IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 3 Issue 9

September2012

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…

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

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

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 3 Issue 9 September 2012
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Associate Editors

Reviewer Board Members

CONTENTS

3D Face Compression and Recognition using Spherical Wavelet Parametrization

Rabab M. Ramadan
College of Computers and Information Technology
University of Tabuk
Tabuk, KSA

Rehab F. Abdel-Kader
Electrical Engineering Department
Faculty of Engineering, Port-Said University
Port-Said, Egypt

***Abstract*— In this research an innovative fully automated 3D face compression and recognition system is presented. Several novelties are introduced to make the system performance robust and efficient. These novelties include: First, an automatic pose correction and normalization process by using curvature analysis for nose tip detection and iterative closest point (ICP) image registration. Second, the use of spherical based wavelet coefficients for efficient representation of the 3D face. The spherical wavelet transformation is used to decompose the face image into multi-resolution sub images characterizing the underlying functions in a local fashion in both spacial and frequency domains. Two representation features based on spherical wavelet parameterization of the face image were proposed for the 3D face compression and recognition. Principle component analysis (PCA) is used to project to a low resolution sub-band. To evaluate the performance of the proposed approach, experiments were performed on the GAVAB face database. Experimental results show that the spherical wavelet coefficients yield excellent compression capabilities with minimal set of features. Haar wavelet coefficients extracted from the face geometry image was found to generate good recognition results that outperform other methods working on the GAVAB database.**

Keywords-3D Face Recognition; Face Compression; Geometry coding; Nose tip detection; Spherical Wavelets.

I. INTRODUCTION

Representing and recognizing objects are two of the key goals of computer vision systems [1-5]. Computing a compact representation of an item is usually an intermediate stage of the vision system, yielding results used by other processes that perform more abstract operations on the data acquired from the objects. Today, the recent development of 3D sensors and sensing techniques stimulated the demand for visualizing and simulating 3D data. The large amount of information involved and the complexity and speed requirements of the processing techniques demand the development of powerful yet efficient data compression techniques to facilitate the storage and transmission of data. The main objective of compression algorithms is to eliminate the redundancy present in the original data and to obtain progressive representations targeting the best trade-off between data size and approximation accuracy [1]. Recently, the interest in 3D face compression techniques has risen as a foundation stage in many areas with a wide range of potential applications such as identification systems in the army, hospitals, universities, and banks to medical image compression and videophones, …etc.

Among numerous biometric modalities, face recognition is one of the most natural and widely accepted authentication and identification methods mainly because of its nonintrusive nature [6-11]. This trend has caught the attention of many academic and research groups and face recognition has become one of the most intriguing and active research areas in pattern recognition and computer vision. In traditional 2D face recognition systems pose and illumination variations always have been challenging problems that severely influence the accuracy of system. In the last decade 3D face recognition is attracting more attention as the increased computing power and 3D scanning technology has enabled the capturing and recognition of faces in 3D [7-8]. The additional knowledge about 3D facial shape has proven to be very useful in eliminating many of the drawbacks of 2D face recognition. This is due to the fact that the acquisition of faces is (to some extent) invariant to changes in illumination during recording and comparison as most equipment based on active stereo vision is robust to illumination variations. In addition, 3D measurements fully preserve the 3D nature of faces and the depth information can easily be used to separate fore- and background. Finally, pose variations can be accounted for by complete transformations (rotation and translations) between different 3D images computed in the 3D space. This efficiently removes the transformation out of the image plane, which is very difficult in 2D face recognition. Therefore, 3D face recognition algorithms are less prone to changes in viewpoint, pose, lighting conditions and subject expressions. The decreasing cost of three-dimensional (3D) acquisition systems and their increasing quality, together with the greater computational power available nowadays, will make real-time 3D systems for face recognition a commonplace in the near future. However, there exist some difficulties in 3D face recognition, such as coping with expression variations, the inconvenience of information capture and large computational costs, these problems have been the focus of recent research [8].

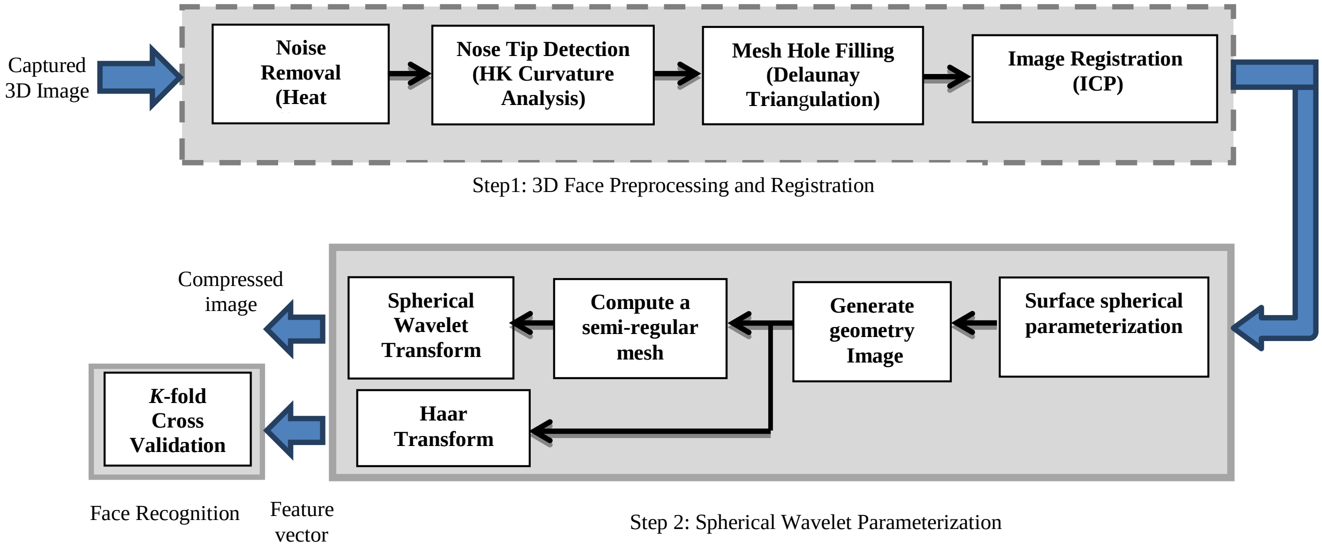

Captured 3D Image
Noise Removal (Heat
Nose Tip Detection (HK Curvature Analysis)
Mesh Hole Filling (Delaunay Triangulation)
Image Registration (ICP)
Step1: 3D Face Preprocessing and Registration
Compressed image
Spherical Wavelet Transform
Compute a semi-regular mesh
Generate geometry Image
Surface spherical parameterization
K-fold Cross Validation
Haar Transform
Face Recognition
Feature vector
Step 2: Spherical Wavelet Parameterization

Figure 1: Block diagram of the proposed 3D face compression and recognition system.

In this paper, a robust and accurate 3D face compression and recognition system is proposed. Gaussian curvature analysis is used for nose tip detection and face region extraction. The Iterative closest point (ICP) is employed to automatically align the face image and to perform the required fine pose correction. The system utilizes discriminative spherical wavelet coefficients which are robust to expression and pose variations to efficiently represent the face image with a small set of features. All processes included in the proposed system are fully automated and can be partitioned into two main stages: 3D preprocessing and registration, and spherical wavelet parameterization. The block diagram of the proposed system is presented in Figure 1. Descriptions of each stage are given as follows:

(1) Preprocessing and registration: First we perform image smoothing using heat diffusion to filter out undesirable distortions and noise while preserving important facial features. Second, the nose tip is detected and used to remove irrelevant information such as data corresponding to the shoulder, neck, or hair areas. Third Delaunay Triangulation is applied to fill holes in the mesh of the extracted face region. Finally, the ICP algorithm is used to align the face image and to normalize the effect of face poses and position variations. This registration process typically applies rigid transformations such as translation and rotation on the 3D faces in order to align them.

(2) Spherical wavelet parameterization: Robust feature representation is very important to the whole system. It is expected that these features are invariant to rotation, scale, and illumination. In our systems, we extract compact discriminative features to describe the 3D Faces based on spherical Wavelet coefficients. First, the 3D face is mapped to the spherical parameterization domain. Second, the geometry image is obtained as a color image and a surface image. Third, the spherical based wavelet coefficients are computed for efficient representation of the 3D face. Two different approaches are utilized for obtaining the wavelet coefficients. In the initial approach, the geometry image is transformed to a semi-regular mesh where the spherical wavelet transform is applied. Alternatively, the Haar wavelet transform can be applied directly to the geometry image.

The rest of this paper is organized as follows: An overview of related work in 3D face compression and recognition is presented in Section II. The preprocessing and normalization tools used in the system are described in Section III. The process of extracting the spherical wavelet coefficients from the 3D face images is explained in Section IV. Section V reports the experimental results and gives some comparisons with existing methods in the literature. Finally, we summarize the paper with some concluding remarks in Section VI.

II. RELATEDWORK

3D meshes are generally used in graphic and simulation applications for approximating 3D Faces. However, Mesh-based surface representations of a face image require large amounts of storage space [1-5]. The emerging demand of applications calling for compact storage, efficient bandwidth utilization, and fast transmission of 3D meshes have inspired the multitude of algorithms developed to efficiently compress these datasets. Image compression has recently been a very active research area but the central concept is straightforward: we transform the image into an appropriate basis and then code only the important expansion coefficients. The problem of finding a good transform has been studied comprehensively from both theoretical and practical standpoints. Excellent survey of the various 3D mesh compression algorithms has been given by Alliez and C. Gotsman in [1, 2]. The recent development in the wavelet transforms theory has spurred new interest in multi-resolution methods, and has provided a more rigorous mathematical framework. Wavelets give the possibility of computing compact representations of functions or data. Additionally, wavelets are computationally attractive and allow variable degrees of resolution to be achieved. All these features make them appear as an interesting tool to be used for efficient representation of 3D objects.

In a typical computer vision system, the compact representation generated from any compression system is used by other processes that perform further operations on the data in the reduced dimension space. Compression algorithms propose a versatile and efficient tool for digital image processing serving numerous applications. 3D Face recognition is one of the imperative applications calling for compact storage and rapid processing of 3D meshes.

Face recognition based on 3D information is not a new topic. It has been extensively addressed in the related literature since the end of the last century [6-11]. Further surveys of the state-of-the-art in 3D face recognition can be found in [7, 8]. Various approaches are reported for extracting and comparing data from facial shapes, each with their own strengths and weaknesses. However, whatever approach is used, three issues always exist that have to be taken into account. (1) The type of facial representation used from which the data is extracted. (2) The way pose or facial orientation differences between different faces are handled which is usually easier in 3D than in 2D but still impose an important challenge. (3) Feature extraction and dimensionality reduction techniques embedded in the system. Several criteria can be adopted to compare existing 3D face algorithms by taking into account the type of problems they address or their intrinsic properties. For example, some approaches perform very well only on faces with neutral expression, whereas other approaches try to address the problem of expression variations. An additional measure of the robustness of the 3D model is its sensitivity to size and pose disparities. This is due to the fact that the distance between the target and the camera can affect the size of the facial surface, as well as its height and depth.

3D Mesh-based surface representation is a popular facial representation strategy used in existing 3D face recognition techniques. In contrast to image-based representations, mesh-based surface representations use a spatially dense discrete sampling across the whole surface, resulting in a 3D point cloud representation of the face. These 3D points can be connected into small polygons resulting in a mesh or wireframe representation of the face. For facial comparison purposes, automated resampling of the facial surface is required to generate consistent and corresponding points. This would be an impossible task manually due to the 1000s of points describing every face. The recognition methods that work directly on 3D point clouds consider the data in their original representation based on spatial and depth information. Point clouds are not properly located on a regular grid therefore a prior registration of the point clouds is usually required. For this purpose, the ICP is the most widely used approach [6]. The classification is generally based on the Hausdorff distance that permits to measure the similarity between different point clouds. Chang et al. [7, 9] register overlapping face regions independently by using an ICP-based multi-region approach. Alternatively, recognition could be performed with “3D Eigen faces” that are constructed directly from the 3D point clouds. Another option is to extract geometrical cues based on Eigen values and singular values of local covariance matrices defined on the neighborhood of each 3D point [7]. The main drawback of the recognition methods based on 3D point clouds however resides in their high computational complexity that is driven by the large size of the data. Spherical representations have been used recently for modeling illumination variations [2, 12-13] or both illumination and pose variations in face images. Spherical representations permit to efficiently represent facial surfaces and overcome the limitations of other methods towards occlusions and partial views. To the best of our knowledge, the representation of 3D face point clouds as spherical signals for face recognition has however not been investigated yet. We therefore propose to take benefit of the spherical representations in order to build an effective and automatic 3D face recognition system.

III. 3D Preprocessing And Registartion

In this paper, each 3D face is described by a three-dimensional surface mesh representing the visible face surface from the scanner viewpoint. In this section, we describe how the original 3D data are preprocessed. The preprocessing of the 3D face images includes image smoothing and noise removal, nose tip detection, hole filling and the registration of the face surface.

A. Image smoothing

Image smoothing is an essential preprocessing stage that significantly affects the success of any image processing application. The main purpose of image smoothing is to reduce undesirable distortions and noise while preserving important features such as discontinuities, edges, corners and texture. Over the last two decades diffusion-based filters have become a powerful and well-developed tool extensively used for image smoothing and multi-scale image analysis. The formulation of the multi-scale description of images and signals in terms of scale-space filtering was first proposed by Witkin [14] and Koenderink [15].Their basic idea was to use convolutions with the Gaussian filter to removes small-scale features, while retaining the more significant ones and to generate fine to coarse resolution image descriptions. The diffusion process (also called heat equation or anisotropic diffusion), is equivalent to evolving the input image under a smoothing partial differential equation using the classical heat equation. Since the diffusion coefficient in the partial differential equation (PDE) smoothing techniques is designed to detect edges [16-17], the noise can be removed without blurring the edges of the image.

In this paper we use the graph spectral image smoothing using the heat kernel proposed by Zhang and Hancock in [18] for smoothing the input image. The approach presents a discrete framework for anisotropic diffusion which is based on the heat equation on a graph instead of using diffusion-based PDEs in a continuous domain. The advantage of formulating the problem on a graph is that it requires purely combinatorial operators and as a result no discretization is required therefore the discretization error is eliminated. Graphs are used to represent the arrangement of image pixels where the vertices in the graph correspond to image pixels. Each edge is assigned a real-valued weight, computed using Gaussian weighted distances between local neighboring windows. This weight corresponds to the diffusivity of the edge. To encode the image structure by a graph without losing information, a function is defined to map changes in the image data to edge weights. The Gaussian weighting function is widely used to characterize the relationship between different pixels. If we encode the intensities of the image as a column vector $\vec{\tau}$ via sequential row or column raster ordering of the image pixels then the weight can be calculated as follows:

$$w(i,j) = \begin{cases} e^{\frac{-d^2(i,j)}{k^2}} & if\ \|X(i) - X(j)\|^2 \leq r \\ 0 & otherwise \end{cases} \qquad (1)$$

Where *X(i)* and *X(j)* are the locations of pixels *i* and *j* respectively, *r* is the distance threshold between two neighboring pixels which controls the local connectivity of the graph, and *d(i, j)* = $|\vec{\tau}(i) - \vec{\tau}(j)|$ is the difference between the intensities $\vec{\tau}(i)$ and $\vec{\tau}(j)$ of the two adjacent pixels indexed *i* and *j*. The adjacency weight matrix *W* is then used to compute the Laplacian matrix L as follows:

$$L(i,j) = \begin{cases} T(i,j)\text{-}w(i,j) & if\ i=j \\ \text{-}w(i,j) & if\ e_{ij} \in E \\ 0 & otherwise \end{cases} \qquad (2)$$

Where *T(i ,j)* is a diagonal matrix computed as follow: $T(i,j) = deg(i) = \sum_{j \in V} w(i,j)$. The spectral decomposition of $L=\phi\Lambda\phi^T$, *where* $\Lambda=diag(\lambda_1,\lambda_2,\ldots\ldots, \lambda_{|V|})$ Is the diagonal matrix with the eigenvalues ascending order. $\phi = (\phi_1, \phi_2, \ldots \ldots, \phi_{|V|})$ is the matrix with the corresponding ordered eigenvectors as columns.

In order to use the diffusion process to smooth a gray-scale image, we inject at each node an amount of heat energy equal to the intensity of the associated pixel. The heat initially injected at each node diffuses through the graph edges as time progresses. The edge weight plays the role of thermal conductivity. According to the edge weights determined from (1), if two pixels belong to the same region, then the associated edge weight is large. As a result heat can flow easily between them. The heat kernel H_t is a $|V|\ x\ |V|$ symmetric matrix for nodes *i*, *j* in the graph the resulting heat element is calculated as follows:

$$H_t(i,j) = \sum_{k=1}^{|V|} e^{-\lambda_k t}\, \phi_k(i)\phi_k(j) \qquad (3)$$

And the heat equation on the graph can be characterized by the following differential equation:

$$\frac{\partial H_t}{\partial t} = LH_t \qquad (4)$$

The algorithm can also be understood in terms of Fourier analysis, which is a natural tool for image smoothing. An image $\in R^2$ normally contains a mixture of different frequency components. The low frequency components are regarded as the actual image content and the high frequency components as the noise. From the signal processing viewpoint, the approach is an extension of the Fourier analysis to images defined in graphs. This is based on the fact that the classical Fourier analysis of continuous signals is equivalent to the decomposition of the signal into a linear combination of the eigenvectors of the graph Laplacian. The eigenvalues of the Laplacian represent the frequencies of the eigenfunctions. As the frequency component (eigenvalue) increases, then the corresponding eigenvector changes more rapidly from vertex to vertex. This idea has been used for surface mesh smoothing in [19]. The image $\vec{\tau}$ defined on the graph *G* can be decomposed into a linear combination of the eigenvectors of the graph Laplacian L, i.e.

$$\tau = \sum_{k=1}^{|V|} a_k \emptyset_k \qquad (5)$$

To smooth the image using Fourier analysis, the terms associated with the high frequency eigenvectors should be discarded. However, because the Laplacian *L* is very large even for a small image, it is too computationally expensive to calculate all the terms and the associated eigenvectors in (5). An efficient alternative is to estimate the projection of the image onto the subspace spanned by the low frequency eigenvectors, as is the case with most of the low-pass filters. We wish to pass low frequencies, but attenuate the high frequencies. According to the heat kernel, the function e^{-tx}acts as a transfer function of the filter such that $e^{-tx}\approx 1$ for low frequencies, and $e^{-tx} \approx 0$ for high frequencies. Therefore, the graph heat kernel can be regarded as a low-pass filter kernel. Figure 2 shows the face image before and after image smoothing.

B. Nose tip detection

Our 3D face compression and recognition system permits the faces to be freely oriented with respect to the camera plane with the only limitation being that no occlusions to hide the major face features such as the eyes, the nose, etc. Having this imperative advantage of being viewpoint invariant requires the detection of some facial features for proper face alignment. In this research alignment was performed automatically in two levels: coarse and fine. The coarse alignment is based on nose tip detection whereas the fine alignment is attained using the ICP registration algorithm. Nose tip is an important face feature point widely used for alignment due to its distinctive features.

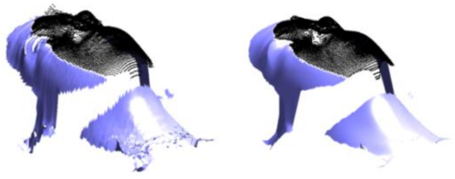


Figure 2. Heat diffusion image smoothing. (a) Input image (b) Image after smoothing using weights from Eq. (1).

The nose is the highest protruding point from the face that is not prone to facial expression. Knowledge of the nose location will enable us to align an unknown 3D face with those in a face database. Besides that, the head pose can be deduced from information obtained from the nose. Therefore nose tip detection is an important part of a 3D face preprocessing [20-25].

Using 2D images, past works have included using luminance values to locate the nose tip [20-25]. This was achieved because the nose tip has a lower luminance value compared to other parts of the face. Besides that, the nostrils are considered as valley regions in a curvature map. However, this method would only work if the face was at a frontal position and looking straight into the camera. Tilted heads and non-frontal faces may cause error in nose detection since the nose tip luminance value might change or the nostrils cannot be detected. Other 2D works include training the computer to detect the nose using Support Vector Machine (SVM) or by using contrast values and edge detection to locate the nose. The drawbacks are SVM has high computational complexity, thus a slow training time while the contrast and edge detection method is affected by expression changes. Using 3D images, one of the methods used was to take horizontal slices of the face and then draw triangles on the slices. The point with the maximum altitude triangle will be considered the nose tip. This method will work for frontal and non-frontal faces. However, for faces tilted to the top, bottom, left or right, errors can occur. This is because in these conditions, the nose tip on the horizontal slices will not be the maximum protruding point. Another method to locate the nose tip from 3D images was proposed by Xu et al. [22]. To locate the nose tip, this method calculates the neighboring effective energy of each pixel to locate suitable nose candidates. It then calculates the neighboring mean and variance of each pixel and then uses SVM to further narrow down the nose tip candidates. Finally, the nose tip is found by choosing the area which has the top three densest nose candidate regions. This method is able to locate nose tip from both frontal and non-frontal faces as well as tilted faces. However, it requires SVM which has high computational complexity.

In this paper the HK curvature analysis is utilized for efficient nose tip detection. To analyze the curvature of 3D faces we let S be the surface defined by a twice differentiable real valued function

$f: U \rightarrow R$ defined on an open set $U \subseteq R^2$

$$S= \{(x, y, z) \,/\, (x,y) \in U, z \in R; f(x,y) =z\} \tag{6}$$

For every point $(x, y, z) \in S$ we consider two curvature measures, the mean curvature (H) and the Gaussian curvature (K) defined as follows:

$$H(x,y) = \frac{(1+f_y^2)f_{xx} - 2f_x f_y f_{xy+}(1+f_x^2)f_{yy}}{2(1+f_x^2+f_y^2)^{3/2}} \tag{7}$$

$$K(x,y) = \frac{f_{xx}f_{yy} - f_{xy}^2}{(1+f_x^2+f_y^2)^2} \tag{8}$$

Where f_x, f_y, f_{xx}, f_{yy}, f_{xy} are the first and second derivatives of $f(x, y)$.

