 min

$$walkingDuration(minutes) = \frac{StepCount}{100} \quad (1)$$

This duration separates the users on the basis of their steps rather then duration. Now if a person performs 1000 steps in 10 minutes then by our formula his/her walkingDuration will be 10 minutes but if a person performs 800 steps in 10 minutes then as per our formula his/her walkingDuration will be 8 minutes rather then 10 minutes which is his/her actual duration.

Based on CDC guidelines [4] and 100 steps per minute threshold as specified by [22] we developed a categorization mechanism based on steps count for measuring walking activity in adults. Table I shows these rules. User's step count are converted into activity duration which is then used to categorize user into different activeness level. walkingDuration is used for categorization instead of step count as all of existing health recommendations [4], [20] describes activeness level with respect to duration for which the activity has been performed. By reflecting user' progress in the form of activeness levels rather then step count will ultimately help users to gauge their performed activities with the existing health recommendations. This idea sets a strong basis for our application and makes our solution distinct from existing systems.

IV. SmartFit

SmartFit is an Android based game that uses sensor enabled mobile phones and tends to change adult's lifestyle by promoting physical activities. The complete workflow of SmartFit is shown in Figure 1. Users enter their goals by using the application interface. SmartFit works by counting user's steps with the help of sensors embedded in the smart phones. This count is then changed to duration for which activity has been performed, a concept portrayed under the section III. Points are awarded if users fulfill their goal while points are deducted if they lags behinds, a method described in detail in section V.

User's total score is then converted into trophies and badges which is then displayed as feedback to user on the application's interface. The feedback is also presented in the form of user's daily, weekly and monthly progress. The objective is to keep users motivated to have a healthy life style as well as engage them for longer period. Beginning with prototype, SmartFit comprises of following components.

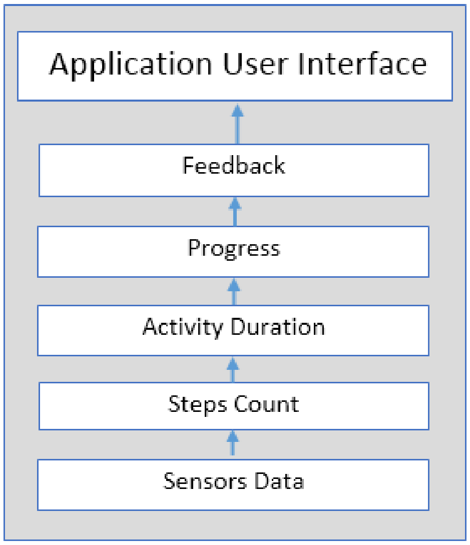


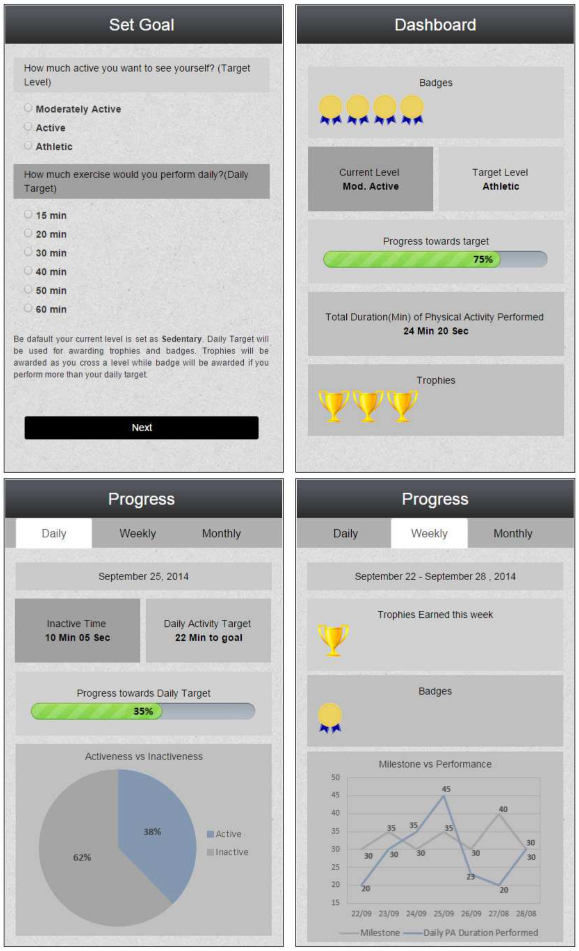
Fig. 1: Workflow for SmartFit

A. Goal Initialization

Users can have one of three goals: to be Moderately active, active or energetic. They set their goals at the start of the application but these goals can be customized in future as per needs. They will select their target level as how much active they want to see themselves in order to have a healthy life style. Goals setting theory [23] suggests that for changing the behavior the target goal must be broken into smaller goals that are easy to achieve rather then focusing on goals that are difficult to complete. Keeping this in mind, target level is divided into sub levels that user has to achieve on daily basis. These sub levels lead users ultimately to the final goal. Users will also be asked to enter the duration of exercise they would perform daily as their daily activity target. This daily activity target serves as the sub goal to the target goal. The daily activity target will be set at the start but can be changed lately. Figure 2 shows the SmartFit's interface for setting the goals.

B. Activity Recognition

SmartFit aims to continuously monitor user's behavior and detect number of steps when he/she has performed some physical activity. Till now we have only focused on walking activity. Google Fit [2] is a health-tracking platform developed by Google for the Android operating system. Google Fit uses sensors in a user's mobile device to record physical fitness activities (such as step count, walking, running or cycling).



How much active you want to see yourself? (Target Level)

○ Moderately Active
○ Active
○ Athletic

How much exercise would you perform daily?(Daily Target)

○ 15 min
○ 20 min
○ 30 min
○ 40 min
○ 50 min
○ 60 min

Be dafault your current level is set as **Sedentary**. Daily Target will be used for awarding trophies and badges. Trophies will be awarded as you cross a level while badge will be awarded if you perform more than your daily target.

Badges

Current Level
Mod. Active

Target Level
Athletic

Progress towards target
75%

Total Duration(Min) of Physical Activity Performed
24 Min 20 Sec

Trophies

September 25, 2014

Inactive Time
10 Min 05 Sec

Daily Activity Target
22 Min to goal

Progress towards Daily Target
35%

September 22 - September 28 , 2014

Trophies Earned this week

Badges

Fig. 2: a) Set Goals b) SmartFit Dashboard c) User's Daily Progress d) User's Weekly Progress

SmartFit makes use of Google Fit API to recognize number of steps taken by user. The objective of activity recognition is to identify user's behavior and then motivate him/her that in order to achieve the daily activity target, they have to do some physical activity.

C. Progress Monitoring Dashboard

Feedback is provided to users by reflecting their progress timely on SmartFit's dashboard. Users can monitor their progress by viewing the dashboard. The dashboard provides complete information about the goal. User's current progress towards achieving the final goal is shown in the form of progress bar on the dashboard and on mobile's notification bar. Dashboard also shows the points earned by the user in the form of trophies and badges as well as total duration of exercise performed by him/her. Users can also view their daily, weekly and monthly progress of activities performed by them. Daily progress shows the time period for which the user is inactive. The progress towards achieving the daily activity target is also shown in the form of progress bar. This allow SmartFit users to monitor their progress and how they can plan their actions for completing their daily target. Figure 2 shows the dashboard and daily, weekly progress screens of

[2] https://developers.google.com/fit/overview

SmartFit.

V. Gamification in SmartFit

Incorporating gamification techniques in the health systems have been proven as the positive source of engagement for users. As stated by Baranowski, et al. [24], duration of physical activity may be increased by integrating fun enhancing procedures. Increasing difficulties with player skills, feedback on performance and awarding achievement badges increases user satisfaction which in turn results in increased physical activity. For calculating users progress, we have developed a scoring system in SmartFit that awards/deducts users points by the function of progress made by him/her. Users set their target level and perform physical activities in order to achieve the desired target. The point system is described as following.

A. Rewards

Users are awarded with points based on physical activities performed by them. If the user's performance is up to expectation then he/she is awarded with points while they are deducted if the user lags behind. Most of gamification techniques also focuses on using awards/badges as intermediate awards for rewarding users. SmartFit represents user's points in the form of trophies and badges. Badges are awarded when user crosses a level while trophies are awarded if the user performs more then the duration of daily activity target. If the user does not perform any activity for the longer period of time, then he/she is penalized and even downgraded to lower level which eventually results in the return of the badge.

B. Levels

SmartFit constitute of five different levels ranging from Sedentary to Energetic. User's initial level is set to sedentary and is updated as he/she gets promoted to higher levels. Users are promoted to higher levels on the basis of points they earned by performing activities. Figure 3 shows the five levels of the game. The minimum level is sedentary while the maximum level of activeness that user can achieve is Energetic. Each level is associated with a milestone that has to be achieved by user in order to cross that level. When user's points crosses his/her current level milestone then user will be promoted to the next level. However if a user's points falls below his/her current level milestone then he/she is downgraded to the lower level. The levels along with their corresponding milestone have been shown in Table II. Value of each level's milestone has been set after carrying out several experiments. These values are set in such a way that they translate the Table I levels into SmartFit's levels. The notion of levels gives the user a sense of accomplishment and progress toward meeting the larger goal.

C. Point System

Points are awarded to users in order to keep them engaged towards the application thus achieving the objective to keep users healthy and active. Points are not only awarded but also deducted by the function of progress made by user. The starting score is initialized to zero and is updated as the user progresses and performs some physical activities. Some of key points of SmartFit's points system are as follow:

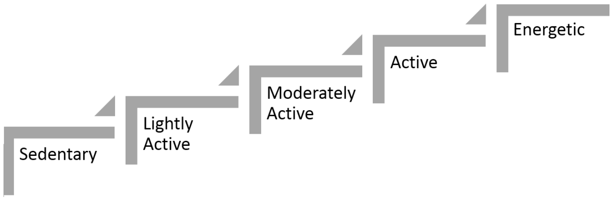


Fig. 3: Game Levels

TABLE II: Game Levels and Milestones for upgrade between levels

Level	Milestone (Points)
Sedentary	30
Lightly Active	70
Mod. Active	130
Active	220
Energetic	380

- User provides daily activity target as his/her preference. Daily activity target is used to award/deduct points. This gives user freedom to perform exercise as per their preferences thus providing them control over the system.
- Steps are counted and converted to walking duration as shown in Equation (1).
- User's daily score is defined as:
  $$dailyScore = walkingDuration$$
- If user's daily Score is less then daily activity target then we define penalty as:
  $$penalty = dailyActivityTarget - dailyScore$$
  However if a user completes daily activity target then penalty is zero as he/she completes the goal successfully so instead of penalizing, smartFit awards user with points.