In our system the face image is represented using an NxM range image. Since we have only a discrete representation of S, we must estimate the partial derivatives. For each point $(x_i, y_j\ )$ on the grid we considered a biquadratic polynomial approximation of the surface:

$$g_{ij}\ (x, y) = a_{ij} + b_{ij}\ (x - x_i\ ) + c_{ij}\ (y - y_j\ ) + d_{ij}\ (x - x_i)(y - y_j\ ) + e_{ij}\ (x - x_i\ )^2 + f_{ij}\ (y - y_j\ )^2, i= 1 \ldots N, j = 1 \ldots M \tag{9}$$

The coefficients a_{ij}, b_{ij}, c_{ij}, di_j, e_{ij}, f_{ij} are calculated by least squares fitting of the points in a neighborhood of (x_i, y_j) . The derivatives of f in (x_i, y_j) are then estimated by the derivatives of g_{ij} :

$$f_x(x_i, y_j\ ) = b_{ij},\ \ fy(x_i, yj\ ) = cij,\ \ f_{xy}(x_i, y_j\ ) = d_{ij}\ ,\ \ f_{xx}(x_i, y_j\ ) = 2e_{ij}, f_{yy}(x_i, y_j\ ) = 2fi_j\ . \tag{10}$$

HK classification of the points of the surface is performed to obtain a description of the local behavior of the surface. HK classification was introduced by Besl in 1986 [25]. Image points can be labeled as belonging to a viewpoint-independent surface shape class type based on the signs of the Gaussian and mean curvatures as shown in Table I.

As proposed by Gordon [26] we use the thresholding process to isolate regions of high curvature and to extract the possible feature points that can be utilized in face alignment during the recognition process. The possible extracted feature points are the two inner corners of the eyes and the tip of the nose. Since the calculation of Gaussian curvature involves the second derivative of the surface function, the noise and the artifacts severely affect the final result and applying a prepressing low-pass filter to smooth the data is required. The surface that either has a peak or a pit shape has a positive Gaussian curvature value ($K > 0$). Points with low curvature values are discarded: $|H(u, v)| \geq T_h$, $|K(u, v)| \geq T_k$, where T_h and T_k are predefined thresholds. A nose tip is expected to be a peak ($K > T_K$ and $H > T_H$), a pair of eye cavities to be a pair of pit regions ($K > T_K$ and $H < T_H$) and the nose bridge to be a saddle region ($K < T_K$ and $H > T_H$). These thresholds were experimentally tested to consider a smaller number of cases and reduce the system pipeline overhead, before choosing values similar to those used by Moreno et al. [27] where (T_h=0.04; T_k=0.0005).

TABLE I. SURFACE CLASSIFICATION AND THE CORRESPONDING MEAN (H) AND GAUSSIAN (K) CURVATURES.

	K<0	***K=0***	***k>0***
H<0	Hyperbolic Concave(saddle ridge)	Cylindrical Concave(ridge)	Elliptical Concave(peak)
H=0	Hyperbolic symmetric (minimal)	Planar(flat)	Impossible
H>0	Hyperbolic Convex (saddle valley)	Cylindrical Convex (valley)	Elliptical Convex (pit)

Once the nose tip is successfully determined as the point with maximum z value, we translate it to the origin and align all the face to it. All the points of the face region are located under the nose tip with negative z values. By choosing a proper z- threshold value the face region can be extracted and irrelevant data can be removed such as points corresponding to the hair, neck and shoulders. Figure 3 shows the result of calculating the Gaussian curvature for one of the sample images in the gallery. After localizing the facial area, the portion of the surface below the detected nose tip is projected to a new image to have the face turned upright and where the nose is taken as the origin of the reference system. As can be seen in Figure 3(d) the detected face region contains holes that need to be filled. The Delaunay triangulations algorithm [28] was utilized in this research to fill missing areas in the detected face region and to place them on a regular grid.

C. Face Registration

The nose tip detection phase described above yields an initial raw position and orientation of the face which is very useful for the registration process. Although nose tip detection is sufficient for coarse face alignment, face registration is essential to ensure that all 3D face images have the same pose before the spherical parameterization stage. The registration process typically applies rigid transformations on the 3D faces in order to align them. The ICP algorithm (originally Iterative Closest Point, and sometimes known as Iterative Corresponding Point) proposed by Besl and McKay [29] is a well-known standard algorithm for model registration due to its generic nature and its ease of application. ICP has become the dominant technique for geometric alignment of three-dimensional models when an initial estimate of the relative pose is known. Many variants of ICP have been proposed, optimizing the performance of the different stages of the algorithm such as the selection and matching of points, the weighting of the corresponding point pairs, and the error metric and minimization strategies [30-31]. An excellent survey of the recent variants of the ICP algorithm has been given by Rusinkiewicz and Levoy [30].

ICP starts with two point clouds of data *X* and *Y*, containing, *N* points in R^3 and an initial guess for their relative rigid-body transform. ICP attempts to iteratively refine the transformation *M* consisting of a rotation *R*, and translation *T*, which minimizes the average distance between corresponding closest pairs of corresponding points on the two meshes. At each ICP iteration, for each point $x_i \in X$ for *i= {1...N}*, the closest point, $y_i \in Y$ is found along with the distance, d_N, between the two points.

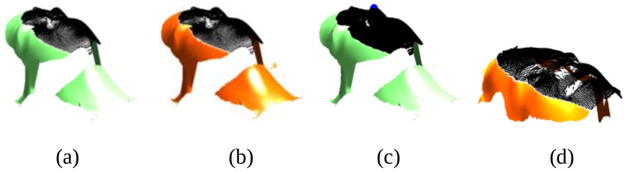


Figure 3. HK classification of face image. (a) Mean curvature (b) Gaussian curvature (c) Nose-tip detection (d) Detected face region.

This is the most time consuming part of the algorithm and has to be implemented efficiently. Robustness is increased by only using pairs of points whose distance are below a predefined threshold. As a result of this first step one obtains a point sequence $Y = (y_1, y_2, \ldots)$ of closest model face points to the data point sequence $X = (x_1, x_2, \ldots)$where each point x_i corresponds to the point y_i with the same index. In the second step, the rigid transformation *M* is computed such that the moved points $M(x_i)$ are moved in a least squares sense as close as possible to their closest points on the model shape y_i, where the objective function to be minimized is:

$$D = \sum_{i=1}^{N} \|M(x_i) - y_i\|^2 \qquad (11)$$

The singular value decomposition of these points is then calculated and rotation/ translation parameters are calculated. After this second step the positions of the data points are updated via $X_{new} = M(X_{old})$. Since the value of the objective function decreases in steps 1 and 2, the ICP algorithm always converges monotonically to a local minimum. This process is repeated either until either the mean square error falls below a predefined threshold or the maximum number of iterations is reached. The generic nature of ICP leads to convergence problems when the initial misalignment of the data sets is large. The impact of this limitation in the ICP process upon facial registration can be counteracted through the use of preprocessing stage that can be used to give a rough estimate of alignment from which we can be confident of convergence. Generating the initial alignment may be done by a variety of methods, such as tracking scanner position, identification and indexing of surface features, “spin-image” surface signatures, computing principal axes of scans, exhaustive search for corresponding points, or user input. In this paper, we assume that a rough initial alignment is always available through the HK curvature analysis performed in the preceding step. Figure 4 presents the face image before and after the ICP registration process.

IV. Spherical Wavelet Parametrization

Wavelets have been a powerful tool in planner image processing since 1985 [1-5, 12, 13, 32-36]. They have been used for various applications such as image compression [1, 2, 5], image enhancement, feature detection [8, 33], and noise removal [32]. Wavelets posse many advantages over other mathematical transforms such as the DFT or DCT as they provide more rigorous mathematical frame work that have the ability of computing accurate and compact representations of functions or data with only a small set of coefficients. Furthermore, wavelets are computationally attractive and they allow variable degrees of detail or resolution to be achieved.

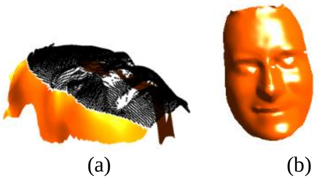


Figure 4. (a) Face image before registration and fill holes (b) Face image after registration and fill holes.

In the signal processing context the wavelet transform is often referred to as sub-band filtering and the resulting coefficients describe the features of the underlying image in a local fashion in both frequency and space making it an ideal choice for sparse approximations of functions. Locality in space follows from their compact support, while locality in frequency follows from their smoothness (decay towards high frequencies) and vanishing moments (decay towards low frequencies). Therefore, 3D wavelet-based object modeling techniques have appeared recently as an attractive tool in the computer However, traditional 2D wavelet methods cannot be directly extended to 3D computer vision environments, possibly for two main reasons: Wavelet representations are not translation invariant [5, 32]. The sensors used in 3D vision provide data in a way which is difficult to analyze with standard wavelet decompositions. Most 3D sensing techniques provide sparse measurements which are irregularly spread over the object’s external surface. This is also important, because sampling irregularity prevents the straightforward extension of 1D or 2D wavelet techniques.

Despite the drawbacks of multi-resolution object representations we believe it have a bright future in 3D computer vision for several reasons [5]. First, the bottom-up scene analysis methods essentially attempt to create hierarchical symbolic representations. Wavelets are excellent for creating hierarchical geometric representations, which can be useful in the image data analysis process. Second, going to 3D implies an important increase in complexity. Wavelet decompositions can provide alternative domains in which many operations can be performed effectively. In this paper we utilize a wavelet transform constructed with the lifting scheme for scalar functions defined on the sphere. Aside from being of theoretical interest, a wavelet construction for the sphere has numerous practical applications since many computational problems are naturally stated on the sphere. Examples from computer graphics include: topography and remote sensing imagery, simulation and modeling of bidirectional reflection distribution functions, illumination algorithms, and the modeling and processing of directional information such as environment maps and view spheres.

A. Spherical Parameterization

Geometric models are often described by closed, genus-zero surfaces, i.e. deformed spheres. For such models, the sphere is the most natural parameterization domain, since it does not require cutting the surface into disk(s). Hence the parameterization process becomes unconstrained [35]. Even though we may subsequently resample the surface signal onto a piecewise continuous domain, these domain boundaries can be determined more conveniently and a posteriori on the sphere. Spherical parameterization proves to be challenging in practice, for two reasons. First, for the algorithm to be robust it must prevent parametric “foldovers” and thus guarantee a 1-to-1 spherical map. Second, while all genus-zero surfaces are in essence sphere-shaped, some can be highly deformed, and creating a parameterization that adequately samples all surface regions is difficult. Once a spherical parameterization is obtained, a number of applications can operate directly on the sphere domain, including shape analysis using spherical harmonics, compression using spherical wavelets [2, 5], and mesh morphing [36].

Given a triangle mesh *M*, the problem of spherical parameterization is to form a continuous invertible map $\varphi: S \rightarrow M$ from the unit sphere to the mesh. The map is specified by assigning each mesh vertex *v* a parameterization $\varphi^{-1}(v) \in S$. Each mesh edge is mapped to a great circle arc, and each mesh triangle is mapped to a spherical triangle bounded by these arcs. To form a continuous parameterization φ, we must define the map within each triangle interior. Let the points {*A*, *B*, *C*} on the sphere be the parameterization of the vertices of a mesh triangle $\{A'= \varphi (A), B'= \varphi (B), C'= \varphi (C)\}$. Given a point $P'= \alpha A'+\beta B'+\gamma C'$ with barycentric coordinates $\alpha+\beta+\gamma=1$ within the mesh triangle, we must define its parameterization $P =\varphi^{-1}(P')$. Any such mapping must have distortion since the spherical triangle is not developable.

B. Geometry Image

A simple way to store a mesh is using a compact 2D geometry images. Geometry images was first introduced by Gu et al. [2, 37] where the geometry of a shape is resampled onto a completely regular structure that captures the geometry as a 2D grid of [*x, y, z*] values. The process involves heuristically cutting open the mesh along an appropriate set of cut paths. The vertices and edges along the cut paths are represented redundantly along the boundary of this disk. This allows the unfolding of the mesh onto a disk-like surface and then the cut surface is parameterized onto the square. Other surface attributes, such as normals and colors, are stored as additional 2D grids, sharing the same domain as the geometry, with grid samples in implicit correspondence, eliminating the need to store a parameterization. Also, the boundary parameterization makes both geometry and textures seamless. The simple 2D grid structure of geometry images is ideally suited for many processing operations. For instance, they can be rendered by traversing the grids sequentially, without expensive memory-gather operations (such as vertex index dereferencing or random-access texture filtering). Geometry images also facilitate compression and level-of-detail control. Figure 5(a)-(d) presents the spherical and geometric representations of the face image.

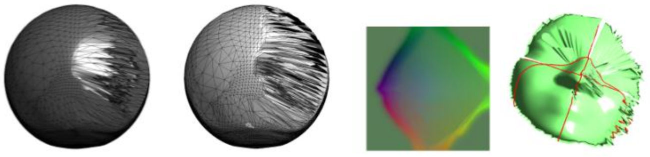

(a) (b) (c) (d)

Figure 5. (a) Initial mapping of face mesh on a sphere (b) Final spherical configuration (c) Geometry image as a color image (d) geometry image as a surface image where the red curves represent the seams in the surface to map it onto a sphere.

C. Wavelet Transform

Haar Transform

Geometry images are regularly sampled 2D images that have three channels, encoding geometric information (*x*, *y* and *z*) components of a vertex in R^3 [37]. Each channel of the geometry image is treated as a separate image for the wavelet analysis. The Haar wavelet transform has been proven effective for image analysis and feature extraction. It represents a signal by localizing it in both time and frequency domains. The Haar wavelet transform is applied separately on each channel creating four sub bands LL, LH, HL, and HH where each sub band has a size equal to 1/4 of the original image. The LL sub band captures the low frequency components in both vertical and horizontal directions of the original image and represents the local averages of the image. Whereas the LH, HL and HH sub bands capture horizontal, vertical and diagonal edges, respectively. In wavelet decomposition, only the LL sub band is used to recursively produce the next level of decomposition. The biometric signature is computed as the concatenation of the Haar wavelet coefficients that were extracted from the three channels of the geometry image.

Spherical Wavelets

To be able to construct spherical wavelets on an arbitrary mesh, this surface mesh should be represented as a multi-resolution mesh, which is obtained by regular 1:4 subdivision of a base mesh [12, 13, 38]. A multi-resolution mesh is created by recursive subdivision of an initial polyhedral mesh so that each triangle is split into four “child” triangles at each new subdivision level. Denoting the set of all vertices on the mesh before the j^{th} subdivision as *K(j)* a set of new vertices

M(j) can be obtained by adding vertices at the midpoint of edges and connecting them with geodesics. Therefore, the complete set of vertices at the $j+1^{th}$ level is given by $K(j+1) = K(j) \cup M(j)$. Consequently, the number of vertices at level j is given by: $10*4^j+2$. This process is presented in Figure 6 (a)-(f) where the face image is shown at 5 different subdivision levels.

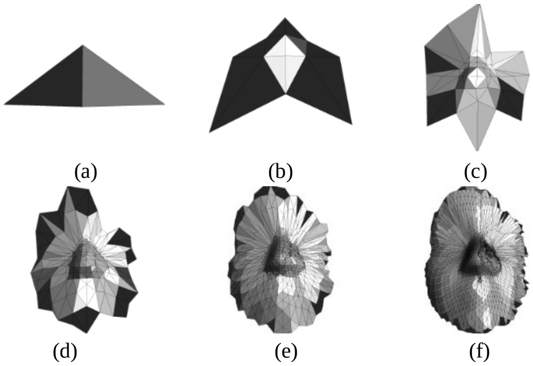


Figure 6. Visualization of recursive partitioning of the face mesh at different subdivision levels. (a) Initial icosahedron (scale 0). (b) Single partitioning of icosahedron (scale 1). (c) Two recursive partitioning of icosahedron (scale 2). (d) Three recursive partitioning of icosahedron (scale 3). (e) Four recursive partitioning of icosahedron (scale 4). (f) five recursive partitioning of icosahedron (scale 5).

In this research, we use the discrete bi-orthogonal spherical wavelets functions defined on a 3-D mesh constructed with the lifting scheme proposed by Schröder and Sweldens [12, 13, 38, 39]. Spherical wavelets belong to second generation wavelets adapted to manifolds with non-regular grids. The main difference with the classical wavelet is that the filter coefficients of second generation wavelets are not the same throughout, but can change locally to reflect the changing nature of the surface and its measure. They maintain the notion that a basis function can be written as a linear combination of basis functions at a finer, more subdivided level. Spherical wavelet basis is composed of functions defined on the sphere that are localized in space and characteristic scales and therefore match a wide range of signal characteristics, from high frequency edges to slowly varying harmonics [38, 40]. The basis is constructed of scaling functions defined at the coarsest scale and wavelet functions defined at subsequent finer scales. If there exist N vertices on the mesh, a total of N basis functions are created, composed of scaling functions and where N_0 is the initial number of vertices before the base mesh is subdivided. An interpolating subdivision scheme is used to construct the scaling functions on the standard unit sphere S denoted by $\varphi_{j,k}$. The function is defined at level j and node $k \in k(j)$ such that the scaling function at level j is a linear combination of the scaling function at level j and $j+1$. Index j specifies the scale of the function and k is a spatial index that specifies where on the surface the function is centered. Using these scaling functions, the wavelet $\psi_{j,m}$ at level j and node $m \in M(j)$ can be constructed by the lifting scheme. A usual shape for the scaling function is a hat function defined to be one at its center and to decay linearly to zero. As the j scale increases, the support of the scaling function decreases. A wavelet function is denoted by $\psi_{j,k}$:S → R. The support of the functions becomes smaller as the scale increases. Together, the coarsest level scaling function and all wavelet scaling functions construct a basis for the function space L^2:

$$L^2 = \{\varphi_{0,k} | k \in N_0\} \cup \{\psi_{j,m} | j \geq 0, m \in N_{j+1}\} \qquad (12)$$

A given function f: $S \rightarrow R$ can be expressed in the basis as a linear combination of the basis functions and coefficients

$$f(x) = \sum_k \lambda_{0,k}\, \varphi_{0,k}(x) + \sum_{0\leq j} \sum_m \gamma_{j,m} \psi_{j,m}(x) \qquad (13)$$

Scaling coefficients $\lambda_{0,k}$ represent the low pass content of the signal f, localized where the associated scaling function has support; whereas, wavelet coefficients $\gamma_{j,m}$ represent localized band pass content of the signal, where the band pass frequency depends on the scale of the associated wavelet function and the localization depends on the support of the function. Figure 7 (a)-(e) presents the spherical wavelets of the face image.

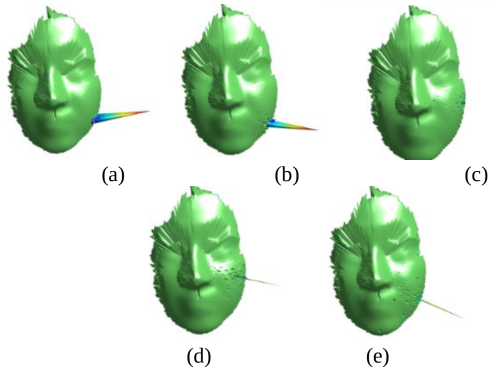


Figure 7. Spherical wavelet transform of face image. (a) using 2% of wavelet coefficients (b) using 5% of wavelet coefficients (c) using 10% of wavelet coefficients (d) using 20% of wavelet coefficients (e) Using all coefficients.

D. *Dimensionality Reduction*

Principal Component Analysis (PCA) [23] is a well-known technique extensively used for dimensionality reduction in computer vision and image recognition. The basic idea of PCA is to find an alternate set of orthonormal basis vectors which best represent the data set. This is to maintain the information content of the original feature space while projecting into a lower dimensionality space more appropriate for modeling and processing. It is possible to use a subset of the new basis vectors to represent the same data with a minimal reconstruction error.

V. Experimental Results

The GAVAB 3D face database [41] was used for the evaluation of the proposed system. GAVAB database contains 549 3D facial surface images corresponding to 61 individuals (45 males and 15 females). Facial surfaces are represented by a mesh of connected 3D points without texture provided by the 3D digitizer VI 700 of Konica-Minolta. Cells of each mesh have four non-coplanar nodes, and sometimes three (in the contour). All subjects in the database are Caucasian with ages between 18 and 40 years. For each individual, there are nine different images containing systematic variations over the pose and facial expression. In particular, for teach subject there are:

two frontal views with neutral expression, two x-rotated views (±30º, looking up and looking down respectively) with neutral expression, two y-rotated views (±90º, left and right profiles respectively) with neutral expression and 3 frontal gesture images (laugh, smile and a random gesture chosen by the user, respectively).

A. 3D Face Compression

The spherical wavelet transform can be used to compress the semi-regular mesh by keeping only the biggest coefficients. Different percentages of the biggest wavelet coefficient were examined and each time the inverse wavelet transform was utilized to reconstruct the approximation face. Figure 8 (a)-(e) shows the reconstructed versions of the original face image using different percentages of wavelet coefficients. As can be seen from Figure 8 there is no visually distinguishable difference between the original image and the corresponding reconstructed images using the various subsets of the wavelet coefficients. The face image can be approximated with a reasonable quality using only 2% of the wavelet coefficients. Figure 9 (a)-(d) presents the original mesh, wavelet transform and the reconstructed images for face images with various pose and facial expressions.

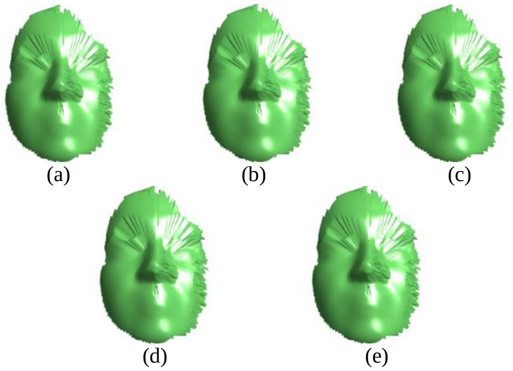
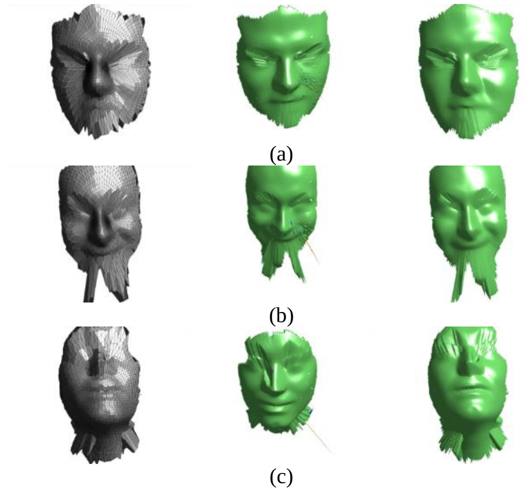


Figure 8. Wavelet approximation of face image. (a) using 2% of wavelet coefficients (b) using 5% of wavelet coefficients(c) using 10% of wavelet coefficients (d) using 20% of wavelet coefficients (e) Using all coefficients.