1) Incorporating Endurance and Fallback: D. Wortley suggests that gamification works by setting goals, rewarding winners and penalizing losers [25]. In order to incorporate the penalizing part in SmartFit, we introduce factors of Endurance and Fallback. Endurance is associated with each level and its value is shown in Table III. The value of fallback is set to 0.3 and it is same for each level. The values of endurance and fallback has been set after carrying out several experiments.

Endurance is used for penalizing user on daily basis. It is included in SmartFit for motivating the user that he/she has to perform recommended level of physical activities regularly. If the user underperforms as compared to his/her current level then due to endurance he/she will be immediately downgraded to lower level. Furthermore endurance also makes the completion of level challenging thus encouraging the user that he/she has to perform walk on daily basis in order to hold the current level.

Fallback is used for penalizing the user on weekly basis. It is introduced in SmartFit for keeping the user's performance consistent over the week. If user performance is found to

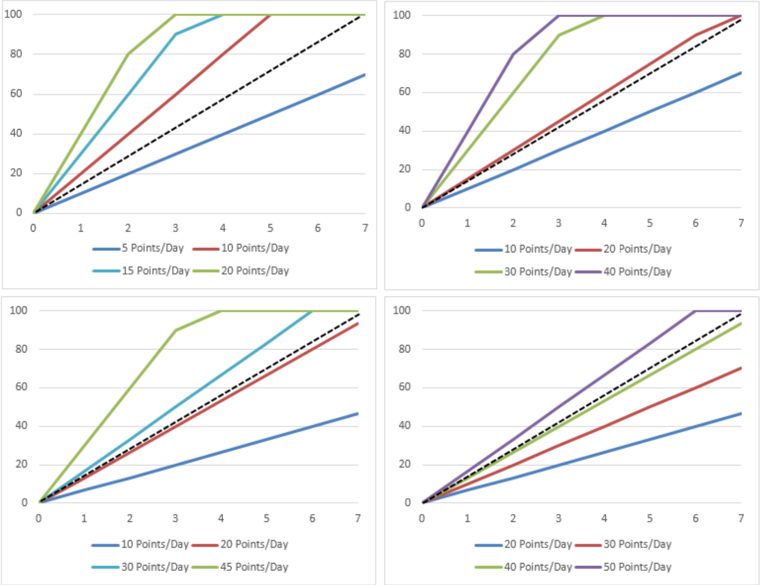


Fig. 4: Simulation Results for identifying level's milestone a)Lightly Active b)Moderately Active c)Active d)Energetic

TABLE III: Endurance for each Game Level

Level	Endurance
Sedentary	0.7
Lightly Active	0.8
Mod. Active	0.8
Active	0.8
Energetic	0.8

be unsatisfactory during the week then his/her level will be dropped. This keeps the user motivated that he/she has to perform consistently in order to hold the current level as well as for progressing towards higher levels. Also through experiments we get to know that penalizing users on weekly basis makes the point system balanced.

2) Score Calculation: Keeping in view the above criteria, user score is calculated by using the following algorithm.

```
**if** $dailyScore \geq dailyActivityTarget$ **then**
    $dailyScore + (totalScore \times currentLevel.Endurance)$
**else**
    $dailyScore - Penalty + (totalScore \times currentLevel.Endurance)$
**end**
```

At the end of week, the user's points are calculated as:

$$totalScore = totalScore - (totalScore * 0.3)$$

VI. Simulation Results and Discussion

The proposed point system was defined after carrying out several experiments in which values of level's milestone, endurance and fallback has been identified. These experiments are carried out with the intention to find out such values of milestone, endurance and fallback that keep the balance between user's skills and user engagement. This balance will serve the purpose of engaging the user for longer period. Furthemore values of level's milestone has been set in such a way that they translate the Table I levels into SmartFit's levels thus making both levels equivalent.

Figure 4 shows the results of experiments performed to determine the milestone of each activeness levels. Table I specifies the duration of exercise required in a week in order to reach a certain level. Thus we have performed the experiment for the time period of seven days so to convert per week duration into daily points. The dotted line in Figure 4 represents the optimal performance. User has to earn points that crosses this line and keep him/her as close as possible to this line. Experiments have been performed by taking different points per day and then the results are compared with the optimal line. If we consider Figure 4 (a) i.e. Lightly Active Level then we can see that the line representing 10 points per day crosses the optimal line as well as it is closest to that line. Thus the milestone for Lightly Active is set to 10 points per day. This means that user has to perform physical activity of 10 minutes on daily basis for maintaining the Lightly Active level. But if a user does not perform 10 points per day then he/she will be downgraded to the lower level which in this case is sedentary. Similarly to reach Moderately Active level user has to perform at least 20 minutes of exercise everyday as the line representing 20 points per day crosses the optimal line and is nearest to that line.

A. Step Recognition Accuracy

As described earlier, SmartFit makes use of Google Fit API for counting the steps taken by the user while walking. So it is necessary to first check the accuracy of step recognition of Google Fit. For this purpose we performed a experiment with 5 subjects, 2 female and 3 male. The subjects were asked

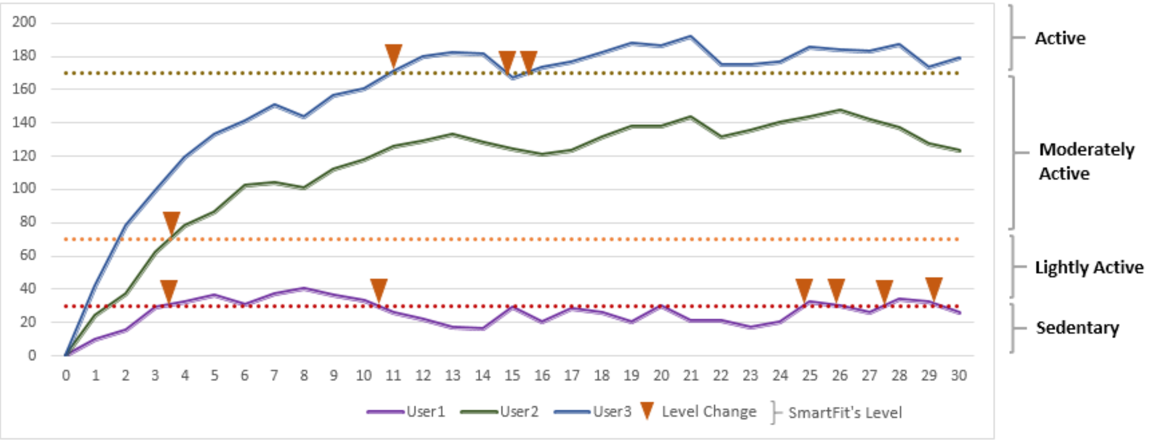


Fig. 5: User Performance vs. Level change

TABLE IV: Summary of Results obtained on testing Google Fit

Gender	Actual Steps	Detected Steps	Accuracy
Female	100	92	92 %
Female	200	185	93%
Male	150	145	97%
Male	220	171	78%
Male	243	172	71%

to count the number of steps taken by them. These steps are then compared with the steps detected by Google Fit so an assessment of accuracy could be made. Users were asked to carry the cell phone running the Google Fit in their hands. The results of the experiment are displayed in Table IV. The average accuracy of Google Fit turns out to be 86%.

B. SmartFit's Evaluation

Now we will present simulation results of our prototype in order to evaluate that whether SmartFit places user into his/her activeness level as specified in Table I. Till now we have evaluated SmartFit on a small scale by considering three users. SmartFit's behavior was simulated for the period of 30 days and results are shown in Figure 5.

User1 takes on average 1500 steps (Min: 1000 , Max: 2000) daily, user2 takes 3000 steps (Min: 2500 , Max: 3500) on average and user3 takes 4000 steps (Min: 3500, Max: 4500) on average while walking. On converting user's step into walking duration then user1 performs 15 minutes of walk on average while user2 do 30 minutes of walk daily and user3 do 40 minutes of walk on average daily. Daily activity target is set to 30 minutes for all the users. What we observe from these readings is that endurance and fallback continuously upgrades/downgrades user level by the function of his/her performance as well as makes the completion of level challenging, thus motivating the user to perform more in order to be active. According to Table III, milestone for Moderately Active level is more as compared to Lightly Active/ Sedentary level. Thus the graph rises very quickly for Moderately Active users while graph for Sedentary or Lightly Active users rises slowly. So the users targeting the higher levels have to perform more in order to reach their desired goal as compared to users targeting the lower levels.

As described above, 15 mins of walk is required daily in order to categorize user as Lightly Active while 30 mins are required to move towards Moderately Active level. User1 behavior has been identified as Lightly Active so SmartFit ensures that he never crosses Lightly Active level. Thus after 28 days he is on Lightly Active level. However variations can be seen between Sedentary and Lightly Active level for user1 in order to motivate him more towards performing activities.

User2 performs walk of on average 30 min of walk daily which is enough to declare him as Moderately Active. Thus SmartFit upgrades the user2's level from sedentary to Lightly Active in two days while level will be changed from Lightly Active to Moderately Active after 4 days and then SmartFit helps user to maintain this level.

User3 performs on average 40 mins of walk daily due to which SmartFit identifies her as Active and moves her quickly to the corresponding level. However variations will be seen between Moderately Active and Active level for User3. These variations are due to SmartFit's scoring system that motivates user that she has to perform more in order to be maintain Active Level. Now if we consider the performance of user1 and user3 on day 14 then we can see the benefit of using fallback factor. At day 14, the total score of user1 falls as his previous week performance is not up to the mark. User1 underperforms during the whole week and as the result points are deducted from his total score. This deduction of point is due to fallback factor and it is used to motivate the user to perform the recommended level of physical activity consistently. But if we see the performance of user3 in 2nd week i.e. from day 7 to day 14 then her performance is up to the mark. User3 has improved her score during the whole week so at the end of week the fraction of points that has been deducted from her score does not cause enough damage to user3's current level and score as compared to user1.

What we see from these simulations is that Smartfit's scoring system not only categorizes users in their respective activeness levels as specified in Table I but by deducting points they also keep them motivated to perform some exercise daily. SmartFit also upgrades/downgrades user's level depending on his/her progress so to keep him engaged and motivated for retaining the current level. However we need to test our system on real users to find out its impact on their behavior.

VII. Conclusion and Future Work

It is recommended that adults should perform 150 minutes of moderate intensity aerobic activity per week to remain healthy and physically fit. However when walk is considered as the performed activity, then activity duration is not enough for categorizing user into any activeness level. Step count is another factor that should be taken into account. Users can perform activity for same duration but they can differ depending on their step count. This article presents an approach for converting users step into the duration for which activity has been performed. We have developed a prototype, SmartFit that uses user step count and converts these steps to activity duration. The objective of SmartFit is to promote physical activities and enable users to achieve their activeness goal by dividing the goal into level and sublevels. Using gamification techniques, users are awarded with points if their performance is up to mark while points are deducted if a user lags behind in completing his/her goal. Unlike existing systems, which focus on the goal itself, our focus is on facilitating the user to achieve the goal.

To ensure that our approach is applicable to wide variety of users, we intend to conduct experiments on larger scale consisting users from different population, age groups, gender etc. After that we will be confident about the impact of SmartFit on user lifestyle.

Moreover, we plan to incorporate user preferences as well as exercise recommendation in SmartFit. User will enter his/her preferences e.g. weight, age, health condition etc. The exercise will be recommended by considering the user's preferences and points are deducted accordingly.

References

[1] Genevieve N Healy, Charles E Matthews, David W Dunstan, Elisabeth AH Winkler, and Neville Owen. Sedentary time and cardio-metabolic biomarkers in us adults: Nhanes 2003–06. *European heart journal*, page ehq451, 2011.

[2] Simon J Marshall and Ernesto Ramirez. Reducing sedentary behavior a new paradigm in physical activity promotion. *American Journal of Lifestyle Medicine*, 5(6):518–530, 2011.

[3] Michael Babyak, James A Blumenthal, Steve Herman, Parinda Khatri, Murali Doraiswamy, Kathleen Moore, W Edward Craighead, Teri T Baldewicz, and K Ranga Krishnan. Exercise treatment for major depression: maintenance of therapeutic benefit at 10 months. *Psychosomatic medicine*, 62(5):633–638, 2000.

[4] Centers for Disease Control and Prevention. Recommendations for physical activities, http://www.cdc.gov/physicalactivity/everyone/guidelines/adults.html.

[5] Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. Gamification. using game-design elements in non-gaming contexts. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 2425–2428. ACM, 2011.

[6] Jane McGonigal. *Reality is broken: Why games make us better and how they can change the world.* Penguin, 2011.

[7] Thijs Franck. How to engineer an app that changes habits effectively. *21st Twente Student Conference on IT*, 2014.

[8] Brian J Fogg. Persuasive technology: using computers to change what we think and do. *Ubiquity*, 2002(December):5, 2002.

[9] Ian Li, Jodi Forlizzi, and Anind Dey. Know thyself: monitoring and reflecting on facets of one's life. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 4489–4492. ACM, 2010.

[10] Ian Li, Yevgeniy Medynskiy, Jon Froehlich, and Jakob Larsen. Personal informatics in practice: improving quality of life through data. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 2799–2802. ACM, 2012.

[11] Saskia Dantzig, Gijs Geleijnse, and Aart Tijmen Halteren. Toward a persuasive mobile application to reduce sedentary behavior. *Personal and ubiquitous computing*, 17(6):1237–1246, 2013.

[12] Laura R Pina, Ernesto Ramirez, and William G Griswold. Fitbit+: A behavior-based intervention system to reduce sedentary behavior. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on*, pages 175–178. IEEE, 2012.

[13] Yuzhong Lin, Joran Jessurun, Bauke de Vries, and Harry Timmermans. Motivate: context aware mobile application for activity recommendation. In *Ambient Intelligence*, pages 210–214. Springer, 2011.

[14] Ashraf Khalil and Suha Glal. Stepup: a step counter mobile application to promote healthy lifestyle. In *Current Trends in Information Technology (CTIT), 2009 International Conference on the*, pages 1–5. IEEE, 2009.

[15] Nael Hirzallah. A simple exercise-to-play proposal that would reduce games addiction and keep players healthy. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 4(2), 2013.

[16] Hamid Mukhtar and Djamel Belaid. Using adaptive feedback for promoting awareness about physical activeness in adults. In *Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC)*, pages 638–643. IEEE, 2013.

[17] Vijay Rajanna, Raniero Lara-Garduno, Dev Jyoti Behera, Karthic Madanagopal, Daniel Goldberg, and Tracy Hammond. Step up life: A context aware health assistant. 2014.

[18] Matthijs Jan Zwinderman, Azadeh Shirzad, Xinyu Ma, Prina Bajracharya, Hans Sandberg, and Maurits Clemens Kaptein. Phone row: a smartphone game designed to persuade people to engage in moderate-intensity physical activity. In *Persuasive Technology. Design for Health and Safety*, pages 55–66. Springer, 2012.

[19] Pavol Bielik, Michal Tomlein, Peter Krátky, Štefan Mitrík, Michal Barla, and Mária Bieliková. Move2play: an innovative approach to encouraging people to be more physically active. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 61–70. ACM, 2012.

[20] World Health Organization. Global recommendations on physical activity for health. W.H.O. report, 2010, ISBN: 9789241599979.

[21] Thomas Stephens, David R Jacobs Jr, and Craig C White. A descriptive epidemiology of leisure-time physical activity. *Public health reports*, 100(2):147, 1985.

[22] Simon J Marshall, Susan S Levy, Catrine E Tudor-Locke, Fred W Kolkhorst, Karen M Wooten, Ming Ji, Caroline A Macera, and Barbara E Ainsworth. Translating physical activity recommendations into a pedometer-based step goal: 3000 steps in 30 minutes. *American journal of preventive medicine*, 36(5):410–415, 2009.

[23] HA Schut and HJ Stam. Goals in rehabilitation teamwork. *Disability & Rehabilitation*, 16(4):223–226, 1994.

[24] Tom Baranowski, Ralph Maddison, Ann Maloney, Ernie Medina Jr, and Monique Simons. Building a better mousetrap (exergame) to increase youth physical activity. *GAMES FOR HEALTH: Research, Development, and Clinical Applications*, 3(2):72–78, 2014.

[25] David Wortley. Gamification and geospatial health management. In *IOP Conference Series: Earth and Environmental Science*, volume 20, page 012039. IOP Publishing, 2014.

Cyberspace Challenges and Law Limitations

Aadil Al-Mahrouqi
School of Computer Science and Informatics
University College Dublin
Dublin, Ireland

Cormac O Cianain
School of Computer Science and Informatics
University College Dublin
Dublin, Ireland

Tahar Kechadi
School of Computer Science and Informatics
University College Dublin
Dublin, Ireland

***Abstract*—Privacy and Data security are heating topic in the modern technologically advanced economy. Technological Innovations have created new forms of electronic data which are more vulnerable to theft or loss when compared to traditional data storage. Moreover, the recent advances in internet technologies have exacerbated the risk of security threats. The Internet brings a whole new set of challenges in terms of data protection. Considering the complexities of modern technological advancements and its impact on data security, this study examines the Irish laws and EU directives for privacy and data security, its effectiveness in managing large scale data breaches and limitations. This paper also simulates attack scenarios that can be done by anonymous users in a complex cyberspace environment and explains how a digital evidence related to the attack scenario can be tracked down.**

Keywords*—Internet anonymous; pseudonymous internet users, electronic discovery; large-scale data breaches

I. Introduction

Cyberspace has become a vital part of individuals and communities worldwide. Many key sectors of the global economy including banking and finance, health sector, communication and the defence relies heavily on cyberspace (United States Department of Defence, 2011). According to Deibert and Rohozinski [1] cyberspace has become an indispensable part of the social, political and economic power worldwide. Cyberspace security threat is a key challenge for the modern society. Many critical infrastructures of the society rely heavily on cyberspace that makes it vulnerable to disruption and exploitation. It represents one of the most serious threat to national security and public security [2]. Any risks from cyberspace are severe since it undermines the safety and security of citizens and cause disruption in social and political life. The constant innovation and advancements in cyberspace technologies continuously generate new forms of security challenges. Users who conform the basic protocol to internet connectivity increases the participation of people from all backgrounds creating a constant flux based on ingenuity [3].

The rise in security threats generated from the development of cyberspace has increased the need for tighter laws and regulations. However, the constant transformation and high degree of complexity of cyberspace creates a major barrier for its regulation. Cyberspace is characterized as a network of interconnected electronic communication channel [4]. The transnational organization of the cyberspace networks makesthe states be fully in control of the entire activities in the cyberspace. This lack of physical proximity and control is a major barrier to states regulations to manage increasing data security breaches. All these special properties and complexities of cyberspace allow cybercrime to elude state control [1]. This paper specifically examines the cyberspace security threats from anonymous and pseudonymous Internet users, electronic discovery challenges, law responses to the problem to these security threats and large-scale data breaches and finally evaluates the current limitations in Irish laws and regulations.

Aadil

August 27, 2015

II. Literature Review

The purpose of this literature review is to provide an overview of the most relevant, previous research done on the legal laws that focus in anonymous and pseudonymous internet users, electronic discovery challenges, law responses to the problem to these security threats and large-scale data breaches and finally to evaluate the current limitations in Irish laws and regulations.