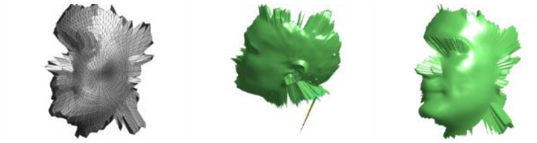


Figure 9. Wavelet approximation of face image for various facial expressions and orientations. (a) Angry expression (b) Laugh expression (c) Looking up (d) Looking left.

The Normalized Error (NE) and Normalized Correlation (NC) were used to evaluate the quality of the reconstructed face image. NE is given as follows:

$$\mathrm{NE}=\frac{\|X-Y\|}{\|X\|} \qquad (14)$$

Where X is the original image and Y is the reconstructed image. i.e. NE is the norm of the difference between the original and reconstructed signals, divided by the norm of the original signal. The NC is given:

$$NC=\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}X(i,j)Y(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{N}X(i,j)X(i,j)} \qquad (15)$$

Where *MxN* is the size of the image. The NE and the NC values of the reconstructed images are presented in Table II for the different wavelet subsets.

TABLE II. NE AND NC FOR VARIOUS WAVELET SUBSETS.

	2%	**5%**	**10%**	**20%**	**100%**
NE	0.67	0.3	0.14	0.06	0
NC	0.9982	0.9997	0.9998	1.0	1.0

The NE and NC values indicate that the reconstructed images are the very similar to the original image. In the case of using only 2% of the wavelets coefficients, the relative error of reconstruction is 0. 67%. The reconstructed signal retains approximately 99.33% of the energy of the original signal.

B. 3D Face Recognition

Two approaches for feature extraction were employed to compare the abilities of the different wavelet transforms applied to the spherical parameterization of the 3D face image. First the spherical wavelet transform is applied to the semi-regular mesh of the face image. For further dimensionality reduction PCA is utilized to reduce the size of the feature vector. Second, the 2-dimentional Haar wavelet transform is applied to each of the three channels of the geometric image. The geometry image regularly samples the face surface and encodes this information on a 2D grid. Each of the *X*, *Y*, and *Z* channels of geometry image are treated as separate images. The concatenation of the Haar wavelet coefficients extracted from the three channels is used as the feature vector (metadata). Each application of the Haar wavelet decomposition reduces the size of the image to 1/4 of its original size. For further data reduction 4-level wavelet decomposition is performed. For example, for the 4-level wavelet decomposition the generated feature vector of 3(8x8) =192 features is the input to the *K*-fold cross validation method.

K-fold cross-validation is a well-known statistical method used to evaluate the performance of a learning algorithm [42]. It outperforms the traditional holdout method that divides the dataset into two fixed non-overlapped subsets: one for training and the other for testing. The major drawback of holdout method is that the results are highly dependent on the choice for the training/test split. Alternatively, in *k*-fold cross validation the data set is partitioned into *k* equally or nearly equal subsets. Subsequently, *k* iterations of the holdout method are performed. In each iteration, one of the *k* subsets is used as the test set and the other *k-1* subsets are put together to form a training set. The average error across all *k* trials is computed. The main advantage of this method is that it is insensitive to how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set *k-1* times. The variance of the resulting estimate is reduced as *k* is increased. The main disadvantage of this method is that the *k* iterations of training are required, which means it takes *k* times as much computation to make an evaluation.

Table III shows the number of the extracted features and the recognition rates for the different feature extraction methods.

TABLE III. RECOGNITION PERFORMANCE FOR VARIOUS FEATURE EXTRACTION METHODS.

Feature Extraction Method	Recognition Rate	Number of Features
Spherical Wavelet +PCA	21%	2000/300
Haar (2-level decomposition)+ PCA	28%	3072/50
Haar (3-level decomposition)+ PCA	31%	768/50
Haar (4-level decomposition)	86%	192

The best average recognition rate of 86% is achieved using the 4-level Haar wavelet decomposition with only 192 features. This is a clear indication that the wavelet features extracted from spherical parameterization are a promising alternative for face recognition. However further research is to be performed to improve the recognition rate.

The performance of the proposed face recognition system based upon the 3D face images of the GAVAB dataset was compared to three different approaches presented by Moreno et al. in [43-45]. In the first approach [43], the range images were segmented into isolated sub-regions using the mean and the Gaussian curvatures. Various facial descriptors such as the areas, the distances, the angles, and the average curvature were extracted from each sub-region. A feature set consisting of 35 best features was selected and utilized for face recognition based on the minimum Euclidean distance classifier. An average recognition rate of 70% was achieved for images with neutral expression and for the images with pose and facial expressions. In the second approach [44], a set of 30 features out of the 86 features was selected and an average recognition rates of 79.1% and 84.03%when the images were classified using PCA and support vector machines (SVM) matching schemas respectively. In the third approach [45], the face images were represented using 3D voxels. An average recognition rate of 84.03% was achieved. Table IV summarizes the results as well as the results obtained from the proposed system.

TABLE IV. COMPARISON OF RECOGNITION RATES FOR VARIOUS 3D FACE RECOGNITION ALGORITHMS BASED ON THE GAVAB DATASET

Technique	Avg. Recognition Rate	Number of Features/ Classifier
Moreno et al. [34]	70%	35/Euclidean Distance
Moreno et al. [15]	79.1% 84.03%	30 features / PCA 30 features / SVM
Moreno et al. [16]	84.03%	3D voxel/ PCA and SVM
Proposed system	86%	192/ K-fold cross-validation

As shown in Table IV, the proposed method based on 4-level Haar wavelet decomposition yields the best recognition rate of 86%. This is a clear indication that the wavelet feature set extracted from spherical parameterization is a promising alternative for 3D face recognition. However further investigation is to be performed to improve the recognition rate.

VI. CONCLUSION

In this paper an innovative approach for 3D face compression and recognition based on spherical wavelet parameterization was proposed. First, we have introduced a fully automatic process for the preprocessing and the registration of facial information in the 3D space. Next, the spherical wavelet features were extracted which provide a compact descriptive biometric signature. Spherical representation of faces permits effective dimensionality reduction through simultaneous approximations. The dimensionality reduction step preserves the geometry information, which leads to high performance matching in the reduced space. Multiple representation features based on spherical wavelet parameterization of the face image were proposed for the 3D face compression and recognition. The GAVAB database was utilized to test the proposed system. Experimental results show that the spherical wavelet coefficients yield excellent compression capabilities with minimal set of features. Furthermore, it was found that Haar wavelet coefficients extracted from the geometry image of the 3D face yield good recognition results that outperform other methods working on the GAVAB database.

REFERENCES

[1] P. Alliez and C. Gotsman, “Recent advances in compression of 3D meshes,” In Proceedings of the Symposium on Mult-iresolution in Geometric Modeling, September 2003.

[2] P. Alliez and C. Gotsman, “Shape compression using spherical geometry images,” In Proceedings of the Symp. Multi-resolution in Geometric Modeling, 2003.

[3] J. Rossignac, 3D mesh compression. Chapter in The Visualization Handbook, C. Johnson and C. Hanson, eds., Academic Press, 2003.

[4] J. Peng, C.-S. Kim, and C.-C. Jay Kuo, “Technologies for 3D mesh compression: A survey,” Journal of visual communication and image representation, Vol. 16, pp. 688–733, 2005.

[5] L. Pastor, A. Rodriguez, J. M. Espadero, and L. Rincon, “3D wavelet-based multi-resolution object representation,” Pattern Recognition, Vol. 34, pp. 2497-2513, 2001.

[6] M. H. Mahoora and M. Abdel-Mottalebb, “Face recognition based on 3D ridge images obtained from range data,” Pattern Recognition, Vol. 42, pp. 445 – 451, 2009.

[7] K.W. Bowyer, K.Chang, and P. Flynn, “A survey of approaches and challenges in 3D and multi-modal3D + 2D face recognition,” Computer Vision and Image Understanding, Vol. 101, No.1, pp. 1-15, 2006.

[8] A.F. Abate, M. Nappi, D. Riccio and G. Sabatino, “2D and 3D face recognition: A survey,” Pattern Recognition Letters, Vol. 28, pp. 1885-1906, 2007.

[9] K. I. Chang, K. W. Bowyer and P. J. Flynn, “Multiple nose region matching for 3D face recognition under varying facial expression,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, pp. 1695-1700, 2006.

[10] G. Günlü and H. S. Bilge, “Face recognition with discriminating 3D DCT coefficients,” The Computer Journal, Vol. 53, No. 8, pp. 1324-1337, 2010.

[11] L. Akarun, B. Gokberk, and A. Salah, “3D face recognition for biometric applications,” In Proceedings of the European Signal Processing Conference, Antalaya, 2005.

[12] P. Schröeder and W. Sweldens, “Spherical wavelets: Efficiently representing functions on a sphere,” In Proceedings of Computer Graphics (SIGGRAPH 95), pp. 161-172, 1995.

[13] P. Schröder and W. Sweldens, “Spherical wavelets: Texture processing,” in Rendering Techniques, New York, 1995, Springer Verlag.

[14] A.Witkin, “Scale-space filtering,” In Proceedings of the International Joint Conference on Artificial Intelligence, pp. 1019-1021, 1983.

[15] J. Koenderink, The structure of images, Biol. Cybern. 50 (1984) pp. 363-370.

[16] G. Sapiro, Geometric partial differential equations and image analysis, Cambridge University Press, Cambridge, 2001.

[17] R. Kondor and J. Lafferty, “Diffusion kernels on graphs and other discrete structures,” In Proceedings of the 19th International Conference on Machine Learning, pp. 315-322, 2002.

[18] F. Zhang and E. R. Hancock, “Graph spectral image smoothing using the heat kernel,” Pattern Recognition, Vol. 41, pp. 3328 - 3342, 2008.

[19] G. Taubin, “A signal processing approach to fair surface design,” In Proceedings of SIGGRAPH, , pp. 351-358, 1995.

[20] A. Colombo, C. Cusano, and R. Schettini, “A 3D face recognition system using curvature-based detection and holistic multimodal classification,” In Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis , pp. 179-184, 2005.

[21] A. Colombo, C. Cusano, and R. Schettini, “3D face detection using curvature analysis,” Pattern Recognition, Vol. 39, pp. 444 – 455, 2006.

[22] C. Xu, T. Tan, Y. Wang, and L. Quan, “Combining local features for robust nose location in 3D facial data," Pattern Recognition Letters, Vol. 27, pp.1487–1494, 2006.

[23] A. Hesher, A. Srivastava, and G. Erlebacher, “Principal component analysis of range images for facial recognition,” In Proceedings of International Conference on Imaging Science, Systems and Technology (CISST 2002).

[24] J. Sergent, “Microgenesis of face perception,” In: H.D. Ellis, M.A. Jeeves, F. Newcombe and A. Young, Editors, Aspects of Face Processing, Nijhoff, Dordrecht (1986).

[25] P. J. Besl and R.C. Jain, “Invariant surface characteristics for 3-d object recognition in range images,” Computer Vision, Graphics Image Process. Vol. 33, pp. 33-80, 1986.

[26] G.G. Gordon, “Face recognition based on depth maps and surface curvature,” SPIE Geom. Methods Computer Vision, Vol.1570,pp. 234-274, 1991.

[27] A. Moreno, A. Sanchez, J. Velez, and F. Diaz, “Face recognition using 3d surface extracted descriptors,” In Proceedings of the Irish Machine Vision and Image Processing, 2004.

[28] D. Cohen-Steiner and J.-M. Morvan, “Restricted delaunay triangulations and normal cycle,” In Proceedings of the nineteenth annual symposium on Computational geometry, pp. 312-321, 2003.

[29] P. J. Besl and N. D. McKay, “A method for registration of 3D shapes,” IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 14, No.2, pp. 239-256, 1992.

[30] S. Rusinkiewicz and M. Levoy, “Efficient variants of the ICP algorithm,” In Proceedings of the. 3rd Int. Conf. on 3D Digital Imaging and Modeling, Quebec, 2001.

[31] J. Cook, V. Chandran, S. Sridharan, and C. Fookes, “Face recognition from 3D data using iterative closest point algorithm and Gaussian mixture models,” In Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization, and Transmission, 2004.

[32] J. Stollnitz, T. D. DeRose, and D. H. Salesin, Wavelets for computer graphics: theory and applications. San Francisco, CA: Morgan Kaufmann, 1996.

[33] Q. M. Tieng and W. W. Boles, “Recognition of 2D object contours using the wavelet transform zero crossing representation,” IEEE Transactions Pattern Analysis and Machine Intelligence, Vol.19, No. 8, pp. 910–916, 1997.

[34] P. Wunsch and A. F. Laine, “Wavelet descriptors for multi resolution recognition of hand printed characters,” Pattern Recognition, Vol. 28 No. 8, pp. 1237–1249, 1995.

[35] Praun and H. Hoppe, “Spherical parameterization and remeshing,” in ACM SIGGRAPH 2003, pp. 340–349, 2003.

[36] M. Alexa, “Recent advances in mesh morphing,” Computer Graphics Forum, Vol. 21, No. 2, pp. 173-196, 2002.

[37] X. Gu, S. J. Gortler and H. Hoppe, “Geometry images,” ACM SIGGRAPH, pp. 355-361, 2002.

[38] P. Schröeder and W. Sweldens, “Spherical wavelets: efficiently representing functions on a sphere,” In Proceedings of Computer Graphics (SIGGRAPH 95), pp. 161-172, 1995.

[39] S. Campagna and H.-P. Seidel, Parameterizing meshes with arbitrary topology. In H.Niemann, H.-P. Seidel, and B. Girod, editors, Image and Multidimensional Signal Processing’ 98, pp. 287-290, 1998.

[40] P. Yu, P. Ellen Grant, Y. Qi, X. Han, F. Ségonne and Rudolph Pienaar, “ Cortical surface shape analysis based on spherical wavelets,” IEEE Transactions on Medical Imaging, Vol. 26, No. 4, pp. 582-597, 2007.

[41] A.B. Moreno and A. Sanchez, “GavabDB: A 3D Face Database,” In Proceedings of 2nd COST Workshop on Biometrics on the Internet: Fundamentals, pp. 77-82, 2004.

[42] A.Blum, A. Kalai, and J. Langford, “Beating the hold-out: bounds for K-fold and progressive cross validation,” In Proceedings of the twelfth annual conference on Computational learning theory, pp. 203-208, 1999.

[43] A.B. Moreno, A. Sanchez, J.F. Velez and F. Dkaz, “Face recognition using 3d surface- extracted descriptors,” In Proceedings of Irish Machine Vision and Image Processing Conference 2003 (IMVIP'03), 2003.

[44] A.B. Moreno, A. Sanchez, J.F. Velez and F.J. Dkaz, “Face recognition using 3d local geometrical features: PCA vs. SVM,” In Proceedings of Fourth International Symposium on Image and Signal Processing and Analysis (ISPA 2005), 2005.

[45] A.B. Moreno, A. Sanchez and J.F. Velez, “Voxel-based 3d face representations for recognition,” in: 12th International Workshop on Systems, Signals and Image Processing (IWSSIP'05), 2005.

AUTHORS PROFILE

Rabab Mostafa Ramadan attended Suez Canal University, Port-Said, Egypt majoring in Computer Engineering, earning the BS degree in 1993. She graduated from Suez Canal University, Port-Said, Egypt with a MS degree in Electrical Engineering in 1999. She joined the Ph.D. program at Suez Canal University, Port-Said, Egypt and earned her Doctor of Philosophy degree in 2004. She worked as an Instructor in the Department of Electric Engineering (Computer division), Faculty of Engineering, Suez Canal University. From 1994 up to1999 , as a lecturer in Department of Electric Engineering (Computer division), Faculty of Engineering , Suez Canal University From 1999 up to 2004, and as an assistant Professor in Department of Electric Engineering(Computer division), Faculty of Engineering , Suez Canal University From 2004 up to 2009. She is currently an Assistant professor in Department of Computer Science, College of Computer & Information Technology, Tabuk University, Tabuk, KSA. Her current research interests include Image Processing, Artificial Intelligence, and Computer Vision.

Rehab Farouk Abdel-Kader attended Suez Canal University, Port-Said, Egypt majoring in Computer Engineering, earning the BS degree in 1996. She graduated from Tuskegee University, Tuskegee, Alabama with a MS degree in Electrical Engineering in 1999. She joined the Ph.D. program at Auburn University and earned her Doctor of Philosophy degree in 2003. She worked as an assistant Professor in the Engineering Studies Program in Georgia Southern University, Statesboro, Georgia from 2003 to 2005. She is currently an Associate professor in the Electrical Engineering department, Faculty of Engineering at Port-Said, Port-Said University, Port-Said, Egypt. Her current research interests include Signal Processing, Artificial Intelligence, and Computer Vision

A Multi-Objective Optimization Approach Using Genetic Algorithms for Quick Response to Effects of Variability in Flow Manufacturing

[1]Riham Khalil
Deputy Director, Centre for Manufacturing organization
De Montfort University
Leicester, United Kingdom

[2]David Stockton
Director, Centre for Manufacturing
De Montfort University
Leicester, United Kingdom

[3]Parminder Singh Kang
Research Assistant, Centre for Manufacturing
De Montfort University
Leicester, United Kingdom

[4]Lawrence Manyonge Mukhongo
Research Assistant, Centre for Manufacturing
De Montfort University
Leicester, United Kingdom

Abstract— **This paper exemplifies a framework for development of multi-objective genetic algorithm based job sequencing method by taking account of multiple resource constraints. Along this, Theory of Constraints based Drum-Buffer-Rope methodology has been combined with genetic algorithm to exploit the system constraints. This paper introduces the Drum-Buffer-Rope to exploit the system constraints, which may affect the lead times, throughput and higher inventory holding costs. Multi-Objective genetic algorithm is introduced for job sequence optimization to minimize the lead times and total inventory holding cost, which includes problem encoding, chromosome representation, selection, genetic operators and fitness measurements, where Queuing times and Throughput are used as fitness measures. Along this, paper provides a brief comparison of proposed approach with other optimisation approaches. The algorithm generates a sequence to maximize the throughput and minimize the queuing time on bottleneck/Capacity Constraint Resource (CCR). Finally, Results are analysed to show the improvement by using current research framework.**

Keywords- Synchronous Manufacturing; Drum-Buffer-Rope; Flow Lines; Multi-Objective Optimisation; Job Sequence.

I. Introduction

All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout the proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example.

A. Drum-Buffer-Rope: A TOC based philosophy

The pull production system is evolved as a revolutionary system, which enabled organizations to the meet the uncertain customer demands at lower production cost and higher profits. Organizations can produce high variability/low volume products more efficiently and at lower cost by controlling critical process parameters, such as work-in-progress (WIP), changeover %, buffer sizes etc. For example, according to [1], accurate WIP management can have the huge impact on organizational performance as it can minimize the inventory levels, which can have direct impact on the throughput levels and queuing times.

Achieving and maintaining system efficiency is not a simple task, it needs close monitoring of critical processes, when it involves high product variability and each product has different resource/processing requirements such as setup time, processing time and routings. Otherwise, it can create organization wide devastating effects. For example, inventory level within system cannot be controlled individually; high inventories can appear in front of the bottleneck machine, just like a push system. Along this, bottleneck resource not only accumulates work in front of constrained resource but also it is one of the root cause for other problems such as extended lead times, missing due dates, higher inventory holding cost etc., which contributes towards increased operational cost, decreased profit and customer dissatisfaction. Therefore, bottleneck/Capacity Constraint Resource (CCR) is one of the decisive factors for controlling the production system. This is well emphasized by [1], [2] and [3], "*If the bottleneck resource wasted one hour, it will be equivalent to one hour wasted for whole system"*. Bottleneck can be seen as a resource with limited capacity to satisfy the demand. Also, there can be more than one bottleneck in a system, but only one bottleneck may be the real constraint. In fact, complexity and randomness involved (such as product dependent setup times, variable processing times, machine failure etc.) in actual system makes it harder to control the bottleneck/CCR.

Researchers have proposed numerous tools to support pull production system and process of continuous improvement (CI). Drum-Buffer-Rope (DBR) is one of the vital tools that are used to maintain the system performance by exploiting the capacity constraint resources (CCR) and bottlenecks. DBR is based on the theory of constraints (TOC) production methodology originated by Goldartt in 1980. Researchers have defined TOC implementation in five steps ([4], [5] and [6]);

1) *System constraint(s) identification.*

2) *Exploitation of identified constraint(s).*

3) *Subordinate everything else according to above decision.*

4) *Elevations of system's constraint(s).*

5) *Go to step 1 if any system's constraint is broken.*

DBR approach represents a set of rules to implement the first three steps of TOC. For instance, the constraints can be market demand, plant capacity or material shortages etc. DBR can be used here to improve the organization performance when it is limited by capacity constraint resources (CCR) or bottleneck i.e. identification and exploitation of CCR/bottleneck processes to maximize their utilization ([5] and [7]). CCR/bottleneck utilization limits the organization throughput, it needs to maximize in order to maintain on time deliveries, minimum WIP level and production cost, maximum profit and quality. As lack of material at CCR or underutilization of CCR can have devastating effects on throughput. In summary, the main aim of DBR is to schedule the material flow through the production line to produce according to customer demand by keeping lead time, WIP and production cost as lower as possible [7]. From current research's perspective dynamic market demand can be seen as one of the causes of variability (in terms of processing times, setup times and routings followed by different work types) in the system, as in high variety/low volume manufacturing variable product demand can cause more often setups without an optimal schedule. This can cause larger queues, decreased throughput and bottleneck shifts.

The main focus of DBR theory is to concentrate on bottleneck constraints to achieve the maximum throughput with minimum lead time, operation expenses and inventory. According to [8], DBR system consists of three main elements (TABLE 1);

1) *Exploitation of CCR* (Drum); Bottleneck defines throughput of production system i.e. capacity of production system must be set what a bottleneck can handle.