A. Anonymity and Pseudonymous Internet

The Internet provides all the users across the globe the freedom and the choice to remain anonymous or pseudonymous. Anonymity has become the cornerstone of Internet communication that promote free speech. Many people prefer to remain anonymous or pseudonymous on the Internet for several reasons which may not always be with criminal intent. For example some people use pseudonymous ID for fun or share information for the benefit of society without revealing their identity. Although not all people misuse the choice of Internet anonymity, people with criminal intent use Internet anonymity techniques to perpetrate cybercrime [5]. Anonymity in blogging is very popular these days. When some bloggers prefer to use their real names some prefer to be pseudonymous to communicate and share their messages and thoughts on blogs. Organizations also use blogs to keep in touch with their customers to obtain feedback about their products and services. It also allows employees to share new ideas which can be used by companies to develop new strategies. Anonymous blogging without any harm to others is permitted among Internet users.

The protection of identities of anonymous and pseudonymous people in Internet varies between countries and depends on the nature of the activities. For example when anonymous

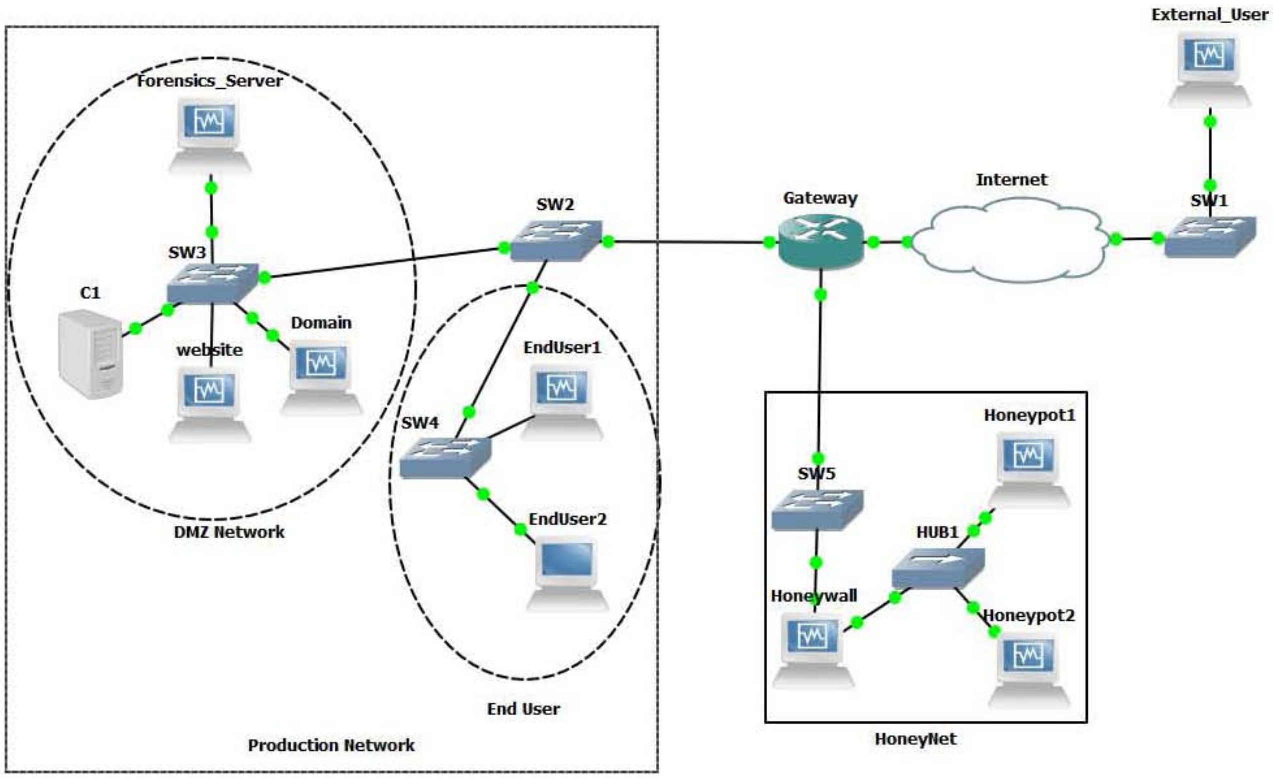


Fig. 1: Simulation Honeynet Network in GNS3

blogging affects the reputation of individuals and organisations, there are laws to reveal the identities of the anonymous person. Although there are exceptions to this law. There are many cases where organisations reacted to anonymous bloggers which when affected their reputation. One such case was when an employee was sacked by her employer for writing an Internet diary under the pseudonym Petite Anglaise which resulted in reputational issues for her employer. The woman was awarded 30,000 for wrongful dismissal [6]. However, blogging can become bullying when the privacy of others are harmed by sharing secrets that affect the reputation of individuals and organisations. Ryanair appealed to the high court of Ireland seeking the disclosure of identities of anonymous individuals using code names such as ihateryanair and cantfly wontfly who made intimidating posts about their pilots [7]. However, a high court in Ireland dismissed the application. A similar case in US, Totalise plc v. Motley Fool was successful where the court ordered to identify the pseudonymous user using the Zeddust ID, under the data protection act and section 10 of the contempt of court act for posting negative comment. [8].

Anonymous downloading has resulted in the illegal downloading of music and film from online website. The first case in Ireland relating to anonymous downloading was brought to the high court in 2008 by EMI and three other major record companies to compel Eircom to prevent its broadband and customers downloading music illegally through file sharing website [9]. In the context of Article 2 and 8 of the Data protection act 1988-2003 [10], Eircom ensured the confidentiality of the users who illegally downloaded music from file sharing to the plaintiff until the courts order released. There has been similar cases in Beijing, China where the ISPs were forced to block the P2P download [11], [12].

B. Electronic Discovery Challenge

Technological changes and development in the information processing are becoming increasingly complex and incredible throughout the history. We have gone through an era of paper based society which is now transformed to a completely electronic based society with the technological advancements. These advances in information processing has brought major changes to the form of record keeping and information processing. The transformation from paper to digital information processing and record keeping has opened out a new set of challenges and complexities to the society. The digital record keeping and information processing has created an explosion of information which is the fundamental reason for the e-discovery problem. E-discovery is the process of identifying, preserving and collecting electronic documents that may be considered relevant to a matter. Electronic information sources are becoming an increasing source of evidence for modern disputes and court can order expensive electronic discovery to form evidences. E-discovery in Ireland is currently governed by the Rules of the Superior Court (Discovery) 2009 S.I. 93 2009 [13]. McCarth v.OFlynn [1979 I.R.127 (x-ray) and Clifford v Minister of justice [2005] IEHC 2008 are examples of cases in Ireland which explained that data discovered in a court can be any item that gives an information as opposed to an item on which writing can be inscribed.

Electronic discovery has been a significant issue in Irish court over the past decade. Based on the significant commercial disputes on electronic discovery, the litigation committee of the law society of Ireland issued a report with recommended changes to the 1999 electronic discovery rules of the superior court of Ireland. The whole recommendations of the litigation committee was accepted and adopted by the rules committee to the superior court of Ireland in 2009. This provided Irish courts with a wider scope and offered significant flexibility in terms of electronic discovery applications. Courts began to consider several complex electronic discovery issues including application concerning Meta data and reasonableness of processes employed. The superior court in a telecom case has stated that the data mining application can be granted as court has the right to order the disclosure of documents or creation of new documents under electronic data discovery.

Organisation has the obligation to preserve document under the new discovery rule. The law society of Ireland has stated clearly that the party seeking discovery to provide a letter to the party against whom the discovery sought specifying the specific category of documents required and the reasons for which the specific category of documents is sought. Failure to discover all the data sought based on the necessary could be a breach of party's obligation to discovery [14]. When a discovery is made, the party against whom the discovery is sought is required to prepare an affidavit that report the statements of the documents. The law society makes it clear that there is not any regulation on the documents that are lost or spoiled. When there is a discovery requirement by Irish law and the documents are lost or spoiled, then there is an obligation to prepare an affidavit that states that I formally had but no longer have the documents required. It should also include an obligation that specify what happened to those document when you last had them and where they are now and obviously an adverse influence can be drawn against you if you lie in this regard or you had these document but recently you have destroyed them or lost them or they went up in the fire.

There are several issues associated with the discovery of documents in the present legal system mainly due to the differences in the traditional rules on document discovery and the increased use of electronic documents by individuals and organisations. The first problem is what constitutes the documents that have to be discovered. This includes the doubts regarding the type of electronic documents whether it includes electronic data that might or might not constitute the content of the document itself, does it include the Meta data and so on. The second problem is that the traditional rules that have been developed for paper documents is not suitable for the discovery of electronic documents.

There are several other challenges relating to the electronic discovery. Once such challenge is the expense incurred by organisations for electronic discovery. The expense of electronic discovery can even exceed the amount discovered. In a case where eircom was asked to produce a report based on the information in its database the expense was higher than the amount discovered. Based on the EMI v Eircom case of electronic discovery where Internet users have been illegally downloading music, the barrister Ronan Lupton reported in the Electronic Discovery Ireland conference in Dublin that the amount spent on electronic discovery was 700,000 EUR while the amount recovered was only 70,000 EUR. The electronic documents stored by an organisation are mostly in proprietary format which means that the data needed to be interpreted. An organisation might need a software to interpret this information which might be expensive. Moreover, rarely the lawyer has IT knowledge necessary to interpret the electronic documents using software making legal proceeding complex and therefore is a major barrier to legal system. Confidentiality of the electronic documents discovered is a major concern. To ensure the confidentiality of the electronic documents, courts have the authority to limit the sight of the documents as in Koger v ODonnell [2009].

Social media sites such as Facebook, twitter, and LinkedIn has been used by people with criminal intent to perpetuate crimes. The electronic communications in social media sites create an extensive electronic information that have become evidences in many litigations. A recent case was where a woman in Ireland was dismissed from her job for insulting her boss on Facebook. Further examination of the Facebook page revealed more critical comments some containing expletives, about her employer. The employment appeals tribunal ruled that the dismissal of the employee was not unfair. Thus social media communications is a major source of evidence that impact the ethical consideration for lawyers. Disrespectful and defamatory comments or message posted on Facebook or other social media site is highly relevant in modern environment and can create adverse consequences.

International transfer of data in a multinational corporation as per the data request from a different country is a major issues in the cross border data protection. For example, when a multinational corporation with operations in US and Europe received a court order to produce personal data stored in its European affiliate, it created an ethical dilemma for the privacy officer. The privacy officer faced the dilemma of whether to satisfy the compulsory US discovery obligations or to comply with the European data protection law which restricts data disclosures for litigation purposes. This has become a complex problem that not only affects the European affiliates of multinational corporations but also the lawyer who deal with the cross border electronic discovery process. Thus there is a need to reconcile the requirement of the US litigation rules and EU data protection laws to provide a precise guidance for businesses on how to manage such conflicting situations. Pre-trial discovery for cross border civil litigation adopted by the Article 29 Data protection working paper (Art.29 DPWP) is a working paper that aims to manage the conflict between the US litigation rules and EU data protection laws [15]. The articles covers the nature of the problem, the legal issues in the EU data protection laws governing the electronic discovery requests and the working paper guidance and further practical steps that can be used by organisations to tackle the issues.

Electronic discovery approach in Irish jurisdiction is less developed especially in the commercial court proceedings which deal with large number of commercial litigations. The admissibility of the electronic evidence was a major problem in Irish courts before the legislative interventions in the Section 22 of the E-Commerce Acts 2000. One major problem with the computer generated evidence is the difficulty to identify the original evidence. This is because the computer generated evidence can be produced multiple times in the same format which is different from the traditional evidence in the paper. In traditional evidence in the paper, there is a single original and the rest can be called copies. Section 22 of the E-Commerce Acts 2000 permits the use of material which is not in its original form provided that it is the best evidence that could be obtained. However, Irish law does not allow the use of material obtained illegally or unconstitutionally since it is considered inadmissible.

C. Law responses to large scale data breaches

Large scale data breaches has become a common in Irish news headlines. There are quite few cases on public and private bodies losing data from their server, laptops and USB keys. One high profile example is of Bank of Ireland losing the personal information of their customers without prior notification. The financial regulator and Billy Hawkes, the Irish Data Protection Commissioner who examined the case identified and examined the security arrangements in place and the exact circumstances that resulted in the delay in reporting to the appropriate personnel for taking further actions. The only justification provided by the Bank of Ireland for its defense was that it ”monitored all of these customer accounts and can confirm that there has been no evidence of fraudulent or suspicious activity” which itself was insufficient and does not justify the fact that the customer information was not protected.

Another famous example was the security breach is by Health Service Executive (HSE). From 2010-2013, there has been over 69 shocking large and small data breaches by HSE in the form of stolen or lost laptop, USB sticks and smartphones. 61 of the electronic records were stolen, with 51 having unspecified sensitive information and 20 without encryption codes [16]. In security breach in 2008 left thousands of HSE staff open to identity theft when an unencrypted laptop containing the personal details had been stolen from the HSE offices at the Carnegie Centre in Dublin’s Lord Edward Street. The staff were not told about the theft of the unencrypted laptop until after 13 days [17]. Similarly in 2010, hundreds of patient records were seriously compromised by a major security breach at the HSE [18]. In 2011 HSE reported another breach of data when documents including sensitive information of over 100 patients including names, addresses and date of births were discovered in a bin outside Roscommon hospital. More recently, June this year, HSE has breached the rights of its employee by disclosing his salary details to his ex-wife [19].

D. Duties under the Data Protection Act

The continuing high profile data breaches has demanded the need for greater accountability by organisations like Bank of Ireland and HSE that holds personal information. Now the major issue is about what can be done to increase the accountability of these organisations and what steps should be taken to prevent such incidents happening in the future. Organisations are not taking the necessary steps to systematically organise and secure personal data. The problem is that the data belongs to different people and not the organisation that hold the data due to which the organisation lack incentive to secure the data appropriately. There is a need to incentivise the holder of the data. The cost associated with the data breach is for the individuals whose information is lost and the cost of securing the data against the breach is held by the organisation holding the data. One method to incentivise the organisation holding the data is to internalise the cost of data loss. There are currently different data security obligations and duties for data controller to secure data under data protection act which is discussed in the following section.

All the business incorporated in Ireland that gather or processes personal data is required to comply with the data protection acts. The data protection acts makes the data controller accountable for the security of the personal data. Data controller has a duty to keep and secure the data. Section 2(1)(d) of the data protection act 1988 and 2003 states that Appropriate security measures shall be taken against unauthorized access to, or unauthorized alteration, disclosure or destruction of, the data, in particular where the processing involves the transmission of data over a network, and against all other unlawful forms of processing. Thus securing data is to prevent unauthorised access and unauthorised disclosure of data to third parties. A data controller is responsible to prevent both internal and external security breach of personal data. In order to ensure effective data security the data protection commissioner issues a code of practice and guidelines on the responsibilities of data controller and actions to take when personal data is exposed to risk of exposure [20].

Section 2 C of the data protection act prescribes the appropriate security measures by the data controllers which includes the measure to ensure the level of security appropriate to prevent the harm that might result from unauthorised or unlawful data processing or accidental or unlawful destruction of the data concerned. There are cases where the data protection commissioner identified inadequate security measures and took necessary steps. One such example was when the data protection commissioner identified inadequate security measures in 2008 when the Credit Union transmitted the personal data of its customers including username and passwords via unsecure mail. The excuse by the credit union for lesser security mesures for usernames and passwords was because they were afraid that the users might not remember their user details at a later date was unjustifiable by the data protection commissioner.

1) Sanctions for Failure to Provide Adequate Data Security: In Ireland, the Data protection commissioner neither have the power to impose a fine for inadequate data security nor for criminal prosecution for data breaches. This allows large data breaches by public and private organisation escape easily. Moreover this is also one main reason why there is continued data breaches in public organisations like HSE. The data stolen from Bank of Ireland in 2008 was only reported to the data protection commissioner only after a year and the actual cost of data lost is not yet notified. UK has much stricter sanctions for failure of adequate data security, For example, the financial services authority UK fined the Nationwide Building Society (UK) 980,000 for the poor data security which resulted in the loss of a laptop containing sensitive information of millions of its customers [21]. The EU directive (95/46/EC) necessitates having an adequate control for data protection in member states and in case of failure to comply with the national data protection law and the person who suffered the damage is entitled to receive compensation from the relevant data protection controller (Article 23).

2) Data Breach Notification and Department of Justice Review Group: Data breach notification in Irish law requires the organisation to voluntarily disclose any data breaches to data protection commissioner. To ensure data breach notification the data protection commissioner approved a code of practise in July 2010 which states that All incidents in which personal data has been put at risk should be reported to the Office of the Data Protection Commissioner as soon as the data controller becomes aware of the incident. In case of doubt in particular any doubt related to the adequacy of technological risk-mitigation measures the data controller should report the incident to the Office of the Data Protection Commissioner. [22].

The code allows departure from the data breach notification when all the following three conditions are met. They are a) $<$ 100 data subjects affected; b) all data subjects affected have been notified without delay; and c) the incident did not involve sensitive personal data or financial data. In all other cases, the data controllers reporting to the data protection commissioner as per the code of practice should contact the data protection commissioners office within two working days of becoming aware of the incident about all the details of the incident include the circumstances surrounding the incident. Based on the details outlined about the incident of data breach the office of the data protection commissioner determine the need for detailed report and/ or subsequent investigation based on the nature of the incident and the appropriate physical and technological security measure to protect the data.

The department of justice commissioned a review group on data protection law. The justice review group on data protection law recommended two necessary changes for effective enforcement of data protection and to avoid any future breach. The recommendation states that The reporting obligations of data controllers in relation to data breaches should be set out in a statutory Code of Practice as provided for under the Data Protection Acts. The Code, broadly based on the current guidelines from the DPC, should set out the circumstances in which disclosure of data breaches is mandatory. Failure to comply with the disclosure obligations of the Code could lead to prosecution by the DPC.

E. Comparison with The US

Data protection issues have become a growing problem in the modern world with the increased use of Internet anonymity. Thus the Internet has become an easy planet to conduct crimes. Governments all over the world are moving towards regulating crimes over the Internet through laws and regulations. However several crimes escape the law due to the complex nature of Internet anonymity techniques. This section examines the laws and regulations in the US compared to the EU directive and Irish laws and regulations throughout history to regulate anonymous and pseudonymous Internet communications to prevent Internet based crimes [23]. However, the 1997 decision by the federal district court of Georgia in the US invalidated a state law that criminalized anonymous and pseudonymous Internet communications, since Pseudonymous and anonymous communications have been part of American tradition and jurisprudence. Pseudonymous and anonymous communication has been a part of US history with many having made rich contributions to political discourse. Thus the issue was consistent with age-old practices in America. Internet anonymity is treated similarly to anonymity in a leaflet or book, treating it as another form of communication media [24].

Decriminalising anonymous and pseudonymous Internet communication has provided new hope to anonymous and pseudonymous Internet users to continue with criminal activities on the Internet. One main difference between the anonymous communication over the Internet and any other media is the ability of anonymous communication through Internet to reach the population all over the world as opposed to reach in a small region. Wider reach of Internet anonymous activities can have a much larger impact than anonymous communication through any other media. The Irish data protection act 2a 1988-2003 does not allow organisations to disclose any personal information without individuals permission, even when there is a physical and security threat. Thus the terms and conditions of many organisations include a clause that allows an organisation to share the personal information of individual when there is a suspected illegal activity or to prevent physical harm or financial loss. One such examples is the privacy policy statement of PayPal that We may share your personal information with Law enforcement, government officials, or other third parties when we are compelled to do so by a subpoena, court order or similar legal procedure. We need to do so to comply with law or credit card rules. We believe in good faith that the disclosure of personal information is necessary to prevent physical harm or financial loss, to report suspected illegal activity, or to investigate violations of our User Agreement. Other third parties with your consent or direction to do so [25].