2) *Protection of CCR from starving* (Time Buffer); Bottleneck should always have work i.e. buffer of jobs should be maintained to accommodate the upstream process interruption. As time wasted onthe bottleneck resource is unrecoverable and can affect throughput of entire system.

3) *A material release schedule (Rope);* Bottleneck processing capacity predicts the arrival of jobs. Jobs should be released only after receiving signal from bottleneck.

This provides one of the major benefits, production accordance to customer demand with a minimal manufacturing lead-time, inventory and production cost. Also, DBR provides ability to maintain flow at high variety and low volume. In Summary, DBR endeavour to achieve the three tasks [8];

1) Very reliable due date performance.

2) Constraint exploitation.

3) Achieving shortest possible response time within imposed limitations by CCR.

B. Multi-Objective Optimization:

Evolutionary computing is a research area within computer science that used for solving combinatorial optimization and complex problems, which they perform base on principles of generic population-based heuristic techniques [9]. Researchers have used various evolutionary optimisation techniques in manufacturing process optimisation; such as [19] has used practical swarm optimisation for flow shop scheduling to minimize the makespan with the limited buffer space. [20] has exemplified the buffer size optimisation using the genetic algorithms in an asynchronous assembly system. The main aim remains to determine the optimal buffer size in order to prevent blocking and waiting for succeeding and proceeding WorkCentre, but proposed method here only considers the single objective i.e. improvement reducing make span might degrade other performance measures. On the other hand, [21] has used variable neighbourhood search approach for flexible job shop scheduling with sequence dependent setup times to minimise the makespan and mean tardiness, where scheduling problem is solved by dividing it into two sub problems i.e. machine selection and sequence assignment. However, proposed algorithm in this research has tested without imposing any constraints on the flow line. Similarly, in the research literature there are other optimisation approaches been used such as ant colony mechanism, chaotic harmony search algorithms, mixed integer goal programming, Makovian analysis, immune algorithms etc. However, most of the approaches are single objective and are not integrated with the simulation model.

Current research has been used genetic algorithms (GA) for optimization process to get the optimal job sequence such that queuing time can be reduced and throughput can be increased. GA's have been applied in wide range of applications. Some of the examples are; Optimization (job shop scheduling), Machine Learning (weather forecasting and prediction of protein structure), Automatic Programming (computer programs evolve for specific task or for other computational structure), Economic Models (development of bidding strategies and emergence of economic markets), Immune System Modelling, Ecological Modelling, Population Genetics Models, Interactions between Evolution and Learning and Social System Models ([9], [10] and [11]). Also, genetic algorithms are always remained as the one of the dominant approach in the optimisation process because the;

1) Adoptability and versatility that almost any problem can be described in GA code.

2) The uncomplicated nature of underlying GA code, as GA mimics the process of natural evolution.

3) Ability to deal with new problems, change in problem definition or change in objective function.

4) Multi-objective optimization (MOO) can be achieved effectively than the traditional techniques.

Fundamentally, genetic algorithms (GA) are the computer programs that mimic the process of biological evolution to solve complex problems and to model evolutionary systems. GA the phenomenon of natural adaptation and this mechanism can be used in evolutionary programming. According to theoretical framework of GA is simply to move from one population of chromosomes to other in order to find an optional solution, where the selection of chromosomes is based upon the genetic operators, known as; crossover, mutation and inversion ([12] and [13]). There are various examples where GA's have proved their effectiveness and efficiency to solve the complex computational problems. For example; algorithm to find a protein structure from large number of amino acids and algorithms to find fluctuation in financial markets. Some of the main advantages of GA's can be listed as ([11], [13] and [14]);

1) GA provides effective use of parallelism i.e. different possibility can be explored simultaneously by using chromosomes.

2) GA as a tool of adaptive programming, where system can maintain its performance level with respect to changing environment.

3) GA provides solution for complex computational problems. For example, creating an artificial intelligent (AI) system from simple rules using bottom up approach, where GA can drive further rules from the simple rules.

GA's are different from traditional optimization tools and based on digital imitation of biological evolution, using basic genetic operators *Selection*, *Crossover*, *Mutation* and *Elitism*. The population comprises a group of chromosomes from which candidates can be selected for the solution of a problem. Initially, a population is generated randomly. The fitness values of the all chromosomes are evaluated by calculating the objective function. A particular group of chromosomes (parents) is selected from the population to generate the offspring by the defined genetic operations and the fitness of the offspring is evaluated in a similar fashion to their parents. Current population is then replaced by newly generated offspring, based on a certain replacement strategy. Such a GA cycle is repeated until a desired termination criterion is reached (for example, a predefined number of generations are produced or objective function has been met). If all goes well throughout this process of simulated evolution, the best chromosome in the final population can become a highly evolved solution to the problem ([11], [13], [14] and [15]).

Current research has focused on multi-objective optimization. The main aim here is to find all the possible trade-offs among the multiple objective functions i.e. finding all the Pareto optimal solutions. Pareto optimal solution can be defines on the basis of domination rule. Researches have exemplified the concept of Pareto optimality based on two domination rules. These can be described as in [16];

A solution "***S1***" is said to be dominate the solution "***S2***" if and only if

1) The solution "S1" is no worse than "S2" in all objectives and,

2) The solution "S1" is strictly better than the solution "S2" in at least one of the objectives.

The domination concept has been used in current research to determine the better solution by combining the multiple objectives using the weighted sum approach. But non-dominance of objective functions has been maintained by generating variable weights for each chromosome [16] and [17]. Current GA implementation can be exemplified as;

1) *Initialization;* Generate an initial random population "*P*" having "m" chromosomes (strings), where "*m*" represents the population size, which can be given as;

Pi = {pi1, pi2, …. , pm-1, pm}, where i=1and i<n

Where "*n*" is the number of generations.

2) *Evaluation;* Evaluate the fitness of each chromosome "*Pi*" against the fitness function "F", where "F" is derived by using weighted sum approach using multiple objectives and weights are generated randomly for each chromosome. Update the tentative set of Pareto optimal solutions and replace the current generation with new population.

3) *Parent Selection;* It emulates the survival-of-the-fittest mechanism in nature. It is expected that a fitter chromosome receives a higher number of offspring and thus has a higher chance of surviving in the subsequent generation. There are many ways to achieve effective selection; including ranking, tournament, and proportionate schemes. Select the pair of chromosomes/chromosome from the current population to take it further to generate new offspring.

4) *Crossover;* A recombination operator is applied to combine subparts of selected pair to produce new offspring. Based on problem complexity, different crossover strategies are proposed, such as single point, multipoint or m-point crossover. However, according to researchers crossover operator to be the determining factor that distinguishes the GA from all other optimization algorithms [10].

5) *Mutation;* It introduces the variation into chromosome to prevent segmentation or premature convergence. Mutation is carried out according to the predefined, which randomly alters the value at specific string position/positions [10].

6) *Elitism;* Elitist strategy where one chromosome or a few of the best chromosomes are copied into the succeeding generation. The elitist strategy may increase the speed of domination of a population by a super chromosome, but on balance it appears to improve the performance [10].

7) *Termination;* Finally, GA can be terminated when stopping condition is satisfied, which can either the population limit "*n*" has been reached or fitness

function has been satisfied. Otherwise go to step 2 for next iteration.

8) The final set of Pareto optimal solutions represents dominated solutions from the each generation and it is up to the decision maker to select a solution according to the selected objectives.

II. Simulation Optimization Using DBR

Current research has opted three phases approach to identify the CCR/Bottleneck resource in selected simulation model and quantify the DBR and GA based multi-objective optimization model. This can be given as;

A. Simulation Modelling

The simulation model has been established using discrete event simulation software Simul8 (Figure I). There are total 240 jobs to complete having 5 work types (TABLE 1). The attributes of simulation models are;

1) Total simulation time was kept equal to the result collection period i.e. 20000 min and simulation warm-up period is kept as 0.

2) Triangular distribution has been used for work entry point to match system closely to real manufacturing environment. Also, inter arrival time has not been changed for different batch sizes.

3) Travelling time between workstations and machine failures are kept as zero. Job loading is kept as first-come first-serve (FCFS) dispatching policy at all stations. This enabled the system to work with the sequence generated using genetic algorithm.

4) Each work type follows a different route, which is defined in job matrix (product routing). Similarly, processing time and setup time with respect to each job and work station is established using job matrix (TABLE 1).

5) Initial buffer sizes are kept as default as set in simulation model, which will allow genetic algorithm to decide the optimal buffer size. This will be implemented at the later stage of research.

B. CCR/Bottleneck Identification

Bottleneck/CCR has been identified by analysis of data from initial runs. It is important to note the in current experiments the inter-arrival time has been kept constant for all batch sizes. A resource is said to be a bottleneck if [4];

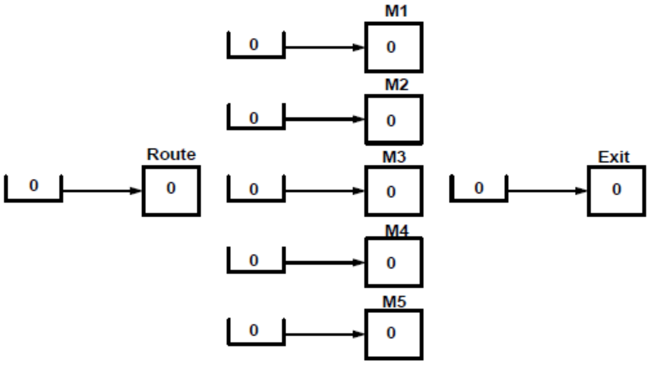
1) Had largest pre-processing queues.

2) Servicing high capacity requirement jobs.

3) Had jobs longest waiting before being processed.

4) Possessed longest cycle time.

These four factors represent a simplistic approach; however in real world, high variety/low volume manufacturing bottleneck can be combination of more than one factor. For example, a bottleneck may not necessarily be the slowest or least capacity operation, but it may be result of combination of more than one factors discussed above or other reasons such as high inter-arrival times, product mix, routings and setups. TABLE 2, shows the bottleneck as machine "***M2***" based on the maximum queue size with in system at any time "***t***", which can related to the physical constraints. The Queue for "***M2***" is always largest than the all other machines. Along this, this argument can be supported by looking at the capacity requirements from the job matrix. "***M2***" is having relatively high capacity requirements then other machines. Similarly, one can argue "***M2***" is bottleneck by considering the logical constraints, such as current policies and procedures involved to process the job. In current scenario, job sequencing can be considered as a logical constraint, for example bad sequencing of jobs effects changeover, which can increase the queuing. Initial results have been collected by using a default sequence generated by program.



[The "route" WorkCentre is there to use the job matrix. The work entry point sets up the "work type" label that is used by job matrix]

Figure 1: Simulation Model

TABLE 1. Job Matrix

Work Type	Job	Location	Timing	Change Over
1	1	M1	5	10
1	2	M2	6	30
1	3	M3	2	10
1	4	M4	3	20
1	5	M5	5	10
1	6	Exit	0	0
2	1	M1	5	20
2	2	M3	3	10
2	4	Exit	0	0
3	1	M1	5	20
3	2	M2	5	30
3	3	M3	4	10
3	4	Exit	0	0
4	1	M2	7	30
4	2	M3	2	15
4	3	M4	3	20
4	4	Exit	0	0
5	1	M2	8	30
5	2	M4	3	10
5	3	M5	4	10
5	4	Exit	0	0

C. Model Optimization

Once the bottleneck/CCR has been identified, the next phase is to improve the system and to make it work near the ideal state (Figure 2). An improved genetic algorithm has been proposed to generate an optimal sequence by using two objective functions i.e. maximizing the throughput and minimizing the queuing length. After running experiments a significant improvement has been shown.

The five primary components of the genetic algorithm used here are;

1) *A chromosomal representation of solutions to the problem i.e. keeping track of job sequence with respect to the work type. The most important point to note here is that chromosome should not lose its integrity in terms of number of jobs encoded when genetic operations are performed.*

2) *Genetic operators that change the composition of the chromosomes.*

3) *A method to initialize a population.*

4) *An evaluation function that represents how well the individual solutions function in the environment, called their "fitness".*

5) *The parameters that are required in order to implement the above components, including population size, number of generations that will be allowed, and stopping criteria.*

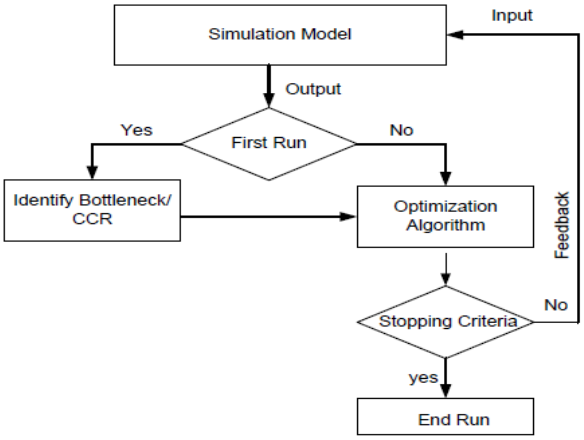


Figure 2. Optimization Model

TABLE 2. BOTTLENECK IDENTIFICATION

EXP. NO.	BATCH SIZE	QUEUE FOR					THROUGHPUT
		M1	M2	M3	M4	M5	
1	20	115	130	46	62	17	240
2	15	96	145	35	66	16	240
3	10	79	135	59	9	35	240
4	5	44	119	42	14	14	240
5	2	9	117	13	4	4	169
6	1	2	96	4	2	2	113

TABLE 3. OPTIMIZED RESULTS

EXP. NO.	BATCH SIZE	QUEUE FOR					THROUGHPUT
		M1	M2	M3	M4	M5	
1	2	15	69	23	10	9	240
2	2	15	65	30	8	6	234
3	1	4	36	10	7	6	234

III. PRELIMINARY RESULTS, DISCUSSION AND CONCLUSSION

From section 2-C, algorithm has shown improvement over the initial results without any optimization. Preliminary results were calculated based on the following parameters TABLE 3;

1) *Batch size was kept 1 and 2 during the optimization process.*

2) *Genetic parameters; population size = 30, Number of generations = 100, Simulation time = 20000 min, No of elite solutions = 2 and crossover, mutation rates are calculated dynamically as solution emerges.*

TABLE 3, shows the preliminary results collected after the optimization process. The results analysis has been shown in the Figure 3. Queue for M2, which was identified bottleneck initially was reduced from 117 to 69 and 65 having throughput 240 and 234 respectively. Similarly, for batch size 1 queue size for M2 has been reduced from 96 to 36 and throughput has been improved from 113 units to 168 units.

In current research a framework has been proposed using drum-buffer-rope and genetic algorithms to achieve identify bottlenecks/CCR, to decide on the optimal buffer sizes and to increase the throughput. It shows an improvement in throughput and noticeable decline in queue sizes. Along this, work is more equally distributed within system (TABLE 3), which will improve system utilization and efficiency. In current implementation, inter-arrival time was kept constant to collect preliminary results. However, algorithm will be improved by adding the facility to deal with the inter-arrival time variations.

Further, improvement will be made on the performance of algorithm and to bring in adaptive inter-arrival times to match exactly with DBR system, as well as the most important factor to determine the optimal buffer sizes and batch sizes to enhance the performance further. As, selecting the appropriate buffer sizes for the flow manufacturing system is a complex task that must account for the random fluctuation in the production rates by individual WorkCentre. If buffer sizes are too large, it can lead to the excessive processing delays and more in-process inventories. On the other hand, if buffer sizes are too small, then small processing delays are small but the small buffer sizes may block upstream WorkCentre from releasing the work.

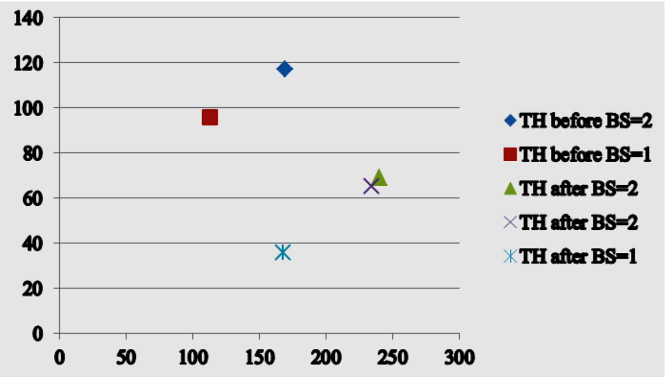


Figure 3. Results Analysis

ACKNOWLEDGMENT

The authors would like to thank TSB – K1532G and De Montfort University for funding the research paper.

REFERENCES

[1] Y. Wang, C. Jun and L. Kong, "Hybrid Kanban/Conwip Control System Simulation and Optimisation Based on Theory of Constraints", IEEE International Conference on Intelligent Computing and Intelligent Systems, November 2009, pp. 666-670.

[2] C. Chen and C. Chen, "A Bottleneck-based Huristic for Minimizing Makespan in a Flexible Flow Line With Unrelated Parallel Machines", Jornal of Computer & Operational Research, 2009, Vol. 36, pp. 3073-3081.

[3] J.D. Blocher, D. Chhajed and M. Leung, "Customer Order Scheduling in a General Job Shop Environment", Jornal of Decision Sciences, 1994, Vol. 29, Iss. 4, pp. 951-980.

[4] M. Umble, E. Umble and S. Murakami, "Implementing Theory of Constraints in a Traditional Japanese Manufactruing Environment: The Case Study of Hitachi Tool Engineeing", International Journal of Prodution Research, 2006, Vol. 44, Iss. 10, pp. 1863-1880.

[5] C. Y. Huang, C. P. Chen, R. K. Li and C. H. Tsai, "Applying Theory of Constraint on Logistic Management in Large Scale Construction Sites: A Case Study of Steel Bar in TFT-LCD Factory Build-up", The Asian journal of Quality, 2008, Vol. 9, Iss. 1, pp. 68 – 93.

[6] K. V. Watson, J.H. Blackstone and S. C. Gardiner, "The Evolution of a management Philosophy: The Theory of Constraints", Journal of Operations Management, 2007, Vol. 25, Iss. 2, pp. 387-402.

[7] H. Wu, C. Chen, C. Tsai and C. Yang, "Simulation and Scheduling Implementation Study of TFT-LCD Cell Plants Using Drum-Buffer-Rope System", Jounal of Expert System with Applications, 2010, Vol. 37, pp. 8127-8133.

[8] J. M. Nicholas, Competitive Manufaturing Management, McGraw-Hill, 1998.

[9] A. Konak, W. C. David and A. E. Smith, "Multi-objective Optimization using Genetic Algorithms: A Tutorial", Journal of Reliability Engineering and System Safety, 2006, Vol. 91, pp. 992 – 1007.

[10] K. S. Tang, K. F. Man, S. Kwong and O. He, "Genetic Algorithms and Their Applications", IEEE Signal Processing Magazine,1996, Vol. 13, Iss. 6, pp. 1053-5888.

[11] A. Norozi, M. K. A. Ariffin and N. Ismail, "Application of Intelligence Based Genetic Algorithm for Job Sequencing Problem on Parallel Mixed-Model Assembly Line", American jounal of Engineering and Applied Sciences, 2010, Vol. 3, Iss. 1, pp. 15-24.

[12] T. Murata, H. Ishibuchi and H. Tanaka, "Genetic Algorithm for Flowshop Scheduling Problems", Jounal of Computer Industrial Engineering, 1996, Vol. 30, No. 4, pp. 1061-1071.

[13] T. Pasupathy, C. Rajendran and R. K. Suresh, "A Multi-Objective Genetic Algorithm for Scheduling in Flow Shops to Minimize the Makespan and Totla Flow Time for Jobs", Industrial Journal of Advanced Manufacturing Technology, 2006, Vol. 27, pp. 804-815.

[14] P. Pongcharoen, C. Hicks ad P.M. Braiden, "The Development of Genetic Algorithms for the Finite Capacity Scheduling of Complex Products with Multiple Levels of Product Structure", European Journal of Operational Research, 2004, Vol. 152, pp. 215-225.

[15] D.E. Goldberg, "Genetic Algorithm in Search, Optimization, and Machine Learning" Addison Wesley Publishing Company, 1989.

[16] T. H. Hou (Tony) and W. C. Hu, "An Integrated MOGA Approach to Determine the Pareto-Optimal Kanban Number and Size for a JIT System", journal of Expert Systems with Applications, 2010, In Press.

[17] J. Jozefowska and A. Zimniak, "Optimization tool for Short-term Production Planning and Scheduling", International Journal of Production Economics, 2008, Vil. 112, pp. 109-120.

[18] T. H. Hou (Tony) and W. C. Hu, "An Integrated MOGA Approach to Determine the Pareto-Optimal Kanban Number and Size for a JIT System", journal of Expert Systems with Applications, 2010.

[19] An effective hybrid PSO-based algorithm for flow shop scheduling with limited buffers, Computers & Operations Research, Volume 35, Issue 9, September 2008, Pages 2791-2806, Bo Liu, Ling Wang, Yi-Hui Jin

[20] Buffer size optimization in asynchronous assembly systems using genetic algorithms, CoTime Series Gene Expression Prediction using

[21] Neural Networks with Hidden Layersmputers & Industrial Engineering, Volume 28, Issue 2, April 1995, Pages 309-322, A.A. Bulgak, P.D. Diwan, B. Inozu

[22] Bi-criteria flexible job-shop scheduling with sequence-dependent setup times—Variable neighborhood search approach, Journal of Manufacturing Systems, Volume 30, Issue 1, January 2011, Pages 8-15, A. Bagheri, M. Zandieh

AUTHORS PROFILE

Riham **Khalil** is a Authors Reader in Manufacturing Science and Deputy Director of Centre for Manufacturing in the Department of Engineering, Faculty of Technology, De Montfort University, United Kingdom. Email: rkhalil@dmu.ac.uk

David **Stockton** is the Head of Commercial Development and Director of Centre for Manufacturing, Professor of Manufacturing Systems Engineering in the Department of Engineering, Faculty of Technology, De Montfort University, United Kingdom. Email: stockton@dmu.ac.uk

Parminder Sing **Kang** is a Research Assistant at the Centre for Manufacturing, Department of Engineering, De Montfort University, United Kingdom. Email: pkang@dmu.ac.uk

Lawrence Mukhongo **Manyonge** is a Research Assistant at the Centre for Manufacturing, Department of Engineering, De Montfort University, United Kingdom. Email: p06249913@myemail.dmu.ac.uk

The Risk Management Strategy of Applying Cloud Computing

Chiang Ku Fan
Department of Risk Management and Insurance
Shih Chien University.
No.70, Dazhi St., Zhongshan Dist., Taipei City 104,
Taiwan (R.O.C.)