The privacy law in the US is much different when compared to the EU directives. When the EU established board standards ensure individual privacy protection [26], the Patriots act in the US allows the US government to search for personal information without users knowledge [27]. Thus the privacy laws in EU and the US are entirely different. The priorities of the EU approach to privacy and data protection law is entirely different from the US privacy and data protection law. The priority of US privacy and data protection law is to ensure justice and security within the country with less priority over the protection of individual priority. However in EU, the privacy and protection of individuals is the main priority. This is evident in the two cases, one in the the US and one in Ireland where the plaintiffs approached the court to reveal the identities of anonymous individuals passing defamatory information over the Internet. The case in US Totalise plc v. Motley Fool was successful and the identities of the anonymous user was identified, whereas the Ryan air attempt to identify defendants who were claimed to intimidate Ryan air pilots was rejected by The High Court, Dublin.

Lawyers in America can preserve and produce electronically stored information without a sanction [28]. However the EU Data Protection and Privacy act (95/46/EC) requires laws to obtain and order to search and preserve evidence. This is because the privacy law in EU is much more comprehensive. The EU privacy act stated that In the EU, privacy is considered a fundamental human right. Therefore, the directive seeks not only to protect E.U citizens privacy generally, but does so within the context of protecting a fundamental right requiring protection of a high degree, which in the Unions language means the maximum possible. Since the US and EU have different privacy and data protection laws, there have been recently a conflict on the privacy and data protection agreement between the US and EU. There is an US - EU agreement to keep the personal information of the transatlantic passengers from EU for up to 15 years [29].

As per S.11 of the Data protection Acts 1988/2003 there are specific conditions that must be satisfied before personal information of European citizens is transferred to third countries. It is stated that Organizations that transfer personal data from Ireland to third countries i.e. places outside of the European Economic Area (EEA) will need to ensure that the country in question provides an adequate level of data protection. Certain third countries have been approved by EU commission that qualified the condition of adequate data protection. The US safe harbour arrangement has been approved for those US companies that have agreed to be bound by the data protection rules. There are currently no cases where the EU member states requested to disclose information of citizens currently residing outside EU. Countries those do not have adequate data protection standards or that are not approved, the data controller ensures data protection rights of individuals in other ways. The controller may use EU approved model contracts that contain data protection standards that are equivalent to EU data protection standards. To ensure data protection by corporation the data controller use EU approved binding corporate rules for international transfer of data and information within the company [30]. The S11 of Data protection Acts 1988/2003 and the Article 29, has set exception to the transferring of data from EU member states to third countries.

The authors [31], [32] proposed the forensics readiness and awareness framework to reconstruct a cybercrime scenario that was previously observed. The proposed framework that contains fifteen different software and database blocks. These blocks works as a single unit in order to forensically process and normalise the captured events. The blocks summarised in four sections, namely, alert logs normalisation, attack scenario reconstruction, information logs normalisation and security awareness and training.

In [33] the authors presented a simulation study network attack scenario. This is the first step towards validating the proposed model. The figure 1 shows the simulation case study used capturing, normalizing and analysing events. The main point of designing virtual network attack environments is to create a sandbox that allows one to perform such experiments, from real assets and at a low cost. Both the capturing and examination of the events were conducted in the simulated case study. The detection of network artifices changes after the executions of SQL-injection attacks were also recorded. The outcome of this experiment can be used as a recommendation in real cyber infrastructure. The core idea of the case study is to examine the website that has been compromised by an SQL injection attack. To simulate this attack scenario many open source tools were used such as Graphical Network Simulator (GNS3), Oracle VM Virtual Box and VMWare workstation. The Wireshark forensics tool was also used to detect criminal activity from the network layer (Layer 3 in OSI model), in addition, the devices memory of victims and attackers examined by using the Volatility Framework 2.4 were also examined.

In addition, simulation approaches helps to graphically simulate an attack for courts, jury and investigators. The simulation approaches also helps to simplify the incidence (An image = 1000 words). The current study [17] proposes investigation learning methodology based on the proposed case study. The learning methodology consists of two stages; stage one is to build a network topology of the proposed case study and stage two is to create a network union matrix.

An in-depth survey for events admissibility in the Irish court of law is carried out in [34]. Overall, the legal review is mainly focused on different primary areas: the admissibility and authentication of digital evidence and focuses mainly on Irish law. Admissibility refers to a set of lawful tests carried out by a judge for forensic assessment of the finding evidence. Trustworthy means that an accurate copy of digital evidence was acquired, and that it has continued to be unchanged since it was recovered. Authentication is a process to check the reliability of digital evidence. The judge summarises five issues that must be considered when evaluating whether evidence will be admitted, namely; not unduly prejudicial, best evidence, not hearsay or admissible hearsay, authenticity and relevance.

The authors also presented an investigation learning methodology based on the proposed case study presented in [34]. The learning methodology consists into stages, stage one is build a network topology of the proposed case study and stage two is to create a network union matrix. Using this setup the authors were then able to simulate specific network devices configuration, perform SQL injection attacks against victim machines and collect network logs. The main motivation of the work is to finally define an attack pathway prediction methodology that makes it possible to examine the network artefacts collected in case network attacks.

Based on this case study the authors proposed a new network forensics model [34] that can makes network events admissible in the court of law. The proposed model presented used to collect available logs from connected network devices, applies decision tree algorithm in order to filter anomaly intrusion, then re-route the logs to a central repository where events management functions are applied.

III. Case study of cyberspace risks and resolutions

The case study scenario is selected based on the issues and problems that are faced in cyber space forensics. In this research, the scenarios have been developed to demonstrate the results and to assist organisations and investigators in dealing with such attacks. There are two attack scenarios that the authors can investigate, the authors have made certain assumptions about the attack strategies used in order to simplify and summarise an attack. One is an internal attack committed by a trusted person within the company, and the other is an external attack committed by an entity whose credentials are unknown to the company. These two scenarios present very different concerns for a company and support, to a point, two differing attack topologies. A third attack type is a hybrid of both of these attack types and can be described as a fuzzy attack. This is one where the attacker is external to the network but establishes a presence within it by compromising a node, gaining a certain degree of control of a node from where he can launch an attack.

A. Incident Summary

Law enforcement received a report that Great International Banks (http://10.55.3.101/) website has been compromised by an unknown attacker. Based on the initial investigation on the website the attacker used different techniques and tools to compromise the victims website such as SQL injection, XSS, broken caching, directory traversal and breaking the local authentication login to the server. Please see the (Figure 1 depicts the Network topology).

Using various tools and methods against the experimental website, an attacker is able to garner lots of information regarding the sites setup, the applications and services running on the device hosting it and back end data meant only for the web application. The system displays many of the vulnerabilities which arise due to poor system administration. The very nature of the Internet is communication. Adding checks, authentication and security can slow down the development of a site and restrict services. Implementing these takes skill and an in-depth knowledge of system, network and application systems. Its often easier to leave a site less secure and use default values for speeding up implementation. Once new features or services are added to an application the site needs to be retested for vulnerabilities. These new features often introduce more security vulnerabilities due to poor error handling and gaining elevated privilege access due to poor authentication and trust between services hosted on the device running the application.

Issues encountered testing this system reveal issues with poor error handling, invalidated user input, cookie poisoning, SQL Injection, XSS broken authentication, cryptography issues, broken caching, directory traversal and a poorly implemented mail relay service. The authors have based the severity on the level of access the authors were able to gain from each of the issues, hear the authors will list the most of the issues which the authors observed in the penetration test as well as a recommendation how to fix these issues.

B. Examples of cyberspace risks

a) Poor error handling (Severity-Medium): Poor error handling is demonstrated by provoking the system into telling you more about the underlying infrastructure. Error messages can be provoked by entering unexpected or unusual input in the application. Using this method against the login page showed that the application was running on IIS (Microsoft Internet Information Services).

HTTP normally runs over port 80, telnetting to ”telnet 10.55.3.101 80” generated the 400 error (HTTP/1.1 400 bad request, server: Microsoft-IIS/5.0). This tells the attacker that the system is probably subject to many of the vulnerabilities with IIS/5.0. It narrows the search of the attacker to this specific application and the attacker can then be tailored towards IIS 5.0. Without this information, an attacker wouldn't be able to leave out attacks for other web servers such as different version of IIS and Apache.

By inserting values via the branch locator page of the application, the authors were able to get the application to accept a value it was not expecting. By using a ”Man in the middle” attack by intercepting traffic to be sent to the application via the browser, the authors were able to insert ”zip= <script >alert(document.cookie) </script >&searchtype=zip” instead of ”zip=1225215&searchtype=zip” which the application was not expecting. This resulted in the application displaying:

```
Microsoft VBScript runtime error '800a000d'
Type mismatch: 'clng'
/locator.asp, line 19
<script >alert(document.cookie) </script >
Line 19 of the code in locator.asp
```

Telling the attacker the name of the file used to interact with the database ”locator.asp”. That the back end database is most likely a version of Microsoft's SQL server, its caused by line 19 in the code for locator.asp and it's been trying is assigned a value that should be an integer.

b) Directory traversal (Severity-Very High): normally a web application would only have access to files in the webroot/<site>directory or shortcuts to cgi/asp directories where the files can only run as a specific web user thus reducing the access to the overall system. From information garnered the site, it was most likely running from a Windows device and seeing a posting in one of the forums $”http : //10.55.3.101 \setminus disclosures.asp?content = .. \setminus .. \setminus winnt \setminus System32\setminus cmd.exe”$, the authors decided to try simple attack and attacker could run. Using Nmap against the site to find it there were any readable directories. The authors mounted a separate image of the web server so the IP address changed as the authors were then able to load a backtrack image and use Nmap. Showing that there were 3 shares, Admin, C$ and IPC” Using the virtual server the authors were then able to navigate to these by running:

$10.55.3.101 \setminus Admin\$$

$10.55.3.101 \setminus C\$$

But both of these required a username and password. The authors were able to find this by using the mounted image and running Nmap again. To reveal that the username ”administrator” and password ”password” would give be accessed. This is a very high vulnerability as it gives the attacker full access to the systems infrastructure and resources. By using this authentication the authors were able to access server files.

c) Default or easily guessable paths: Using Burp and some research into the ISS setup, the authors were able to determine the location of directory readable folders:

$http : //10.55.3.101/Images/$
$http : //10.55.3.101/css/$
$http : //10.55.3.101/html/$
$http : //10.55.3.101/includes/$
$http : //10.55.3.101/js/$

A sample list in the:

$http : //10.55.3.101/js/$

Friday, July 26, 2002 7:12 PM 2514 rollovers.js
Wednesday, December 25, 2002 10:32 PM 23 test.txt
Wednesday, July 31, 2002 4:05 PM 1807 validate.js

$http : //10.55.3.101/images/$

$Friday, July26, 200211 : 06PM2548eycu.gif$
$Friday, July26, 20027 : 19PM644forum_off.gif$
$Friday, July26, 20027 : 17PM644forum_on.gif$
$Tuesday, July03, 20017 : 48AM1101icon_browser_ie.gif$
$ResultsTruncated..$

Most of the folders don't hold sensitive information, primarily used for cascading style sheet, images used in the rendering of the web application but access to the js folder and includes would give the attackers a chance to trick the script to accept values from the user and are executable on the server.

d) SQL injection (Severity-High): These vulnerabilities are found in the area of the website that accepts input from the user and then uses this in the underlying database. By looking around the website, the authors were able to determine the variables accepted by looking at the client side form checks validate.js. The customer number had to be numeric with no letters or special characters:

$functionCheckNumbers(TheNumber)$
$varvalid = true$
$varGoodChars = ”1234567890”$
$vari = 0$
$if(TheNumber == ””)$
$valid = false;$
$for(i = 0; i <= TheNumber.length - 1; i++)$
$if(GoodChars.indexOf(TheNumber.charAt(i))$
$== -1)valid = false;$

The password side accepted the following combination of collection of characters:
$(varGoodChars = 1234567890qwertyuioplkjhg$
$fdsazxcvbnmQWERTYUIOPLKJHGFDS$
$AZXCVBNM!@\#()\ -_; : |? ><, .)$

Using Burp again as the Man in the Middle attack, the attacker is able to bypass the validate.js and replace:
$acctnum = 123123\&txtPassword = sdfsf12\&action = login$; with: $acctnum =' or'1' =' 1\&txtPassword =' or'1' =' 1\&action = login$

This will allowed us to successfully access to the personal account of first record in the database customer accounts. In summary, been able to manipulate the data held within the database itself, using basic enumeration and SQL code. Constrain input by listing acceptable characters. The use of parameterised SQL for data access. Using an account with the minimum level of access which has restricted access to the database. The use of stored procedures with parameterised SQL. Good coding practice would suggest, never concatenate user input with application SQL for form the SQL been sent to the database. Constant vigilance and checking logs for any suspect attacks. Its rare that an attacker would be able to compromise a system on their first attempt. This may give you time to realise potential errors in your code and fix them before they are exploited.

e) Column enumeration (Severity-High): Once the attacker has this level of access, its much easier to gain more access and reveal more information using column enumeration. Using the branch locator page, as a starting point, the authors were able to get the application to reveal the next column by adding 1=1 to the URL giving the response ($http : //10.55.3.101/locator.asp?searchtype = state\&state = z'having1 = 1-$) Giving the attacker the next column in the table. Therefore, the next column name is locator and then branch number. Using column enumeration after every column revealed allowed us as an attacker to from the following ($http : //10.55.3.101/locator.asp?searchtype = state\&state = z'groupbylocator.branchnohaving1 = 1$) Building from this and using the output from the back end database an attacker would get all of the columns available (Branch, Address, City, State, Zip and Telephone).

f) Cross Site Scripting: The authors took information from the forum and tried obvious passwords like admin,password,Aladdin,null, etc against them. One of the accounts (Maria Orlando) had the password for the account 103645516 set to password. Taking the example, and after logging in as Maria Orlando:103645516. This was easy to enumerate in the Burp suite by sending the output to Intruder and building a payload of each password into the account numbers found on the forum. Knowing that the account number was numeric and 9 characters in length made the task easier. The authors were able to gain access using the supplied credentials 103645516:password. The authors then logged into the customer forum as Maria and posted the following java script in the message content:

$< scripttype = ”text/javascript” >$
$alert(”BOOM!!!!”); < /script >$

So this will producing error message in the forum page once that page refresh.

g) Cookie poisoning (Severity-Medium/High): The ability to steal another users identity. Normally used to track a users preferences but can be used once authorisation has been granted to allow the user to login without a password. Very few applications would allow this where financial or medical data is involved, but it has been known to happen. Depending on what the cookie is used to give access to affects the severity. Been able to post messages as someone else, social engineer could allow an attacker to gain an elevated level of access. There is no silver bullet solution but implementing short session timeouts, deleting cookies once a user logs out, setting HTTP only flag and trying to eliminate cross-site scripting from your site will help alleviate issues surrounding cookie poisoning of the stealing of cookies. Hashing the token by using unique features from the client like IP or not unique but the browser been used would make the attackers attempt far more difficult. The authors were able again logged in as Maria but then the authors were able to post into the customer forum as (Michael Nancarrow 10364818), the poisoning code that has been used into the message field of Marias post:

$< script > document.cookie =$
$”custnum = 103646818” < /script >$

However, after executing this attack, the authors had successfully stolen Michaels cookie by using his customer number instead of Marias thus allowing the attacker to post to the forum as Michael. Once the authors clicked back into the customer forum the authors were posting as Michael without having to log out or know Michaels password.

h) Cryptography issues: Would the authors log onto a banking application over HTTP? No chance. HTTP is the telnet of SSH. All data would be in plain text and wouldnt offer any challenge to the attacker once this data is intercepted via a sniffer or other device. A key logger would be one of the few way an attacker could retrieve a users passwords but even if the attacker was able to run a TCP dump on either the network port connecting the victims machine or via airsnort, depending on the RSA or SSL key used, there is little risk of the attacker been able to convert this into usable data. This site should be using HTTPS. The lack of any form of cryptography on the site on any of the services provided by the site makes a well-placed attackers objective far easier by been able to read every transaction in plain text.

i) Broken caching: Neither cache-control or pragma are set on this website. By default, a response is cached if the requirements of the request method, request header fields, and the response status indicate that it is cached. Any form of expiry reduces the time frame an attacker can exploit a vulnerability. Not allowing for a cookie to expire means that an attacker can lift the cookie from any machine the real user used to authenticate to an application and use that.

j) Pragma: Is set on the server to tell the client that its not to cache any of the information locally. Every time the client must request the data required from the server which increases network traffic but ensures the data is the most update. If cookies arent set and stored locally, an attacker cannot use this method to gain access. The server doesnt accept cached information in this case.

k) Un-validated user input (Severity-Medium/High): The checks should be performed on the server side at the very least. Implementing them on the client side is useful and would reduce the amount of traffic between the client and the server but as discovered during this assignment, its trivial to bypass client checks. Expecting all the data been sent to you in a particular format is a bad idea. As will be demonstrated later, the authors were able to by the account number validate check and password.

l) XSS broken authentication (Severity-High): The most common of all the publicly reported security vulnerabilities.

m)Mail relay service (Severity-Medium): If the VM hosting this platform had Internet access this would have been exploitable. If the site relayed it would become blocked by ISPs and SPAM services. Valid mail would be dropped us less the IP was white-listed. An attacker could spoof mail to an e-mail address, a savvy user would look at the headers of the mail received and reasonably assume it was from the Credit Union.

C. Security Recommendation

1- Poor Error Handling generic errors should be returned to the client side without references. All errors can still be logged to help the system administrator to troubleshoot any user errors by searching through logs. These, however, shouldnt be displayed to the client. Having generic error pages with Please contact the helpdesk for almost every error can frustrate the client but its a small price to pay for thwarting an attacker from gaining valuable information.

2- Invalidated user input the checks should be performed on the server side as well as the client side. Setting the lengths of variables and what characters can be entered on the server side reduces the risk of an attacker been able to cause a stack overflow or been able to input a script.

3- Cookie poisoning there is no silver bullet solution but implementing short session timeouts, deleting cookies once a user logs out, setting HttpOnly flag and trying to eliminate cross-site scripting from your site will help alleviate issues surrounding Cookie poisoning of the stealing of cookies. Hashing the token by using unique features from the client like IP or not unique but the browser been used would make the attackers attempt far more difficult.

4- XSS broken authentication implement HttpOnly cookies. A set of strong authentication and session management controls.

5- Broken session management restricting the number of attempts a user may try and authenticate. Implementing a time-out period before a user is able to log on again. On the social side, many applications now send an e-mail to a completely separate account informing the real users that another user is trying to authenticate as them. Taking IP location and where a user normally logs on from. If its different, further preset questions should be asked of the user to ensure the right person is gaining access. This may be a small inconvenience on a valid user but will thwart many attackers.

6- Broken access control setting a minimum requirement for passwords. Ensuring that a password must contain upper and lower case, numbers or special characters and have a minimum length of 7 characters. Not allowing easily guessable password like password, ensuring that passwords expire and cannot be reused. Never store passwords in plain text and avoid using the same root password for all systems for Ease of use. If a password is been transmitted it should be over a signed SSL connection. Allowing administrators to only log on from certain IP ranges by implementing ACLs. Use of one-time used passwords like RSA key for Administrator access. Avoid using usernames like admin/root/boh. Avoid trust relationships between components.

7- Cryptography issues the lack of any form of cryptography on the site on any of the services provided by the site makes a well-placed attackers objective far easier by been able to read every transaction in plain text. Network sniffers and Airsnort type applications would give an attacker a very easy method of gaining access that would be difficult for an administrator to differentiate against. It raises the question if the valid user initiated a transaction or an attacker who sniffed/found/hacked the valid users credentials.

8- Broken caching enable cache control and for vital pages like login etc, ensure these arent cached by setting pragma.

9- Directory traversal setting a minimum requirement for passwords. Not allowing directory listing by configuring the .htaccess file and configuring the /etc/conf/httpd.conf . Ensure the Option Indexes is set correctly so an attacked cannot browse every file in the directory. In IIS which this system is running from, directory listing is disabled by default so don't enable it. If for some reason it was enabled, it can be removed by double-clicking on directory browsing from the user interface of ISS, click on Actions, and disable from there.

10- Mail relay service ensures only authenticated users can real messages. Disable anonymous access and allow all only computers which successfully authenticate to the mail service (probably running MS Exchange) to send mail. At the very least basic authentication should be checked but integrated Windows authentication maybe preferred depending on the setup.