Tien-Chun Chen
Department of Risk Management and Insurance
Shih Chien University.
No.70, Dazhi St., Zhongshan Dist., Taipei City 104,
Taiwan (R.O.C.)

Abstract— **It is inevitable that Cloud Computing will trigger off some loss exposures. Unfortunately, not much of scientific and objective researches had been focused on the identification and evaluation of loss exposures stemming from applications of Cloud Computing. In order to fill this research gap, this study attempts to identify and analyze loss exposures of Cloud Computing by scientific and objective methods which provide the necessary information to administrators in support of decisions of risk management. In conclusion, this study has identified "Social Engineering", "Cross-Cloud Compatibility" and "Mistakes are made by employees intentionally or accidentally" are high priority risks to be treated. The findings also revealed that people who work in the field of information or Cloud Computing are somehow ignorant of where the risks in Cloud Computing lie due to its novelty and complication.**

Keywords- Cloud Computing; Risk Assessment; Risk Management; Insurance.

I. Introduction

Since 1980's, the functions of Personal Computer (PC) have indicated that their capabilities have been widely developed to serve human beings with all kinds of daily works. After that, networks have reached every single corner on this planet. During the Past decade, there are more than 2 billion network users now around the world, 5 times of the number of year 2000. The quality and quantity of PC fall behind the pace of fast growing network and PC users. This was the reason why traditional functionality of computer could no longer satisfy PC users and scientists. As a result, engineers strived to devise new technologies to meet different needs all over the world.

Fortunately, Cloud Computing system, a revolutionary architecture of computer system, has been emerged in recent years. Statistics data shows that 66% of USB sticks are lost and around 60% of those lost contain commercial data. Feigenbaum said [1], the enterprise security director of Google, stated that data is typically lost when laptops and Universal Serial Bus(USB) flash drives are lost or stolen, however, local storage is no longer necessary if a company uses cloud-based apps.

This new development has brought computer users' interests back to the information technology. Thanks to the rapidly increased popularity, Cloud Computing services are destined to be the next generation of information technology. Cloud computing providers offer individuals, enterprises and government agencies a variety of services that allow users to apply Cloud Computing for saving and sharing information, database management, data mining as well as their far-reaching web services ranging from mega datasets processing for complex scientific problems to utilizing clouds to administrate and supply access to certain records [2]. In other words, Cloud Computing, which yields highly scalable computing application, storage and platforms, is playing more and more important role through-out business information technology strategy [3]. The computing utility, like all other four existing utilities: water, gas, telephony, and electricity, will provide the basic level of computing service that is considered essential to meet the everyday needs of the general community [4].

While it is true that Cloud Computing services are a modern trend and around 75% of companies and public sectors intend to reallocate or increase their budgets to finance secure Cloud Computing and "Software as a Service" (SaaS) according to some surveys conducted within 2010, however, certain concerns about Cloud Computing and services do exist nowadays. For example, International Data Corporation's (IDC) report shows that 30% of respondents were seeking data security and non-stop support from their cloud providers. Moreover, issues regarding reliability, security, availability, privacy, performance and the management of service level agreements of software services are deeply concerned by the users in the cloud [5][6][7]. In addition, the Chief Information Security officers (CISOs) pointed out that their particular concerns are about the lack of standards for working in the cloud, SaaS, and the secure internet access. Because lack of standards not only makes companies unable to back up their data from one Cloud Computing service providers to another, but also makes it difficult to handle the service interruption of Cloud providers .

We are never short of stories about Cloud Computing service interruptions. For example, Amazon put their users out of service for six hours in February 2008 while their Simple Storage Service (S3) and Elastic Compute Cloud (EC2) suffered a three hours outage. In July, the same year, an eight hours outage was again caused by Amazon's S3 [8]. Google's Webmail service "Gmail" went down for three hours in early 2009, thus prevented its 113 million users from accessing their emails or documents they stored online as "Google Docs" [9].

Base on the discussion above, taking the advantages of Cloud Computing may obviously an ideal solution that leads to both cost-efficiency and flexibility. However, it is inevitable that Cloud Computing will trigger off some loss exposures need to be treated. Unfortunately, there were rare scientific and objective researches focused on identifying and evaluating the loss exposures from applications of Cloud Computing. Insurers or enterprises have only limited information to refer to when they attempt to plan an appropriate risk management program. In order to fill the blank with regard to the research on loss exposures identification and evaluation in Cloud Computing services, the purposes of this study are:

1) to identify loss exposures of Cloud Computing services by scientific and objective methods;

2) to measure and analyze the loss exposures with regard to application of Cloud Computing;

3) to provide the necessary information to administrators in support of decision making risk management with regard to employment of Cloud Computing;

4) to support management's authorization of Cloud Computing based on objectively and scientifically risk-focused assessments; and

5) to recommend essential risk management strategies could be employed to control or reduced losses attributable to the application of Cloud Computing.

II. Literature Review

The major purposes of this research are to identify loss exposures of Cloud Computing services by scientific methods and evaluate the loss exposures with regard to application of Cloud Computing. Therefore, this study is going to review the prior literatures related to the definition of Cloud Computing, risk management, and risks of applying Cloud Computing service.

A. *The Definition of Cloud Computing*

Throughout scientific literatures, many different definitions of Cloud Computing can be found. Svantesson and Clarke [10] defined Cloud Computing as typically a technical arrangement under which users store their data on remote servers under the control of other parties, and rely on software applications stored and perhaps executed elsewhere, rather than on their own computers. In other words, the Cloud Computing appears to be a single point of access for all the Information Technology (IT) requests from consumers.

Based on the observations of Knorr & Gruman [11] and Ward & Sipior [12], the essence of Cloud Computing may be an updated version of utility computing which includes virtual servers delivered over internet; while a broader definition encompasses IT resources outside of the firewall including conventional outsourcing. To draw a conclusion from the above definitions, Cloud Computing, obviously, is not a new technology but a new concept and a new business model. Moreover, Cloud Computing is an evolving term that describes the development into something different. Meanwhile, many studies or reports described Cloud Computing. The most common descriptions are “agility”, “scalability”, “availability”, “cost-efficiency” , “elasticity”, “extensibility” [3][13][14][15].

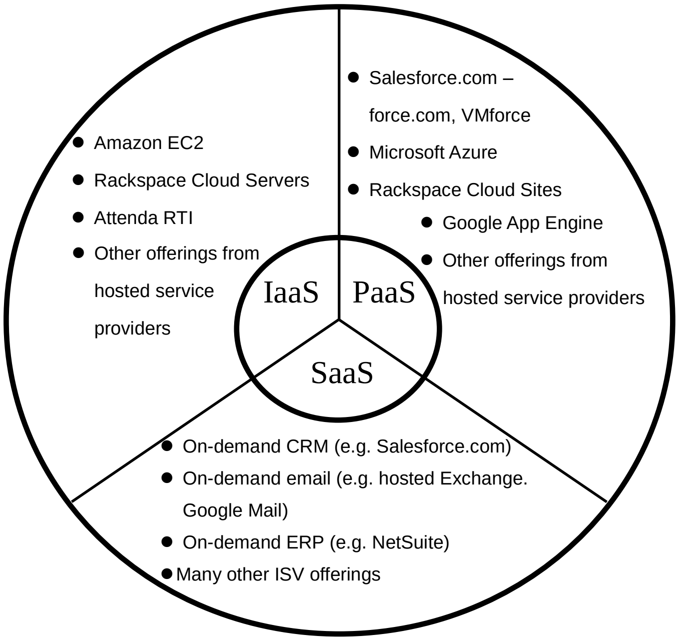

Amazon EC2
Rackspace Cloud Servers
Attenda RTI
Other offerings from hosted service providers
Salesforce.com – force.com, VMforce
Microsoft Azure
Rackspace Cloud Sites
Google App Engine
Other offerings from hosted service providers
IaaS
PaaS
SaaS
On-demand CRM (e.g. Salesforce.com)
On-demand email (e.g. hosted Exchange. Google Mail)
On-demand ERP (e.g. NetSuite)
Many other ISV offerings

Figure 1. Products in Different Cloud Computing Service Level. [16]

There is not much doubt that Cloud Computing improves a company’s ability to flexibly scale services up and down. In detail, Cloud Computing can be classified into three services layers. The lowest level is Infrastructure as a Service (IaaS). This is where pre-configured hardware is provided via a virtualized interface or hypervisor.

There is no high level infrastructure software provided such as an operating system, this must be provided by the buyer embedded with their own virtual applications. Platform as a Service (PaaS) goes a stage further and includes the operating environment included the operating system and application services.

PaaS suits organizations that are committed to a given development environment for a given application but like the idea of someone else maintaining the deployment platform for them. Software as a Service (SaaS) offers fully functional applications on-demand to provide specific services such as email management, Customer Relationship Manage, Enterprise Resource Planning, web conferencing and an increasingly wide range of other applications [12][15][17]. Fig. 1, for examples, shows products of every Cloud Computing service level.

B. *Risk Assessment and Plotting in Risk Management Matrix*

In the research of Marshall and Alexander [18], participants were asked to think of the risks their businesses faced and list these risks. The participants were then asked to evaluate the probability and the consequence of the risks on a scale of 1 to 10, where 1 is low and 10 is high.

The consequence of the risks can be evaluated by terms of severity and cost to the business. Once the participants have rated the probability and consequence of all risks, it can be placed on the risk management matrix shown in Fig. 2.

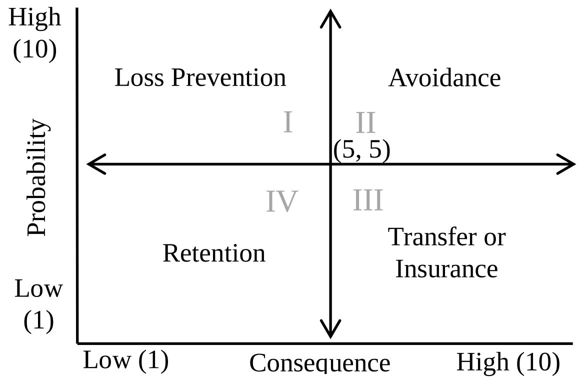


Figure 2. Risk Management Matrix

Many variations of above risk management matrix model are in use. Descriptions of consequence and probability along the two axes may vary [19][20], as many descriptions in particular cells, depending on the context and requirements of the organization using the model. Some versions incorporate references to the organization's decision-making structure [21]. This facilitates assessment of where a particular risk falls in terms of consequence and probability (many other axes such as “frequency” and “severity” or “likelihood” and “impact” are also used, but changing names does not affect the logic.) and helps establish the organizational response to manage the risk [22][23]. Based on the prior studies, the standard and major approach of assessing and characterizing risk is to use matrices which categorize risks by consequence and probability of occurrence. In other words, a two-dimensional risk matrix was usually used to analyze exposures (also see Fig. 2). The consequence (severity) is displayed on the horizontal axis, and the probability (frequency) is displayed on the vertical axis. The resulting four quadrants are with the risk characteristics of high frequency and high severity; high frequency and low severity; low frequency and high severity; and low frequency and low severity. In determining the appropriate technique or techniques for handling losses, a matrix can be used that classifies the various loss exposures according to frequency and severity [24].

There is a widespread belief that the qualitative ranking provided by matrices reflects an underlying quantitative ranking. Typically these matrices are constructed in an intuitive (but arbitrary) manner. Unfortunately, it is impossible to maintain perfect congruence between qualitative (matrix) and quantitative rankings [21].This is essentially due to the impossibility of representing quantitative rankings accurately on a rectangular grid [25]. Moreover, Categorizations of severity cannot be made objectively for uncertain consequences. Inputs to risk matrices (e.g., frequency and severity categorizations) and resulting outputs (e.g., risk ratings) require subjective interpretation, and different users may obtain opposite ratings of the same quantitative risks. Therefore developing an appropriate risk assessment approach may enable risk managers to plot risk on matrices in a more logically sound manner. Fortunately, some studies provided good references which may deal with the problems of tradition of quantitative risk assessment [26][27][28][29]. Their common approach is to employ relative severity and frequency to assess risks while severity and frequency information come from review of the literature and export elicitation.

C. *Risks of Cloud Computing Service*

The sensitive data of each enterprise, which is in a traditional on-premise application deployment model, continues to reside within the enterprise boundary and is subject to its physical logical and personnel security and access control policies [14]. However, the enterprise data is stored outside the enterprise in the most of Cloud Computing service model. Therefore, the Cloud Computing vendor is usually suggested to adopt additional security checks to prevent breaches. This is because malicious users can exploit weakness in the data security model to gain unauthorized access to data. In other words, applying Cloud Computing service has the risk of system vulnerability through malicious employees [30]. Unfortunately, not all security breaches in the Cloud Computing are the fault attributable to the Cloud Computing service provider. Mistakes made by employees intentionally or accidentally are the risk which results in breaches [31]. For example, the use of poor passwords or company’s default password to log on to their network or e-mail platform [30][31].

Utilizing Cloud Computing service, enterprises may get into legal troubles which are caused by the risks of privacy, jurisdiction, and agreement or contract. The cloud infrastructure needs to suffer challengers beyond the traditional issues of remote access, data transfer, and intrusion detection and control through constant system monitoring [3]. The unique schema for physical data storage may well house multiple clients’ data on one physical device. This shared physical server model requires the vendor to ensure that each separate customer’s data remains segregated so that no data bleeding occurs across virtual servers [32]. Further, enterprises and individuals interested in applying Cloud Computing services must ensure they are aware of the privacy risk associated with using the product and take this risk into account when deciding whether to use it [33]. In many cases, vendors’ servers span multiple countries, due to compliance and data privacy laws in various countries, whose jurisdiction the data falls under, when an investigation occurs [3][14]. There is also another law issue raised by applying Cloud Computing between cloud users and cloud provider [32][34]. Such as to sign an unclear delineation of liability in a Cloud Computing service contract, or to get locked into a contractual arrangement that does not cater for the user’s needs.

Besides law issues, cross cloud compatibility is another risk need to be concerned as utilizing Cloud Computing service. An online storage service called The Linkup shut down on August 8, 2008 after losing access as much as 45% of customer data. The Linkup’s 20,000 users were told the service was no longer available and were urged to try out another storage site. In addition to mitigating data lock-in concerns, developing a new generalized usage model in which the same software infrastructure can be used in cross-cloud. Therefore, before developing interoperability technology and improving portability of data and resources between parts of the cloud, the risk of cross cloud compatibility actually is a significant uncertainty that will impact the efficiency of utilizing Cloud Computing service [3].

In practice, there is no specialized policy designed to cover Cloud Computing risks. However, the traditional policies (e.g. Cyber Security Liability Insurance, Cyber breach Insurance, Privacy-Data Breach Insurance, and Network Security and Privacy Insurance) provide partial coverage regard to the risks of applying Cloud Computing such as information, cyber or internet security. Thus, by reviewing coverage or exclusion noted in the policies, some risks related to utilizing Cloud Computing service could be recognized. The risks covered and excluded by policies can be classified into six categories, including legality, system vulnerability, social engineering, administrative or operational mistakes, damage cause by rogue employee, and damage cause by natural disaster.

III. Methodology

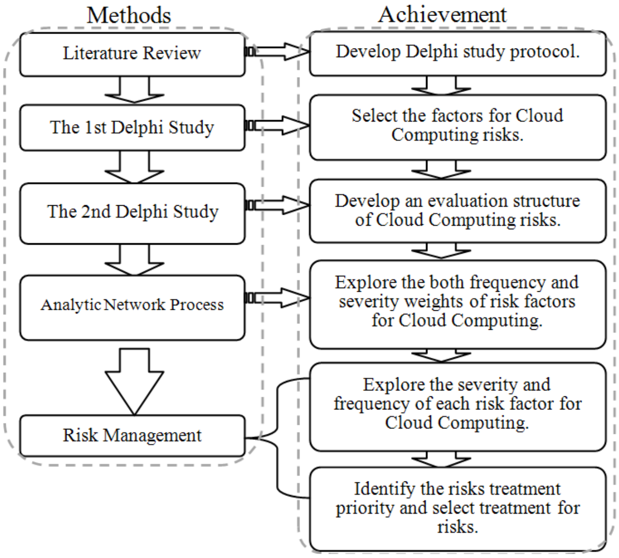

Methods
Achievement
Literature Review
Develop Delphi study protocol.
The 1st Delphi Study
Select the factors for Cloud Computing risks.
The 2nd Delphi Study
Develop an evaluation structure of Cloud Computing risks.
Analytic Network Process
Explore the both frequency and severity weights of risk factors for Cloud Computing.
Risk Management
Explore the severity and frequency of each risk factor for Cloud Computing.
Identify the risks treatment priority and select treatment for risks.

Figure 3. Theoretical Approach Adopted in This Study

The purpose of this paper is to identify and analyze the Cloud Computing related risks. More importantly, combine risk management matrix and ANP's result to provide practical implication. In order to achieve the objectives of this study, the estimation model in this study consists of three phases. In the first phase, the key success factors for Cloud Computing and hierarchical structure of evaluation are identified by using the modified Delphi method. In the second phase, risks' weights of frequency and severity of Cloud Computing are also used as the evaluation criteria and are calculated effectively by employing the "Analytic Network Process" (ANP). In the third phase, the gaps between risks' weights of Cloud Computing and the risk treatment priorities are recognized by using the "Risk Management Matrix". Theoretical approaches adopted herein are described as Fig. 3.

A. Analytic Network Process

After Delphi study, this paper adapts ANP methodology is adapted herein for the propose of identifying risks related to Cloud Computing due to its suitability in offering solutions in a complex multi-criteria decision environment since ANP uses a network without a need to specify levels in hierarchy. This study integrate the process of ANP comprises four major steps [35][36]. But we are not attempt to select the best alternatives in this study.

1) Step 1: Model construction and problem structuring.

The configuration decision problem needs to be stated clearly and structured into its important components. In this case, the relevant criteria is structured in the form of a control hierarchy where the higher the component level. A control hierarchy is simply a hierarchy of criteria and sub-criteria where priorities are derived with respect to the overall goal of the system being analyzed [35]. The highest elements are decomposed into sub-components and attributes. The model development will require the determination of attributes at each level and a definition of their relationships. The model can be obtained by seeking the opinions of the decision makers through brainstorming or other appropriate methods. In this study, the ultimate objective is to determine risks of Cloud Computing.

2) Step 2: Pairwise comparisons matrices between component levels

In this step, elicitation of the decision maker's priorities is completed. The decision maker is asked to respond to a series of pairwise comparisons. In ANP, like AHP, decision elements at each component are compared pairwise with respect to their importance for their control criterion, and the components themselves are also compared pairwise with respect to their contribution to the goal. In the case of interdependencies, components within the same level may be viewed as controlling components for each other, or levels may be interdependent on each other. We leave the interdependencies' evaluation until Step 3.

Saaty [36] has suggested a scale of 1 to 9 when comparing two components. A score of 1 represents the criteria have same importance or indifference where a score of 9 indicates complete dominance to the comparison criteria in a pairwise comparison matrix. If criteria have some level of weaker impact than its comparison criteria the range of the scores will be from 1 to 1/9, where 1 indicates indifference and 1/9 represents an extreme importance by one criterion (row component in the matrix) compared to the other criteria (column component in the matrix). Thus, the value a_{12} for=2, whereas a_{21}=1/2.When scoring is conducted for a pair, a reciprocal value is automatically assigned to the reverse comparison within the matrix. That is, a_{ij} is a matrix value assigned to the relationship of ith element to jth element, then denotes $a_{ij} = 1/a_{ji}$, Once all the pairwise comparisons are complete, the relative importance weight for each component is determined (these results are shown in Table 6 and Table 7). Given that A is the pairwise comparison matrix; the weights can be determined by expression (1).

$$A \cdot w = \lambda_{max} \qquad (1)$$

Where A is the matrix of pairwise comparison, w is the eigenvector or priority vector, and λ_{max} is the largest eigenvalue of A. Saaty [36] provides several algorithms for approximating w. In this study a two-stage algorithm to solve for the largest eigenvalue: the first one is the construction of the network (step 3), and the second one is the calculation of the priorities of the elements (step 4). In order to construct the structure of the problem, all of the interactions among the elements should be considered. This procedure is referred to as the process of averaging over normalized columns. The procedure may be algebraically represented as follows (Formula 2):

$$w_i = \frac{\sum_{i=1}^{I}\left(\frac{a_{ij}}{\sum_{j=1}^{J} a_{ij}}\right)}{I} \qquad (2)$$

where

w_i = the weighted priority for component i,
J = index number of columns (components),
I = index number of rows (components).

Given an initial determinant of risks control hierarchy network, pairwise comparisons need to be made between the applicable attributes within a given risks dimension cluster.

3) Step 3: Pairwise comparisons matrices of interdependencies

To reflect the interdependencies which occur in the network, pairwise comparisons need to be created among all the risk factors of Cloud Computing. We have not included the influence of risks on itself yet. When the elements of a component Y depend on another component X, represent this relation with an arrow from component X to Y. All of these relations are evaluated by pairwise comparisons and a super-matrix, which is a matrix of influence among the elements, is obtained by these priority vectors. The super-matrix is raised to limiting powers to calculate the overall priorities, and thus the cumulative influence of each element on every other element with which it interacts is obtained [37]. If self-controlling linkages are allowed, the graphical representation (which would show in Fig. 7 and Fig. 8) would be a loop from the controlling attribute to itself. The example question asked of the decision maker for evaluating the interdependencies is "When considering risks of Cloud Computing, with regards to increasing robustness, what is the relative impact of criteria A when compared to criteria B?" For example, "When considering risks of Cloud Computing, with regards to improving robustness, what is the relative impact of Hardware when compared to Legality?" For the criteria cluster, this procedure is repeated three times to account for all the applicable risk factors as shown in Fig. 4.

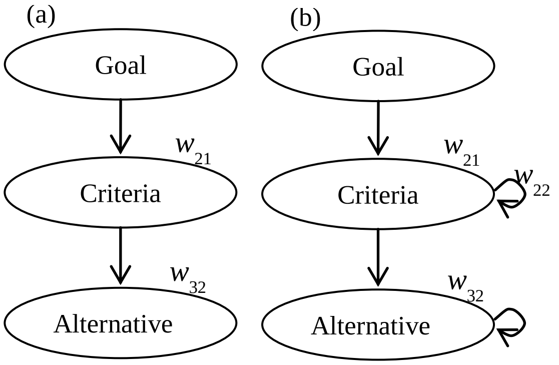


Figure 4. A hierarchy. (b) A nonlinear network.

To obtain global priorities in a system with interdependent influences, the local priority vectors are entered in the appropriate columns of a matrix known as a super-matrix. A super-matrix is actually a partitioned matrix, where each matrix segment represents a relationship between two nodes (components or clusters) in a system [36]. Let the components of a decision system be C_k, k = 1,2,3,…,n and let each component k have m_k elements, denoted by $e_{k1}, e_{k2}, \ldots, e_{kn}$.The local priority vectors obtained in Step 2 are grouped and located in appropriate positions in a super-matrix based on the flow of influence from a component to another component, or from a component to itself, as in the loop. A standard form of a super-matrix is shown in formula (3) [35].

$$W = \begin{array}{cc|ccccc} & & C_1 & \cdots & C_k & \cdots & C_n \\ & & e_{11}\cdots e_{1m_1} & \cdots & e_{k1}\cdots e_{km_k} & \cdots & e_{n1}\cdots e_{nm_n} \\ \hline C_1 & \begin{matrix} e_{11} \\ \vdots \\ e_{1m_1} \end{matrix} & \boldsymbol{w_{11}} & \cdots & \boldsymbol{w_{1k}} & \cdots & \boldsymbol{w_{1n}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ C_k & \begin{matrix} e_{k1} \\ \vdots \\ e_{km_k} \end{matrix} & \boldsymbol{w_{k1}} & \cdots & \boldsymbol{w_{kk}} & \cdots & \boldsymbol{w_{kn}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ C_n & \begin{matrix} e_{n1} \\ \vdots \\ e_{nm_n} \end{matrix} & \boldsymbol{w_{n1}} & \cdots & \boldsymbol{w_{nk}} & \cdots & \boldsymbol{w_{nn}} \end{array} \qquad (3)$$

For example, the super-matrix representation of a hierarchy with three levels is as shown in Fig. 4(a) is as follows (Formula 4) [32]:

$$W_h = \begin{bmatrix} 0 & 0 & 0 \\ w_{21} & 0 & 0 \\ 0 & w_{32} & I \end{bmatrix} \qquad (4)$$

Where w_{21} is a vector that represents the impact of the goal on the criteria; w_{32} is a matrix that represents the impact of criteria on each of the alternatives; I is the identity matrix; and entries of zero correspond to those elements that have no influence. For the above example, if the criteria are interrelated among themselves, the hierarchy is replaced by a network, as shown in Fig. 4(b). The (2, 2) entry of w_n given by w_{22} would indicate the interdependency, and the super-matrix would be as follows (Formula 5):

$$W_n = \begin{bmatrix} 0 & 0 & 0 \\ w_{21} & w_{22} & 0 \\ 0 & w_{32} & I \end{bmatrix} \qquad (5)$$

Note that any zero in the super-matrix can be replaced by a matrix if there is an interrelationship of the elements in a component or between two components. Since there usually is interdependence among clusters in a network, the columns of a super-matrix usually sum to more than 1. The super-matrix must first be transformed to make it stochastic; that is, each column of the matrix sums to unity. An approach recommended by Saaty [34] is to determine the relative importance of the clusters in the super-matrix with the column cluster (block) as the controlling component [33]. That is, the row components with nonzero entries for their blocks in that column block are compared according to their impact on the component of that column block [32]. Through pairwise comparison of the row components with respect to the column component, an eigenvector can be obtained for each column block. For each column block, the first entry of the respective eigenvector is multiplied by all the elements in the first block of that column, the second by all the elements in the second block of that column, and so on. In this way, the block in each column of the super-matrix is weighted. The result is known as the weighted super-matrix, which is stochastic. Raising a matrix to powers gives the long-term relative influences of the elements on each other.

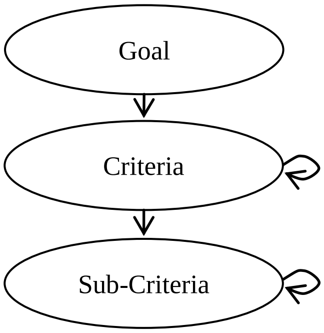


Figure 5. Network form for this paper.

To achieve convergence on the importance weights, the weighted super-matrix is raised to the power of 2*k*+ 1, where *k* is an arbitrarily large number. This new matrix, called the limit super-matrix [32], has the same form as the weighted super-matrix, but all the columns are the same. By normalizing each block of the super-matrix, the final priorities of all the elements in the matrix can be obtained. The network model of this study is described in Fig. 5.

B. *Combine ANP Results with Risk Management Matrix*

This study use ANP method that combines traditional technique of risks treatment and risk management matrix (Frequency/Severity matrix) to measure and understand the big picture of Cloud Computing related risks. First step, obtain relative weight of Cloud Computing risks' frequency and severity separately by running ANP method twice. Second step, put relative weight of risks' frequency and relative weight of risks' severity on the risk management matrix.

IV. RESULTS

A. *Result of Delphi method*

The goal of the first Delphi study is to identify the risk factors for Cloud Computing. Delphi panelists were asked to justify their answers to interview questions and rate their level of agreement toward risk factors, ranging from strongly agree (SA) (5) to strongly disagree (SD) (1).

The interview protocol was developed based on the literature review. The interview explored more fully the perceptions of experts about the risk factors for Cloud Computing. These qualitative responses helped to elaborate the quantitative responses to the standardized questions, and qualitative themes were indicative of opinions raised by a large majority of the Delphi panelists.

Descriptive statistics of attitude toward each key factor at interview were showed as Table 1. In the final round (3rd round), 6 Delphi panelists strongly agreed that "Normal Wear and Tear or Malfunction", "Natural Disaster", "System Vulnerability", and "Social Engineering" are risk factors for Cloud Computing. Moreover 5 Delphi panelists strongly agreed that "Privacy", "Agreement or Contract", "Burglary", and "Cross-Cloud Compatibility" are risk factors for Cloud Computing. There were no undecided (UD) (3), disagree (D) (2) and strongly disagree (SD) (1) answers for key factor item at round 3.

TABLE 1: DESCRIPTIVE STATISTICS OF ATTITUDE TOWARD EACH KEY FACTOR AT INTERVIEW ROUND 2 AND ROUND 3

Key factors of risk	**Attitude toward key factors of risk**									
	SA		***A***		***UD***		***D***		***SD***	
	R2	***R3***	***R2***	***R3***	***R2***	***R3***	***R2***	***R3***	***R2***	***R3***
Agreement or Contract	5	5	1	1	0	0	0	0	0	0
Privacy	4	5	2	1	0	0	0	0	0	0
Jurisdiction	4	4	1	2	1	0	0	0	0	0
Burglary	4	5	2	1	0	0	0	0	0	0
Damaged or spoiled by employees result from intention or accidental	3	5	2	1	1	0	0	0	0	0
Natural Disaster	5	6	1	0	0	0	0	0	0	0
Normal Wear and Tear or Malfunction	4	6	1	0	1	0	0	0	0	0
System Vulnerability	5	6	0	0	1	0	0	0	0	0
Social Engineering	5	6	1	0	0	0	0	0	0	0
Mistakes are made by employees intentionally or accidentally	3	4	3	2	0	0	0	0	0	0
Cross-Cloud Compatibility	3	5	2	1	1	0	0	0	0	0

*Five Attitudes toward Necessary Competencies: Strongly Agree (SA), Agree (A) Undecided (UD), Disagree (D), and Strongly Disagree (SD).

TABLE 2: COMPARISON OF INTERVIEW ROUND 2 AND ROUND 3

Delphi Panelist Attitude toward Each Risk Factor Between R2 and R3	**Z**	**Sig.(2-tailed) (α=0.05)**
Agreement or Contract	0.000	1.000
Privacy	-1.000	0.317
Jurisdiction	-1.000	0.317
Burglary	-1.000	0.317
Damaged or spoiled by employees result from intention or accidental	-1.732	0.083
Natural Disaster	-1.000	0.317
Normal Wear and Tear or Malfunction	-1.342	0.180
System Vulnerability	-1.000	0.317
Social Engineering	-1.000	0.317
Mistakes are made by employees intentionally or accidentally	-1.000	0.317
Cross-Cloud Compatibility	-1.732	0.083

As stated in the methodology chapter, the issues of divergence and convergence of opinion are fundamental to a Delphi study. Based on the result of a Wilcoxon Signed Rank test, no significant attitude difference toward each key success factor was found between R2 and R3. Thus, the 10 items proposed by this study can be identified as risk factors for Cloud Computing.

TABLE 3: DESCRIPTIVE STATISTICS OF THE 3RD ROUND INTERVIEW

Risk Factor	N	Max	Min	Mean	Std. Dev.
Agreement or Contract	6	5	4	4.8	0.4
Privacy	6	5	4	4.8	0.4
Jurisdiction	6	5	4	4.7	0.5
Burglary	6	5	4	4.8	0.4
Damaged or spoiled by employees result from intention or accidental	6	5	4	4.8	0.4
Natural Disaster	6	5	5	5.0	0.0
Normal Wear and Tear or Malfunction	6	5	5	5.0	0.0
System Vulnerability	6	5	5	5.0	0.0
Social Engineering	6	5	5	5.0	0.0
Mistakes are made by employees intentionally or accidentally	6	5	4	4.7	0.5
Cross-Cloud Compatibility	6	5	4	4.8	0.4

As a result, according to Table 2 and 3, all items proposed by this study in first round of Delphi method can be identified as risk factors for Cloud Computing.

The goal of the second Delphi study was to develop an evaluation hierarchical structure for risks of Cloud Computing. Delphi panelists were asked to justify their answers to interview questions and rate their level of agreement toward hierarchical evaluation structure developed by this research (see Fig. 6). These qualitative responses helped to elaborate the quantitative responses to the standardized questions, and qualitative themes were indicative of opinions raised by a large majority of the Delphi panelists.

B. Result of ANP method

The ANP questionnaire was developed based on the result of the second Delphi study and distributed to 6 experts same as the panelists in Delphi studies. The following is general sub-Matrix notation for the risk analysis of Cloud Computing:

$$W= \begin{array}{r} \\ Goal(G) \\ Criteria(C) \\ Sub\text{-}criteria(S) \end{array} \begin{array}{c} \begin{array}{ccc} G & C & S \end{array} \\ \begin{bmatrix} 0 & 0 & 0 \\ w_{21} & w_{22} & 0 \\ 0 & w_{32} & w_{33} \end{bmatrix} \end{array}$$

This study asked experts to figure the relations among criterion as well as sub-criterion. Table 4 represents pairwise comparison and eigenvectors (e-Vector) of the criteria.

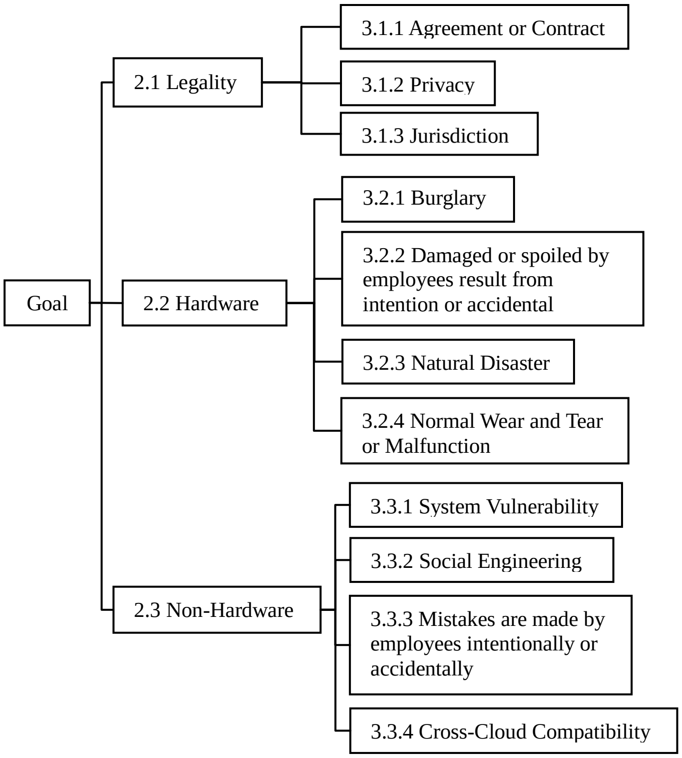

Goal
2.1 Legality
2.2 Hardware
2.3 Non-Hardware
3.1.1 Agreement or Contract
3.1.2 Privacy
3.1.3 Jurisdiction
3.2.1 Burglary
3.2.2 Damaged or spoiled by employees result from intention or accidental
3.2.3 Natural Disaster
3.2.4 Normal Wear and Tear or Malfunction
3.3.1 System Vulnerability
3.3.2 Social Engineering
3.3.3 Mistakes are made by employees intentionally or accidentally
3.3.4 Cross-Cloud Compatibility

Figure 6. Structure of criteria and sub-criteria.

TABLE 4: CRITERIA PAIRWISE COMPARISON MATRIX OF FREQUENCY OF RISK FACTORS

Goal	Legality	Hardware	Non-Hardware	Weights (e-Vector)
Legality	1	1/2	1/3	0.0938
Hardware	2	1	1/4	0.1666
Non-Hardware	3	4	1	0.7396

$$\text{Thus, in Frequency, } W_{21} = \begin{bmatrix} 0.0938 \\ 0.1666 \\ 0.7396 \end{bmatrix}$$

The respective of the three evaluative criteria in frequency (F) are “Legality” (F=0.0938), “Hardware” (F=0.1666) and “Non-Hardware” (F=0.7396). Then, same as criteria’s procedure, this study obtained sub-criteria’s e-Vector shown as table 5.

TABLE 5: PAIRWISE COMPARISON WEIGHTS (E-VECTOR) FOR SUB-CRITERIA

Sub-Criteria	Frequency
Agreement or Contract	0.4934
Privacy	0.3108
Jurisdiction	0.1958
Burglary	0.1936
Damaged or spoiled by employees result from intention or accidental	0.3564
Natural Disaster	0.1243
Normal Wear and Tear or Malfunction	0.3257
System Vulnerability	0.1906
Social Engineering	0.4182
Mistakes are made by employees intentionally or accidentally	0.1205
Cross -Cloud Compatibility	0.2707

The e-Vector for "Legality" ($W_{32\ (Column\ 1)}$), "Hardware" ($W_{32\ (Column\ 2)}$) and "Non-Hardware" ($W_{32\ (Column\ 3)}$) are organized into matrix W_{32}. W_{32} represents the relative importance of sub-criteria with respect to criteria in frequency and severity as follows:

$$\text{In frequency, } W_{32} = \begin{array}{c} \\ 3.1.1 \\ 3.1.2 \\ 3.1.3 \\ 3.2.1 \\ 3.2.2 \\ 3.2.3 \\ 3.2.4 \\ 3.3.1 \\ 3.3.2 \\ 3.3.3 \\ 3.3.4 \end{array} \begin{array}{c} \begin{array}{ccc} \text{Legality} & \text{Hardware} & \text{Non-Hardwa} \end{array} \\ \begin{bmatrix} 0.4934 & 0 & 0 \\ 0.3108 & 0 & 0 \\ 0.1958 & 0 & 0 \\ 0 & 0.1936 & 0 \\ 0 & 0.3564 & 0 \\ 0 & 0.1243 & 0 \\ 0 & 0.3257 & 0 \\ 0 & 0 & 0.1906 \\ 0 & 0 & 0.4182 \\ 0 & 0 & 0.1205 \\ 0 & 0 & 0.2707 \end{bmatrix} \end{array}$$

The inner dependence network maps of criteria and sub-criteria were illustrated by experts as follows. (see Fig. 7 and Fig. 8)

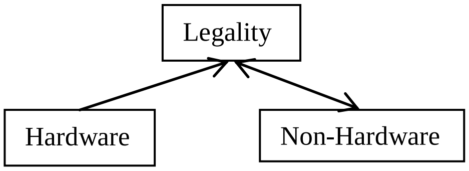


Figure 7. Inner dependence among criteria.

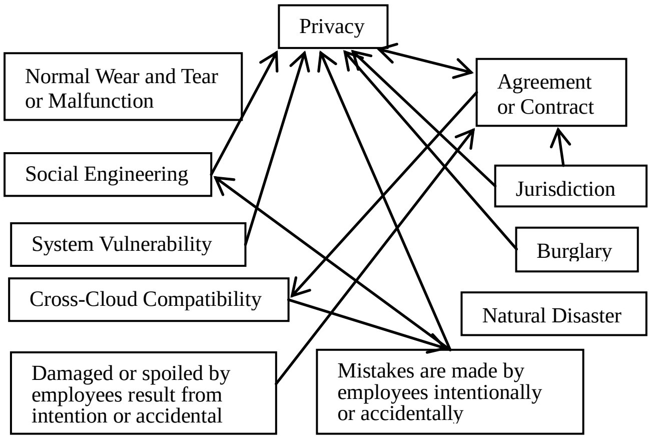

Privacy
Normal Wear and Tear or Malfunction
Agreement or Contract
Social Engineering
Jurisdiction
System Vulnerability
Burglary
Cross-Cloud Compatibility
Natural Disaster
Damaged or spoiled by employees result from intention or accidental
Mistakes are made by employees intentionally or accidentally

Figure 8. Inner dependence among sub-criteria.

After data collection, the weights of risks of Cloud Computing were obtained by ANP. According to the results on Table 6, the highest weight of the criteria, both frequency (F) and severity (S), is "Non-Hardware" (Weight: F=0.60355, S=0.53099). The 2nd is "Legality" (Weight: F=0.25, S=0.40702) and the least important criteria is "Hardware" (Weight: F=0.14645, S=0.06199). Among the criteria of "Legality", "Hardware" and "Non-Hardware", the most important sub-criterion of frequency are: 1st "Cross-Cloud Compatibility" (Weight=0.24803), 2nd "Social Engineering" (Weight=0.17236), 3rd "Mistakes are made by employees intentionally or accidentally" (Weight=0.14337) respectively. The least important sub-criterion of frequency are "Natural Disaster" (Weight=0.01007, Rank=11th), 10th is "Normal Wear and Tear or Malfunction" (Weight=0.0264) and 9th is "Burglary" (Weight=0.04295). Moreover, the most important sub-criterion of severity are "Cross-Cloud Compatibility" (Weight=0.25841), "Mistakes are made by employees intentionally or accidentally" (Weight=0.19201) and "Social Engineering" (Weight=0.09754). The lowest sub-criterion of severity are "Normal Wear and Tear or Malfunction" (Weight=0.00405, Rank=11th), 10th is "Natural Disaster" (Weight=0.01246) and 9th is "Burglary" (Weight=0.03061). (see table 7).

TABLE 6: CRITERION'S RELATIVE WEIGHT OF FREQUENCY AND SEVERITY

Criteria	Frequency (Rank)	Severity (Rank)
2.1 Legality	0.25 (2)	0.40702 (2)
2.2Hardware	0.14645 (3)	0.06199 (3)
2.3 Non-Hardware	0.60355 (1)	0.53099 (1)
Geometric mean	**0.280617152**	**0.237505995**

TABLE 7: SUB-CRITERION'S RELATIVE WEIGHT OF FREQUENCY AND SEVERITY

Sub-Criteria	Frequency (Rank)	Severity (Rank)
3.1.1 Agreement or Contract	0.06947 (7)	0.07497 (7)
3.1.2 Privacy	0.07854 (4)	0.09689 (4)
3.1.3 Jurisdiction	0.06377 (8)	0.09575 (5)
3.2.1 Burglary	0.04295 (9)	0.03061 (9)
3.2.2 Damaged or spoiled by employees result from intention or accidental	0.0723 (6)	0.04421 (8)
3.2.3 Natural Disaster	0.01007 (11)	0.01246 (10)
3.2.4 Normal Wear and Tear or Malfunction	0.0264 (10)	0.00405 (11)
3.3.1 System Vulnerability	0.07274 (5)	0.09311 (6)
3.3.2 Social Engineering	0.17236 (2)	0.09754 (3)
3.3.3 Mistakes are made by employees intentionally or accidentally	0.14337 (3)	0.19201 (2)
Geometric mean	**0.067289379**	**0.057189441**

V. RESULT OF RISK MANAGEMENT

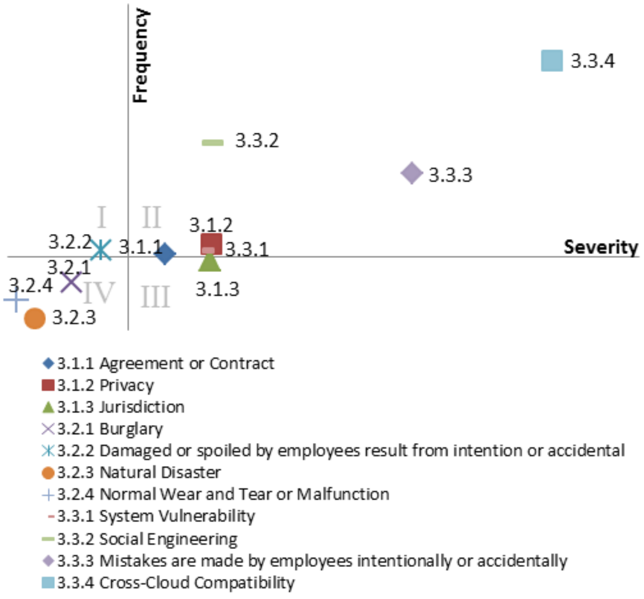


Figure 9. Sub-Criteria Risk Management Matrix

Fig. 9 represents a risk management matrix that illustrates a clear priority of risk management. In Fig. 9, the highest frequency of risk among all the eleven risks is "Cross-Cloud Compatibility" (3.3.4), the second highest frequency is "Social Engineering" (3.3.2), while the third one is "Mistakes are made by employees intentionally or accidentally" (3.3.3). In addition to the above-mentioned risks, "Privacy" (3.1.2), "System Vulnerability" (3.3.1), "Damaged or spoiled by employees result from intention or accidental" (3.2.2) and "agreement or Contract" (3.1.1) are merely greater than the geometric mean. The frequencies of the rest four risks are below geometric mean .From the aspect of severity, however, the sequence of the top three risks is contrary to that of frequency. The top severity is "Cross-Cloud Compatibility" (3.3.4), followed by "Mistakes are made by employees intentionally or accidentally" (3.3.3), then "Social Engineering" (3.3.2). As a result, "agreement or Contract" (3.1.1), "Privacy" (3.1.2), "System Vulnerability" (3.3.1), "Social Engineering" (3.3.2), "Mistakes are made by employees intentionally or accidentally" (3.3.3) and "Cross-Cloud Compatibility" (3.3.4) all fall within quadrant II. It is noteworthy that the severity of "Mistakes are made by employees intentionally or accidentally" (3.3.3) following "Cross-Cloud Compatibility" (3.3.4) by a very insignificant difference. Additionally, risks such as "System Vulnerability" (3.3.1), "Privacy" (3.1.2), "Damaged or spoiled by employees result from intention or accidental" (3.2.2) are all above geometric mean, but only "Damaged or spoiled by employees result from intention or accidental" (3.2.2) in quadrant II. Thus, "Jurisdiction" (3.1.3) fall within quadrant III. "Natural Disaster" (3.2.3), "Burglary" (3.2.1) and "Normal wear and tear or Malfunction" (3.2.4) represent relatively low risks on severity and frequency that lie on quadrant IV.