IV. Conclusion

Data breaches are increasingly becoming a major issue with the advances in technologies. Increased storage of electronic data and better technologies to gain access to different forms of data has increased the susceptibility of organisations such as financial services firms and health-care providers who store sensitive information vulnerable to data security breaches. Data breaches are a major cost to the society, and the data protection act aims to protect personal data. The data protection commissioner has developed a code of practice and guidelines to ensure the security of personal data held by the organisation in Ireland. Despite the efforts to ensure data security, the number of data security breaches continues to rise over the years. Statistics shows that the total security breaches in Ireland rose by 47% in two years time. Data security breaches are not just a problem in Ireland but a global issue. The largest data breach of the century was by Heartland payment systems where almost 130 million records were lost [35].

The data security threat is only expected to rise over the coming years with the rise in new technologies such as cloud computing. The modern technology trend of cloud computing that are considered as an effective method for data storage is vulnerable to data theft thereby increase the data security threat. Some countries in EU are already reporting concern on data security with cloud technology. Although, EU directive and Irish data protection law provide more security to individual privacy and data security when compare to the US, there is still a continuous need to review the data security policies and guidelines. The Irish data protection acts are currently developing in line with security requirements, and this trend needs to be continued with the new technological advancements. In time, there might needs to review the data protection act in Irish law and in the EU and can even force to add new data protection and privacy legislation based on the changing data security requirements.

In addition, this paper presents a simulation study network attack scenarios. The main point of designing virtual network attack environments is to create a sandbox that allows the authors to perform such experiments from the real assets and at a low cost. The outcome of this experiment can be used as a recommendation in the real IT infrastructure. The core idea of the case study is to examine the website that has been compromised by various attacks. To simulate this attack scenario, the authors used many open source tools like Graphical Network Simulator (GNS3), Oracle VM Virtual Box and VMWare workstation. The authors used different forensics tools to detect criminal activity from the victim machines, for example, Wireshark, Volatility, Linux dd and HxD.

The future of authentication in the authors opinion is by no means clear. The Internet is about communication. Its primary function is the transmitting, storing and availability of information.

With alternate paths to various destinations, ensuring data transmitted over these paths isnt copied, altered or compromised is against the very protocols designed which established the Internet. The Internet was not designed to be secure, it was designed to be resilient. When the fundamental principals which allowed this were based on open protocols its only a matter of time before any secured connection over it is breached. Users are demanding that authentication methods become to be simpler and more secure. The average end-user doesnt know or care if its via IPv4, IPv6, IPSec, Token ring or 2 tin cans linked together by a taut piece of string. Its just supposed to work. Every form of encryption to date has been broken from ROT13 to SSL. The Enigma code was broken. The latest SSL certs will be compromised. Its the authors opinion that in time anything can be broken.

Acknowledgment

The authors would like to acknowledge the support of by the Insight Centre for Data Analytics and UCD centre for cyber security and cyber crime Investigation.

References

[1] R. J. Deibert and R. Rohozinski, "Risking security: Policies and paradoxes of cyberspace security," *International Political Sociology*, vol. 4, no. 1, pp. 15–32, 2010.

[2] U. DoD, "Department of defense strategy for operating in cyberspace," *July. www. defense. gov/news/d20110714cyber. pdf (accessed 14 September 2013)*, 2011.

[3] J. H. Saltzer, D. P. Reed, and D. D. Clark, "End-to-end arguments in system design," *ACM Transactions on Computer Systems (TOCS)*, vol. 2, no. 4, pp. 277–288, 1984.

[4] G. I. Zekos, "Personal jurisdiction and applicable law in cyberspace transactions," *The Journal of World Intellectual Property*, vol. 3, no. 6, pp. 977–1016, 2000.

[5] H. L. Armstrong and P. J. Forde, "Internet anonymity practices in computer crime," *Information management & computer security*, vol. 11, no. 5, pp. 209–215, 2003.

[6] H. Samuel. (2007) Petite anglaise blogger wins sacking case. [Online]. Available: http://www.telegraph.co.uk/news/1547113/Petite-anglaise-blogger-wins-sacking-case.html

[7] T. McIntyre. (2006) Online anonymity - ryanair edition (continued). [Online]. Available: http://www.tjmcintyre.com/2006/05/online-anonymity-ryanair-edition.html

[8] L. news and guidance from pinsent Masons. (2002) Totalise v motley fool. [Online]. Available: http://www.out-law.com/page-8699

[9] C. J. (2012) Emi records (ireland) limited & ors v the data protection commissioner, [2013] iesc 34 (2013). [Online]. Available: http://www.out-law.com/page-8699

[10] S.-A. Hinfey, "blockingprogress: the irish high court decision in emi v upc," *Journal of Intellectual Property Law & Practice*, vol. 6, no. 7, pp. 494–501, 2011.

[11] J. Delahunty. (2007) Belgium court orders isp to block illegal downloads. [Online]. Available: http://www.afterdawn.com/news/article.cfm/2007/07/04/belgium-court-orders-isp-to-block-illegal-downloads

[12] D. E. Bambauer, R. J. Deibert, J. G. Palfrey, R. Rohozinski, N. Villeneuve, and J. Zittrain, "Internet filtering in china in 2004-2005: A country study," *Berkman Center for Internet & Society at Harvard Law School Research Publication*, no. 2005-10, 2005.

[13] S. Collins, A. Harbison, V. Mee, R. Moore-Vaderaa, C. Murphy, O. OConnor, and D. Moore, "Good practice guide to electronic discovery in ireland," *eDiscovery Group of Ireland*, 2013.

[14] L. S. of Ireland. (2007) Civil litigation discovery in the electronic age: Proposals for change. [Online]. Available: http://www.lawsociety.ie/documents/news/Law%20Society%20Report.pdf

[15] D. P. Commissioner. (2011) Breach notification guidance. [Online]. Available: https://www.dataprotection.ie/docs/Data-Breach-Handling/901.htm

[16] I. Examiner. (2015) Serious hse data breaches risk patient safety. [Online]. Available: http://www.irishexaminer.com/viewpoints/analysis/serious-hse-data-breaches-risk-patient-safety-239828.html

[17] C. Sheehy. (2008) Stolen hse laptop leaves staff open to identity theft. [Online]. Available: http://www.herald.ie/news/stolen-hse-laptop-leaves-staff-open-to-identity-theft-27887535.html

[18] R. Burke. (2010) Hse 'rocked' by security breach on 1,500 patient records. [Online]. Available: http://www.independent.ie/business/irish/hse-rocked-by-security-breach-on-1500-patient-records-26690497.html

[19] E. Edwards. (2015) Hse breached rights of employee by disclosing salary to ex-wife. [Online]. Available: http://www.irishtimes.com/news/ireland/irish-news/hse-breached-rights-of-employee-by-disclosing-salary-to-ex-wife-1.2259714

[20] D. P. Commissioner. (2015) Responsibilities of data controllers. [Online]. Available: https://www.dataprotection.ie/docs/Responsibilities-of-data-controllers/1243.html

[21] P. Williamson. (2007) Nationwide fine for stolen laptop: The nationwide building society has been fined 980,000 by the city watchdog over security breaches. [Online]. Available: http://news.bbc.co.uk/2/hi/business/6360715.stm

[22] D. P. R. Group. (2010) Data protection. [Online]. Available: http://www.justice.ie/en/jelr/dprgfinalwithcover.pdf/Files/dprgfinalwithcover.pdf

[23] D. J. Karl, "State regulation of anonymous internet use after aclu of georgia v. miller," *Ariz. St. LJ*, vol. 30, p. 513, 1998.

[24] J. D. Wallace, *Nameless in cyberspace: Anonymity on the internet*. Cato Institute, 1999.

[25] Paypal. (2013) Privacy policy. [Online]. Available: https://cms.paypal.com/uy/cgi-bin/marketingweb?cmd=-render-content-content-ID=ua/Privacy-popup-locale.x=en-US

[26] J. T. Soma and N. A. Norman, "International take-down policy: a proposal for the wto and wipo to establish international copyright procedural guidelines for internet service providers," *Hastings Comm. & Ent. LJ*, vol. 22, p. 391, 1999.

[27] D. of Justice. (2011) The usa patriot act: Preserving life and liberty (uniting and strengthening america by providing appropriate tools required to intercept and obstruct terrorism). [Online]. Available: http://www.justice.gov/archive/ll/highlights.htm

[28] K. Gates. (2006) E-discovery amendments to the federal rules of civil procedure go into effect today. [Online]. Available: http://www.ediscoverylaw.com/2006/12/articles/news-updates/e-discovery-amendments-to-the-federal-rules-of-civil-procedure-go-into-effect-today

[29] A. Travis. (2011) Air passenger data plans in us-eu agreement are illegal, say lawyers. [Online]. Available: http://www.theguardian.com/world/2011/jun/20/air-passenger-data-plans-illegal

[30] M. D. Birnhack, "The eu data protection directive: an engine of a global regime," *Computer Law & Security Review*, vol. 24, no. 6, pp. 508–520, 2008.

[31] A. Al-Mahrouqi, S. Abdalla, and T. Kechadi, "Cyberspace forensics readiness and security awareness model," *International Journal of Advanced Computer Science and Applications*, vol. 6, pp. 123–127, 2015.

[32] ——, "Network forensics readiness and security awareness framework," in *International Conference on Embedded Systems in Telecommunications and Instrumentation (ICESTI 2014), Algeria, October 27-29 2014*, 2014.

[33] A. Al-Mahrouqi, P. Tobin, S. Abdalla, and T. Kechadi, "Simulating sql-injection cyber-attacks using gns3," *International Journal of Computer Theory and Engineering*, vol. 8, no. 3, pp. 213–2017, 2016.

[34] A. Al-Mahrouqi, S. Abdalla, and T. Kechadi, "Efficiency of network event logs as admissible digital evidence," in *Science and Information Conference 2015, London, United Kingdom, 28-30 July 2015*, 2015.

[35] N. Yau. (2011) Largest data breaches of all time. [Online]. Available: http://flowingdata.com/2011/06/13/largest-data-breaches-of-all-time