VI. CONCLUSION AND MANAGEMENT IMPLICATION

The major contribution of this paper lies in the identification and verification of Cloud Computing services' risk factors in which no research has ever been conducted before. ANP method is capable of solving complicated problems. Implementation of ANP method enables decision-makers to visualize the impact of various criteria in the final result as well as measure the severity and frequency of risk of Cloud Computing services. Apply ANP method to evaluate relative weight separately. Then put weight into risk management matrix, a general technique for measuring risk, and proceed to prioritize risks.

As the results of risk management matrix illustrated in Fig. 9 that seven of the criteria, "Agreement or Contract", "Privacy", "System Vulnerability", "Social Engineering", "Mistakes are made by employees intentionally or accidentally" and "Cross-Cloud Compatibility", located in the quadrant II needs to be handled with extra caution. Generally speaking, it is recommended to avoid the risk located in quadrant II. However, neither Cloud Computing users nor providers can escape risk such as "Cross-Cloud Compatibility" or "Social Engineering" in Fig. 9. In this situation, one may try to lower its frequency (which means loss prevention) so that it can be handled by insurance or transfer. This study suggests companies who plan to apply Cloud Computing technique or Cloud Computing service provider treat risk that falls within quadrant II as their first priority. Same reason, criterion in quadrant I, like "Damaged or spoiled by employees result from intention or accidental", requires extra caution as well to prevent loss. But it should be prioritized after quadrant II. Generally, risks in quadrant III can be covered by insurance or transfer, because risks in quadrant III occur higher loss than frequency. Likewise, the same coverage applies to the risk in quadrant III identified in this study (namely "Jurisdiction"). In term of quadrant IV, it is supposed to be the least priority that does not even deserve further processing. All that we need to do is monitor and trace the risk in quadrant IV then respond to any frequency and severity changing/unchanging. This study also found interesting implication that people who work in the information field tend to underestimate some risks. For example, this questionnaire shows that most experts outweighed "Legality" over "Privacy " and "Jurisdiction".

Furthermore, “Privacy” becomes the most severe and frequent risk among all risks in questionnaire. However, after adjust weight of criterion by ANP method base on connections drew in the questionnaire, the consequence reflects differently: “Privacy” becomes minor and “Cross-Cloud compatibility” becomes more important. That is the contribution of this study and also the reason why this study applies ANP method.

This study suggests further researches that focus on specific field or industry such as bank applying Cloud Computing service or insurance company applying Cloud Computing service, to acquire more certain, practical and clearer result.

REFERENCES

[1] Ashford, W. (2009), Cloud computing more secure than traditional IT, says Google. Computer Weekly. Retrieved on Sep. 13, 2011 from http://www.computerweekly.com/Articles/2009/07/21/236982/cloud-computing-more-secure-than-traditional-it-says.htm

[2] Hand, J. D. (2007), Principles of Data Mining, Adis Data Information BV.

[3] Paquette, S.; Jaeger, P. T. and Wilson, S. C. (2010), Identifying the security risks associated with governmental use of Cloud Computing, Government Information Quarterly 27 , p.p. 245-53.

[4] Buyya R. and Parashar M. (2010), User requirements for cloud computing architecture, Proc. 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, Melbourne, Australia, 17-20 May 2010, p.p. 625-30.

[5] Sultan, N. (2011), Reaching for the “cloud”: How SMEs can manage, International Journal of Information Management, 31, p.p.272-8

[6] Stinchcombe, N. (2009), Cloud computing in the spotlight, Retrieved on Nov. 8, 2011 from http://www.infosecurity-magazine.com/view/4755/cloud-computing-in-the-spotlight/

[7] Chow, R.; Golle, P.; Jakobsson, M.; Masuoka, R.; Molina, J.; Shi, E. and Staddon, J. (2009), Controlling data in the Cloud: Outsourcing computation without outsourcing control, CCSW’09, November 13, 2009.

[8] Leavitt, N. (2009), Is Cloud Computing really ready for prime time? Computer, 42(1),p.p.15–20.

[9] Naughton, J. (2009), There's a silver lining to Google's Cloud Computing glitch, Retrieved on Aug. 15, 2011 from http://www.guardian.co.uk/technology/2009/mar/01/gmail-outage-cloud-computing

[10] Svantesson, D. and Clarke, R. (2010), Privacy and consumer risks in Cloud Computing, Computer Law & Security Review, 26, p.p. 391-7.

[11] Knorr, E. and Gruman, G. (2011), What cloud computing really means. Retrieved on Dec. 5, 2011 from http://www.infoworld.com/d/cloud-computing/wht-cloud-computing-really-means-031

[12] Ward, B. T. and Sipior, J. C. (2010), The internet jurisdiction risk of cloud computing, Information Systems Management, 27, p.p. 334-9.

[13] Tisnovsky, R. (2010), Risk versus value in outsourced Cloud computing, Financial Executive, November, p.p. 64-5.

[14] Subashini, S. and Kavitha, V. (2011), A survey on security issues in service delivery models of Cloud computing, Journal of Network and Computer Applications, 34, p.p. 1-11.

[15] Zissis, D. and Lekkas, D. (2011), Addressing cloud computing security issues, Future Generation Computer Systems, 28, p.p. 583-92.

[16] Tarzey, B. (2011), The cloud and the channel, Retrieved on Oct. 5, 2011 from http://www.quocirca.com/media/presentations/032011/572/Quo%20-%20cloud%20for%20CL%20-%20March%204th%202011%20V3.pdf

[17] Lackermair, G. (2010), Hybrid cloud architectures for the online commerce, Procedia Computer Science, 3, p.p.550-5.

[18] Marshall, M. I., & Alexander, C. (2006), Using a contingency plan to combat human resource risk, Journal of Extension , 44(2) Article 2IAW 1. Retrieved on Sep. 22, 2011 from

http://www.joe.org/joe/2006april/iw1.shtml

[19] Federal Aviation Administration (2009), Risk Management Handbook, Retrieved on Dec. 3, 2011 from http://www.faa.gov/library/manuals/aviation/media/FAA-H-8083-2.pdf

[20] GAIA R&D Limited (2010), Project Governance Seminars Workshops & Training, Retrieved on Dec. 20, 2011 from http://www.gaiainvent.com/services.html

[21] Awati, K. (2009), Cox’s risk matrix theorem and its implications for project risk management. Retrieved on Dec 18, 2011 from http://eight2late.wordpress.com/2009/07/01/cox%E2%80%99s-risk-matrix-theorem-and-its-implications-for-project-risk-management/

[22] Sinha, P. R.; Malzahn, D. and Whitman, L. E. (2004), Methodology to mitigate supplier risk in an aerospace supply china, Supply Chain Management: An International Journal, 9 (2), p.p. 154-68.

[23] Mac Crimmon K. R. and Wehrung, D. A. (1986), Taking risks: The management of uncertainty, Free Press, New York.

[24] Rejda, G. E. (2011), Principles of Risk Management and Insurance. 11th Edition, New Jersey: Prentice Hall.

[25] Cox, L. A. (2008), What’s wrong with risk Matrices? Risk Analysis, 28(2), p.p.497-515.

[26] Lim, S.H. (2011), Risks in the north korean special economic zone: Context, identification, and assessment, Emerging Markets Finance & Trade, 47(1), p.p.50-66.

[27] Picado, F.; Barmen, G.; Bengtsson, G. Cuadra, S.; Jakobsson, K.; and Mendoza, A. (2010), Ecological, Groundwater, and Human Health risk assessment in a mining region of nicaragua, Risk Analysis: An International Journal, 30(6), p.p.916-33.

[28] Pintar, K. D. M.; Charron, D. F.;Fazil, A.; McEwen, S. A.; Pollari, F.; Waltner-Toews, D. (2010), A risk assessment model to evaluate the role of fecal contamination in recreational water on the incidence of cryptosporidiosis at the community level in ontario, Risk Analysis: An International Journal, Jan2010, 30(1), p.p.49-64.

[29] Aven, T. and Renn, O. (2009), The role of quantitative risk assessments for characterizing risk and uncertainty and delineating appropriate risk management options, with Special Emphasis on Terrorism Risk, Risk Analysis: An International Journal, 29(4), p.p.587-600.

[30] Casale, J. (2010), Social networking, cloud computing bring new risk exposures, Business Insurance, 9/27/2010, 44(38), p.17.

[31] Bublitz, E. (2010),Catching the Cloud: Managing risk when utilizing Cloud Computing, National Underwriter Property & Casualty November 8, 2010, p.12, p.13, p.16 .

[32] Jaeger, P. T.; Grimes, J. M.; Lin, J. and Simmons, S. N. (2009), Where is the Cloud? Geography, Economics, Environment, and Jurisdiction in Cloud Computing. First Monday, 14(5), p.p.4-15.

[33] Armburst, M.; Fox, A.; Griffith, R.; Joseph, A. D.; Katz, R. and Konwinski, A. et al. (2009), Above the clouds: a Berkley view of Cloud Computing. Retrieved on Dec. 5, 2011 from http://radlab.cs.berkekey.edu/

[34] Saaty T. L. (1996), Decision making with dependence and feedback: The analytic network process, RWS Publications, Pittsburgh.

[35] Meade, L. M. and Sarkis, J. (1999), Analyzing organizational project alternatives for agile manufacturing processes-An analytical network approach, International J. Prod. Res., 37(2), p.p.241-61.

[36] Saaty T. L. (1980), The Analytic Hierarchy Process, McGraw Hill Publications.

[37] Saaty, T. L. and Vargas, L. G. (1998), Diagnosis with dependent symptoms: Bayes theorem and the analytic hierarchy process. Operations Research, 46(4), p.p.491–502.

A Survey on Models and Query Languages for Temporally Annotated RDF

Anastasia Analyti
Institute of Computer Science, FORTH-ICS,
Heraklion, Greece

Ioannis Pachoulakis
Dept. of Applied Informatics & Multimedia, TEI of Crete,
Heraklion, Greece

Abstract— **In this paper, we provide a survey on the models and query languages for temporally annotated RDF. In most of the works, a temporally annotated RDF ontology is essentially a set of RDF triples associated with temporal constraints, where, in the simplest case, a temporal constraint is a validity temporal interval. However, a temporally annotated RDF ontology may also be a set of triples connecting resources with a specific lifespan, where each of these triples is also associated with a validity temporal interval. Further, a temporal RDF ontology may be a set of triples connecting resources as they stand at specific time points. Several query languages for temporally annotated RDF have been proposed, where most of which extend SPARQL or translate to SPARQL. Some of the works provide experimental results while the rest are purely theoretical.**

Keywords- Temporal RDF; provenance; semantics; query languages.

I. Introduction

RDF ("Resource Description Framework") [1], [2] is a growing semantic web standard for the specification of ontologies. An RDF ontology contains a set of triples (s,p,o), denoting that subject s is associated with object o by property p. However, this information is static meaning that either does not change over time or the whole RDF ontology corresponds to a particular time point. However, the truth of statements often changes with time and Semantic Web applications often need to represent such changes and reason about them. For example, statements regarding airline flights are valid only in certain time intervals. Validity time should also be integrated in the query language allowing to retrieve "flights from London to Paris during Mary's summer vacation". Some additional example temporal queries are the following:

1. Who are the *foaf*:*Persons* whose lifespan overlaps with Einstein's?
2. What is the temperature in Chicago at sunrise of July 20th, 2008?
3. What are the names of the engineers who committed code to a particular software in the first half of 2008?
4. What is the salary of Tom during the interval [2007-01-01, 2009-12-31]?
5. Who was the head of the german government before and after the unification of 1990?
6. Who are the service providers that provide web services for more than 4 consecutive years?

Who are the house members who sponsored a bill after April 2, 2008?

In this paper, we provide a survey on the models and query languages for temporally annotated RDF. In most of the works, a temporally annotated RDF ontology is essentially a set of RDF triples associated with temporal constraints, where, in the simplest case, a temporal constraint is a validity temporal interval. However, a temporally annotated RDF ontology may also be a set of triples connecting resources with a specific lifespan, where each of these triples is also associated with a validity temporal interval. Further, a temporal RDF ontology may be a set of triples connecting resources as they stand at specific time points. Several query languages for temporally annotated RDF have been proposed, where most of which extend SPARQL [3] or translate to SPARQL, the most widely accepted query language for RDF. Some of the works provide experimental results while the rest are purely theoretical.

We divide reviewed works into three main categories: (a) works that they have their own model theory (Section 2), (b) works that they extend RDF simple entailment [2] (Section 3), and (c) works that they extend RDFS entailment [2] (Section 4). Works that extend RDF simple entailment are further divided into works that directly translate into RDF and those that do not. Section 5 concludes the paper and provides a comparison of the presented approaches

II. Works With Their Own Model Theory

In [4], a *temporal RDF* (*tRDF* for short) *database* is a set of triples of the form $(s, p{:}\{T\}, o)$, $(s, p{:}{<}n{:}T{>}, o)$, $(s, p{:}[n{:}T], o)$, and (p *rdfs*:subPropertyOf p'), where s is a URI reference from a set U, p,p' are URI references from a set P, o is an entity from $R= U \cup L$, where L is a set of literals, n is a natural number, and T is a temporal interval. Intuitively, the triple $(s, p{:}\{T\}, v)$ indicates that the association (s, p, o) holds at every time point in T, the triple $(s, p{:}{<}n{:}T{>}, o)$ indicates that the association (s, p, o) holds at least n time points within T, and the triple $(s, p{:}[n{:}T], o)$ indicates that the association (s, p, o) holds at most n time points within T. An *interpretation I* of a *tRDF* database is a function from the set of time points to $U \times P \times R$. Satisfaction of a *tRDF* triple is defined in such a way that intuitive meaning is preserved. Obviously, a *tRDF* database may be inconsistent due to the temporal constraints imposed to RDF triples.

A *tRDF* query over a *tRDF* database D is a set of triples of the form $(s, p{:}\{T\}, o)$, $(s, p{:}{<}n{:}T{>}, o)$, $(s, p{:}[n{:}T], o)$, where s,p,o,T are possibly variables, with the constraint that each temporal variable appears only once. An answer to a *tRDF* query q is the set of all possible substitutions to the variables in q such that all triples in q after proper substitutions are entailed by D.

To efficiently answer *tRDF* queries, a *tGRIN* index structure is proposed such that temporally closed resources and resources close in the *tRDF* graph are stored in the same index node. Query answering using the *tGRIN* index is shown to outperform query answering using R^+-trees, *SR*-trees, and the *ST*-index, the most promising representatives of valid-time indexing methods, according to [5].

In [6], the authors extend RDF triples with an annotation from a set A which is a partially ordered set. We consider the case that A is the set of all temporal intervals $[t,t']$, where t,t' are natural numbers. The inclusion ordering $\subseteq$ is the partial ordering in this set. An *annotated RDF theory* (*aRDF-theory* for short) is a finite set of triples (s, p:a, o), where s is a resource from a set R, p is a property from a set P, o is a resource in R, and $a \in A$. In addition, an *aRDF* theory contains statements (*p, rdfs:subProperyOf, p'*), where $p,p' \in P$, and statements indicating which properties are *transitive*.

Let O be an *aRDF* theory, let p be a transitive property in O, and let $r,r' \in R$. Then, there is a *p-path* between r,r' if there exists a set of triples $t_1=(r,p_1{:}a_1,r_1),\ldots,\ t_k=(r_{k-1},p_k{:}a_k,r')$ such that for all $i \in [1,k]$, (p_i *rdfs*:*subPropertyOf* * p). A p-path Q is indicated by the set of triples $\{t_1,\ldots,t_k\}$ that form the path.

An *interpretation I* is a mapping from the set of triples (s,p,o), where $s,o \in R$ and $p \in P$, to A. An interpretation I satisfies ($s,p{:}a,o$) iff $a \subseteq I(s,p,o)$. I satisfies an *aRDF* theory O iff (i) I satisfies every $(s,p{:}a,o) \in O$ and (ii) for all transitive properties $p \in P$, for all p-paths $Q=\{t_1,\ldots,t_k\}$ in O, where $t_i=(r_i,p_i{:}a_i,r_{i+1})$, and for all $a \in A$ such that $a \subseteq a_i$, it is the case that $a \subseteq I(r_1,p,r_{k+1})$, for all $i \in [1,k]$.

A *simple aRDF query q* has the form ($s,p{:}a,o$), where s,p,a,o can be variables. $A_O(q)$ consists of all ground instances of q that are entailed by O. However, $A_O(q)$ may contain redundant triples. For example, if $(a,p{:}[1,100],o) \in A_O(q)$, then there is no point including redundant triples such as $(a,p{:}[1,10],o)$ in it. $Answer_O(q)$ eliminates all redundant triples from $A_O(q)$.

A *conjunctive query Q* is a set of simple *aRDF* queries such that for any simple query $q \in Q$, there is a variable in q that appears in another simple query $q' \in Q$.

The authors present efficient algorithms for simple and conjunctive query answering, showing that the time complexity for answering a conjunctive query is in $O((|R|^2*|P|)^{|Q|})$, where $|Q|$ is the number of simple queries in Q.

The authors also provide experimental results showing the efficiency of their approach.

III. Works That Extend RDF Simple Entailment

A. Approaches that translate to RDF

In [7], instead of having RDF triples associated with their validity temporal interval, named graphs [8] are used both for saving space and for querying the temporal RDF database using standard SPARQL. In particular, each created named graph g is associated with a temporal interval i and all RDF triples whose validity interval is i become members of g (in this process blank nodes are replaced by URIs). The authors introduce through examples a query language, named τ-SPARQL which extends the SPARQL query language for RDF graphs. Each τ-SPARQL query can be translated into a SPARQL query.

A τ-SPARQL query that retrieves all *foaf*:*Persons* whose lifespan overlaps with Einstein's is:

```
SELECT ?s2, ?e2 ?person WHERE {
        [?s1, ?e1] ?einstein foaf:name "Albert Einstein"
        [?s2, ?e2] time:intervalOverlaps [?s1, ?e1]
        [?s2, ?e2] ?person a foaf:Person.}
```

This query is translated into a SPARQL query, as follows:

```
SELECT ?s2, ?e2 ?person WHERE{
        GRAPH ?g1 {?einstein foaf:name "Albert Einstein".}
        ?g2 time:intervalOverlaps ?g1.
        GRAPH ?g2 {?person a foaf:Person.}
        ?g2 time:hasBegining ?s2.
        ?g2 time:hasEnd ?e2.}
```

Temporal relationships between named graphs, such that *time*:*intervalOverlaps* are derived from a temporal reasoning system. Additionally, the authors propose an index structure for time intervals, called *keyTree index*, assuming that triples within named graphs have indices by themselves. The proposed index improves the performance of time point queries over an in-memory ordered list that contains the intervals' start and end times.

Experimental results are provided.

In [9], the *time-annotated RDF* framework is proposed for the representation and management of time-series streaming data. In particular, a TA-RDF graph is a set of triples $<s[t_S], p[t_p], o[t_o]>$, where $<s,p,o>$ is an RDF triple and t_S, t_p, and t_o are time points. In other words, a TA-RDF graph relates streams at certain points in time. To translate a TA-RDF graph into a regular RDF graph, a data stream vocabulary is used, where (i) *dvs*:*belongsTo* is a propery that indicates that a resource is a frame in a stream, (ii) *dvs*:*hasTimestamp* is a property indicating the timestamp of a frame, and (iii) *dvs*:*Nil* is a resource corresponding to the Nil timestamp.

An RDF graph G is the *translation* of a TA-RDF graph G^{TA} iff (B is the set of blank nodes):

$$<s[t_S], p[t_p], o[t_o]> \in G^{TA} \Leftrightarrow \exists\ r_S, r_p, r_o$$

$$[(<r_S, dvs{:}belongsTo, s> \in G \wedge <r_S, dvs{:}hasTimestamp, t_S> \in G \wedge r_s \in B) \vee (t_S = dvs{:}Nil \wedge r_S = s)] \wedge$$
$$[(<r_p, dvs{:}belongsTo, p> \in G \wedge <r_p, dvs{:}hasTimestamp, t_p> \in G \wedge r_p \in B) \vee (t_p = dvs{:}Nil \wedge r_p = p)] \wedge$$
$$[(<r_o, dvs{:}belongsTo,o> \in G \wedge <r_o, dvs{:}hasTimestamp, t_o> \in G \wedge r_o \in B) \vee (t_o = dvs{:}Nil \wedge r_o = o)] \wedge$$
$$< r_S, r_p, r_o> \in G.$$

A query language for the time-annotated RDF, called TA-SPARQL, is proposed which has a formal translation into normal SPARQL. For example, a TA-SPARQL query requesting the temperature in Chicago at sunrise of July 20th, 2008 is:

```
SELECT ?temperature WHERE {
          <urn:OHARE> <urn:hasTemperatureSensor> ?x.
          ?x[“2008-07-20T05:34:00Z”^^xsd:dateTime]
                    <urn:hasReading>  ?temperature.}
```

The above TA-SPARQL query is translated into the SPARQL query:

```
 SELECT ?temperature WHERE {
          <urn:OHARE> <urn:hasTemperatureSensor> ?x.
          ?F <urn:hasReading> ?temperature.
          ?F dvs:belongsTo ?x.
          ?F dvs:hasTimestamp ?F_T.
          FILTER(?FT= “2008-07-20T05:34:00Z”^^xsd:dateTime)}
```

The system has been implemented on top of the Tupelo[1] semantic middleware. However, no experimental results are provided.

In [10], the authors consider temporal RDF graphs which is a set of triples of the form (*s*,*p*:[*start*,*end*],*o*), where (*s*,*p*,*o*) is an RDF triple and *p*:[*start*,*end*] is a shorthand for a URI that identifies a temporal property which has *base property p*, beginning *start* and ending *end*.

The authors define a *simple temporal interpretation* by extending an RDF simple interpretation as follows:

1. *T* is a subset of the set of resources.
2. *NT* is a value representing no time.
3. The set of properties contains *tb*:*property*, *tb*:*begin*, and *tb*:*end*.
4. *BP* is a subset of resources, called the set of *base properties*.
5. *PT* is a mapping from $BP \times (T \cup \{NT\}) \times (T \cup \{NT\})$ into the set of properties.
6. If $tp=PT(bp,t_1,t_2)$ then (i) $(tp,bp) \in$ IEXT(*tb*:*property*) , (ii) if $t_1 \neq NT$ then $(tp,\ t_1) \in$ IEXT(*tb*:*begin*), otherwise IEXT(*tb*:*begin*) contains no pair (*tp*, *t*), for any *t*, and (iii) if $t_2 \neq NT$ then $(tp,\ t_2) \in$ IEXT(*tb*:*end*), otherwise IEXT(*tb*:*end*) contains no pair (*tp*,*t*), for any *t*.

As temporal RDF graphs are ordinary RDF graphs they can be queried using normal SPARQL. However, it is helpful to the writer of temporal queries to provide some extra syntax to enable queries to be written more compactly and to hide the details of the underline representation.

For example, a query asking for the names of the engineers who committed code to a particular software in the first half of 2008 is the following:

```
    SELECT ?name WHERE {
          ?module rdfs:label “module name”.
          ?module f:updatedBy: (?uBegin, ?uEnd) ?person.
          FILTER (tb:intervalsIntersect(?uBegin,?uEnd,
                    “2008-01-01”^^xsd:date,
                    “20008-7-01”^^xsd:date))
          ?person ex:hasName ?name.}
```

This query can be expressed in normal SPARQL, as follows:

```
    SELECT ?name WHERE {
          ?module rdfs:label “module name”.
          ?updatedBy tb:property f:updatedBy.
          ?updatedBy tb:begin ?uBegin.
          ?updatedBy tb:end ?uEnd.
          FILTER (tb:intervalsIntersect(?uBegin,?uEnd,
                              “2008-01-01”^^xsd:date,
                              “20008-7-01”^^xsd:date))
          ?module f:updatedBy ?person.
          ?person ex:hasName ?name.}
```

Though an implementation of a prototype is mentioned, no experimental results are provided.

B. *Other approaches*

In [11], an *N*-dimensional time domain has the form: $\mathcal{T}=T_1 \times \ldots \times T_N$, where each T_i is a set of intervals. A *multi-temporal RDF triple* is defined as (*s*,*p*,*o* | *T*), where <*s*,*p*,*o*> is an RDF triple and $T \subseteq \mathcal{T}$. Note that since *T* is a set, some compression is achieved in the storage of multi-temporal RDF triples.

As a query language, the authors propose T-SPARQL, an extension of SPARQL that has many features of TSQL2 [12] (a query language designed for temporal relational databases). As in TQL2, if *T* is a multi-dimensional time element, the expression VALID(*T*) and TRANSACTION(*T*) can be used to express conditions on the valid and transaction components of *T*.

T-SPARQL is demonstrated through examples and a query that requests the salary of Tom during the interval [2007-01-01,2009-12-31] is the following:

```
    SELECT ?salary INTERSECT(?t, “ [2007-01-01,2009-12-31]”)
    WHERE {
    ?emp rdf:type ex:employee;
                        ex:Name “Tom”;
                        ex:Salary ?salary | ?t.
    FILTER (VALID(?t) OVERLAPS
    “ [2007-01-01,2009-12-31]”^^xs:period)}
```

No implementation of T-SPARQL is provided.

In [13], an *uncertain temporal knowledge base* is a pair *KB* = <*F*, *C*>, where *F* is a set of weighted temporal RDF triples and *C* is a set of first-order temporal consistency constraints. In particular, a fact in *F* has the form: $p(s,o,i)_d$, where $p(s,o)$ is an RDF triple, *i* is a temporal interval, and $d \in [0,1]$ is a confidence degree that $p(s,o)$ is true during interval *i*. Additionally, a temporal consistency constraint in *C* has the form:

$$p_1(?s,?o_1,?i_1) \wedge p_2(?s,?o_2,?i_2) \wedge relA(?o_1,?o_2) \rightarrow rel_T(?i_1,?i_2)$$

or of the form:

$$p_1(?s,?o_1,?i_1) \wedge p_2(?s,?o_2,?i_2) \wedge relA(?o_1,?o_2) \rightarrow false$$

where $?i_1$, and $?i_2$ are temporal interval variables, *relA* is an (optional) arithmetic relation, such as = and ≠, and rel_T is a

[1] Tupelo. http://tupeloproject.ncsa.uiuc.edu

temporal predicate such as *overlap* and *before* (see Allen's temporal relations among intervals [14]).

For example, the fact that a player can only play for one club at a time is expressed by the query:

$$playsForClub(?s,?o_1,?i_1) \wedge playsForClub(?s,?o_2,?i_2) \wedge ?o_1 \neq ?o_2 \rightarrow disjoint(?i_1,?i_2)$$

A query Q is a conjunction of triples $p(s,o)$, where s and o can be variables. To answer a query Q, all matches from the *KB* at collected into a set F_Q. Then, all facts possibly conflicting with them are also added to F_Q. To resolve the conflicts, a consistent subset $F_{Q,C}$ of F_Q is selected such that the sum of the weights of the facts in $F_{Q,C}$ is maximized. Then, the matches to Q within $F_{Q,C}$ are returned as answer to the query. The query answering problem is shown to be NP-hard. A scheduling algorithm for query answering is provided, as well as an efficient approximation algorithm with polynomial performance. Experimental results show the efficiency of the proposed approach.

In [15], the authors extend RDF with temporal features and evolution operators. In addition, in contrast to the rest of the reviewed works, they associate concepts with their lifespan. In particular, an *evolution base* Σ is a set of RDF triples and a mapping τ from the set of considered RDF triples and considered resources to the set of temporal intervals. In addition, Σ may contain statements of the form (*c*, *term*, *c'*), where *term* is one of the special evolution properties *becomes*, *join*, *split*, *merge*, and *detach*.

The expression (*c, becomes, c'*) expresses that the concept *c'* originates from the concept *c* and should hold $\tau(c).end < t(c').start$. The expression (*c*, *join*, *c'*) expresses that a part of concept *c'* born at time *t* comes from a part of concept *c*. The expression (*c*, *spilt*, *c'*) expresses that a part of concept *c* ending at time *t* becomes a part of a new concept *c'*. The expression (*c*, *merge*, *c'*) indicates that a part of concept *c* ending at time *t* becomes part of an existing concept *c'*. The expression (*c*, *detach*, *c'*) indicates the new concept *c'* is formed at time *t* with at least one part from *c*.

An evolution base Σ is consistent, if for all $(s,p,o) \in \Sigma$ it holds that $\tau(s,p,o) \subseteq \tau(s)$ and $\tau(s,p,o) \subseteq \tau(o)$. Additionally, if $p \in$ {type, subClassOf, subPropertyOf} then it should hold that $\tau(s) \subseteq \tau(o)$.

To support evolution-aware querying, the authors define a navigational query language to traverse temporal and evolution edges in an evolution graph. This language is analogous to nSPARQL [16], a language that extends SPARQL with navigational capabilities based on nested regular expressions. nSPARQL uses four different axes, namely *self* , *next*, *edge*, and *node*, for navigation on an RDF graph and node label testing. The authors extend the nested regular expressions constructs of nSPARQL with temporal semantics and a set of five evolution axes, namely *join*, *split*, *merge*, *detach*, and *become*s that extend the traversing capabilities of nSPARQL to the evolution edges. The extended query language is formally defined.

An example query is "who was the head of the German government before and after the unification of 1990". The query is expressed as follows:

SELECT ?*Y*, ?*W*
(?*X*, **self**::*Reunified Germany*/**join**$^{-1}$[1990]/
next::*head*[1990], ?*Y*) AND
(?*Z*, **self**::*Reunified Germany*/**next**::*head*[1990], ?*W*)

The first triple finds all the heads of state of the *Reunified Germany* before the unification by following **join**$^{-1}$[1990] and then following **next** :: *head*[1990]. The second triple finds the heads of state of the Reunified Germany after the unification.

No implementation results of this theory are provided.

IV. Works That Extend RDFS Entailment

In [17], a *temporal graph* is a set of temporal triples of the form $(s,p,o)[t]$, where (s,p,o) is an RDF triple and t is a time point. Given a temporal graph G, $G(t)$ denotes the set of RDF triples in G corresponding to time point t.

The authors define *temporal entailment* between two temporal graphs G, G' as follows:

1. For ground temporal RDF graphs G, G', define $G \models_\tau G'$ iff $G(t) \models_{RDFS} G'(t)$, for each t.
2. For general temporal graphs G, G', $G \models_\tau G'$ iff for every ground instance $v(G)$ of G, there exists a ground instance $v'(G')$ of G' such that $v(G) \models_{RDFS} v'(G')$.

It is shown that temporal entailment is NP-complete. To test temporal entailment, the authors define the *slice closure* of G, as follows $scl(G) = \cup_t (cl(G(t)))^t$, where $cl(H)$ is the RDFS closure [18] of an RDF graph H and $H^t = \{(s,p,o)[t] \mid (s,p,o) \in H\}$. In particular, it is proved that $G \models_\tau G'$ iff there is a mapping v such that $v(G')$ is a subgraph of $scl(G)$.

The authors extend their theory to support also anonymous timestamps.

A query is defined as a pair $(H, B \cup A)$, where H and B are temporal RDF graphs without blank nodes and with some elements replaced by variables and A is a set of usual arithmetic built-in predicates over time point variables and time points. All variables appearing in H should also appear in B. For deriving maximal validity intervals a special structure is used. For example a query that asks for the service providers that have web services for more than 4 consecutive years is:

$$(?X, interval, ?t_e\text{-}?t_s) \leftarrow (?Y, provided\ by, ?X) \parallel ?t_s, ?t_e \parallel, ?t_e - ?t_s > 4.$$

No implementation of this theory is provided.

In [19], the authors extend the work in [17] and they define a *temporal graph* as a set of temporal triples of the form $(s,p,o){:}i$, where (s,p,o) is an RDF triple and i is a temporal interval variable or a temporal interval. A *temporal constraint* is an expression of the form $i\ \omega\ i'$, where i, i' are temporal intervals or temporal interval variables and ω is one of the relationships of Allen's temporal interval algebra [14]. A *temporal graph with temporal constrains* (called *c-temporal graph*) is a pair $C = (G, \Sigma)$, where G is a temporal graph and Σ is a set of temporal constraints over the intervals of G.

The authors define entailment between two *c*-temporal graphs *C*, *C'* as follows: $C \models_{\tau(const)} C'$ iff for each time ground instance *v*(*C*) of *C*, there is a time ground instance *v'*(*C'*) of *C'* such that $v(C) \models_{\tau} v(C')$. The authors define the *c-slice closure* of *C*, denoted by *cscl*(*C*), extending the definition of slice closure of [17]. It is proved that $C \models_{\tau(const)} C'$ iff there is an interval map γ from *C'* to *C* and a mapping *v* s.t. *v*(γ(*C'*)) is a subgraph of *cscl*(C). Entailment between two *c*-temporal graphs is shown to be NP-complete. No query language or implementation is provided.

In [20], [21], [22], the authors consider an extension of RDFS with spatial and temporal information. Here, we consider only the extension with temporal information. Assume a set *D* of RDF triples associated with their validity temporal interval *i*. Starting from *D*, the authors apply the inference rules *A*:?*i*, *B*:?*i'*→ *C*: ?*i* ∩ ?*i'*, where *A*, *B* → *C* is an RDFS entailment rule [2] and ?*i*, ?*i'* are temporal interval variables, until a fixpoint is reached. Then, the temporal intervals of the same RDF triple are combined, creating maximal temporal intervals.

Based on these maximal temporal intervals, a formal extension of the SPARQL language is proposed, called SPARQL-ST, supporting however only the AND and FILTER operations. The TEMPORAL FILTER condition is precisely defined supporting all interesting conditions between temporal intervals including Allen's temporal interval relations.

An example SPARQL-ST query that returns all house members who sponsored a bill after April 2, 2008, along with the temporal interval that the bill was sponsored is:

```
SELECT ?p, intersect(#t1, #t2, #t3, #t4) WHERE {
          ?p gov:hasRole ?r #t1.
          ?r gov:forOffice ?o #t2.
          ?o gov:isPartOf gov:congress_house #t3.
          ?p gov:sponsor ?b #t4.
          TEMPORAL FILTER (
          after(intersect(#t1, #t2, #t3, #t4), interval(04:02:2008,
                    04:02:2008, MM:DD:YYYY)))}
```

SPARQL-ST has been implemented by extending a commercial relational database system and experimental results are provided.

In [23], [24], the authors extend the RDFS and ter-Horst entailment rules [25] (which extend RDFS with terms from the OWL [26] vocabulary) with temporal information. Four example inference rules are presented, including:

```
?s ?p ?o ?b ?e
?p rdfs:domain ?dom
→
?s rdf:type ?dom ?b ?e
?p rdf:type owl:FunctionalProperty
?p rdf:type owl:ObjectProperty
?x ?p ?y ?b1 ?e1
?x ?p ?z ?b2 ?e2
→
?y owl:sameAs ?z
provided that [?b1,?e1] and [?b2,?e2] overlap
?p rdf:type owl:FunctionalProperty
?p rdf:type owl:DatatypeProperty
?x ?p ?y ?b1 ?e1
?x ?p ?z ?b2 ?e2
→
?x rsd:type owl:Nothing
provided that ?y ≠?z and  [?b1,?e1] and [?b2,?e2] overlap
```

Note that in the above rules ?*b*, ?*e*, ?*b1*, ?*e1*, ?*b2*, and ?*e2* are time point variables. The last rule indicates that inconsistency is expressed by assigning the bottom type *owl*:*Nothing* to individuals. For checking consistency two additional rules must be added addressing a combination of *owl*:*sameAs* and *owl*:*differentFrom*, as well as *owl*:*disjointWith* together with two *rdf*:*type* statements.

The proposed extension has been implemented using the forward chaining engine *HFC* [27], which supports arbitrary tuples, user defined tests, and actions. Some experimental results are provided. However, no query language is provided.

In [28], a general framework for representing, reasoning, and querying annotated RDFS data is presented. The authors show how their unified reasoning framework can be instantiated for the temporal, fuzzy, and provenance domain. Here, we are concerned with the temporal instantiation. We define ⊥={{}} and ⊤={[-∞,+∞]}. Let *L*={*t* | *t* is a finite set of disjoint temporal intervals} ∪ {⊥, ⊤}. On *L*, the authors define the partial order:

$t \leq t'$ iff for all $i \in t$, there is $i' \in t'$ such that $i \subseteq i'$.

Obviously, (*L*, ≤, ⊥, ⊤) is a bounded lattice. Between the elements of *L*, the authors define the operations + and × are follows: $t_1 + t_2 = inf(t \mid t_i \leq t, i=1,2\}$ and $t_1 \times t_2 = sup(t \mid t \leq t_i, i=1,2\}$. For example, {[2,5],[8,12]} + {[4,6],[9,15]} = {[2,6],[8,15]} and {[2,5],[8,12]} × {[4,6],[9,15]} = {[4,5], [9,12]}. An annotated *RDFS graph G* is a set of temporal triples (*s*,*p*,*o*) :*t*, where (*s*,*p*,*o*) is an RDF triple and $t \in L$. The models of *G* are formally defined extending ρRDF semantics, where ρRDF [29] is a subset of RDFS keeping its essential features.

The authors present a set of sound and complete inference rules of the general form:

$(s_1, p_1, o_1) : ?t_1, \ldots, \; (s_n, p_n, o_n) : ?t_n,$
$\{(s_1, p_1, o_1), \ldots, (s_n, p_n, o_n)\} \vdash_{\rho RDF} (s, p, o)$
$\rightarrow$
$(s,p,o) : (?t_1 \times \ldots \times ?t_n)$

For example:

$(c_1, rdfs{:}subClassOf, c_2){:}?t_1, \; (c_2, rdfs{:}subClassOf, c_3){:}?t_2$
$\rightarrow$
$(c_1, rdfs{:}subClassOf, c_3){:}?t_1 \times ?t_2$

Additionally, the inference rules contain the generalization rule:

$(s,p,o) : ?t, \; (s,p,o) : ?t' \rightarrow (s,p,o) : (?t + ?t').$

The generalization rule is destructive, meaning that this rule removes its premises as the conclusion is inferred.

An extension of SPARQL is presented for querying an annotated RDF graph. A *basic annotated pattern* is an expression (*s*,*p*,*o*):*t*, where *s, p, o, t* can be variables. Let *P* be a basic annotation pattern and *G* be a temporal graph. The authors define the evaluation $[P]_G$ as the list of substitutions that are solutions of *P*, i.e., $[P]_G=\{\theta \mid G \text{ entails } \theta(P)\}$. Based on $[P]_G$ the evaluations $[P \text{ AND } P']_G$, $[P \text{ UNION } P']_G$, $[P \text{ FILTER } R]_G$, $[P \text{ OPTIONAL } P'[R]]_G$ are formally defined, where *R* is a filter expression.

An example query asking for the employees of eBay during some time period that optionally owned a car at some point during their stay is:

```
SELECT ?p, ?t, ?c WHERE {
        (?p type ebayEmp): ?t
        OPTIONAL {(?p hasCar ?c): ?t'
                FILTER (?t' ⊆ ?t)}}
```

Note that the definition of $[P]_G$ is not based on maximal temporal intervals and, thus all temporal intervals that satisfy the query are returned. Therefore, the authors define an ordering between substitutions: $\theta' \leq \theta$ iff (i) $\theta \neq \theta'$, (ii) $domain(\theta) = domain(\theta')$, (iii) $\theta(x) = \theta'(x)$, for any non-temporal variable *x*, and (iv) $\theta'(t) \subseteq \theta(t)$, for any temporal variable *t*. Then, for any $\theta \in [P]_G$, remove any $\theta' \in [P]_G$ such that $\theta' \leq \theta$.

No implementation is provided for this theory.

In [30], a *temporal graph G* is a set of temporal triples $(s,p,o)[t,t']$, where (s,p,o) is an RDF triple and $[t,t']$ is its corresponding validity temporal interval. The semantics of a temporal graph *G*, assuming an entailment relation *X* (such as RDF, RDFS, and OWL2 RL/RDF [31] entailment) are formally defined using multi-sorted first order logic.

A *basic graph pattern* (*BGP*) is a set of triples (s,p,o), where s,p,o can be variables. A *temporal group pattern* (*TGP*) is an expression defined inductively, as follows:

B at t_3, *B* during $[t_1,t_2]$, *B* occurs $[t_1,t_2]$
B maxinterval $[t_1,t_2]$, *B* mintime t_3, *B* maxtime t_3,
P_1 and P_2, P_1 union P_2, P_1 optional P_2,
P_1 filter *R*,

where *B* is a *BGP*, P_1 and P_2 are *TGPs*, *R* is a built-in expression, and t_1,t_2, and t_3 are either time points or variables. Note that a *TGP* query is an extension of a SPARQL query. For example, a *TGP* query that retrieves all events *z* in London having at least one time point in common with Oktoberfest is:

```
SELECT ?z WHERE {
{(Munich, hosts, Oktoberfest)} maxint [?x,?y].
(London, host,?z)} occurs [?x,?y].}
```

The evaluation of a *TGP* query w.r.t. a temporal graph *G* and an entailment relation *X* is formally defined using multi-sorted first-order logic. Yet, evaluation of a *TGP* using this definition can be inefficient. Therefore, the authors describe an optimization.

Assume that the entailment relation *X* is characterized by a set of definite rules of the form: $A_1,..,A_n \rightarrow B$.

Then, the rules:

$$A_1[x_1,y_1], \ldots, A_n[x_n,y_n], max(x_1,\ldots,x_n) \leq min(y_1,..,y_n) \rightarrow B[max(x_1,\ldots,x_n), min(y_1,..,y_n)]$$

are applied until a fixpoint is reached, where x_i and y_i are time point variables. Then, based on the result, derived RDF triples are associated with their maximal validity intervals. Now, based on these maximal intervals the evaluation of a *TGP* query is efficiently defined.

Though the authors state that they have implemented their framework using the PostgreSQL database system, no implementation results are provided.

V. Conclusion-Discussion

In this paper, we have reviewed models and query languages of temporally annotated RDF. Below, we compare these models and query languages on various aspects. First, we would like to state that approaches that have their own model theory or extend RDF simple entailment miss important inferences made from the works that extend RDFS entailment. For example, an object *o* may be an instance of class *c* during a temporal interval *i* and the class *c* may be subclass of a class *c'* during an interval *i'*. Only works that extend RDFS entailment are able to derive that *o* is instance of class *c'* during the intersection of the intervals *i* and *i'*.

From the works that extend RDFS entailment, the approach in [17] seems less efficient since it computes the RDFS closure of RDF triples at each time point. Additionally, [28] considers all temporal intervals that satisfy the query and then selects the maximal ones. In contrast, [22] and [30] achieve query answering using directly maximal temporal intervals achieving a higher performance.

In our opinion, the approach in [28] does not give always the desirable results: For example, assume that an annotated RDFS graph consists of the triples (s,p,o):[1998, 2009] and (s',p',o'):[2008, 2012]. Consider now the query:

```
SELECT ?t, ?t' WHERE {
(s,p,o): ?t.
(s',p',o'): ?t'.
FILTER (before(?t,?t'))}
```

Then, [28] will return the answers (i) ?*t*=[1998, 2007], ?*t'*=[2008, 2012], (ii) ?*t*=[1998, 2008], ?*t'*=[2009, 2012], and (iii) ?*t*=[1998,2009], ?*t'*=[2010, 2012]. In contrast [22], will return no answer to this query since it works with maximal temporal intervals and 2009 > 2008.

In [30], the query:

```
SELECT ?t, ?t' WHERE {
        (s,p,o) during ?t.
        (s',p',o') during ?t',
        FILTER (before(?t,?t'))}}
```

will return the answers of [28], as well as the intervals *t,t'* such that $t \subseteq$ [1998,2009], $t' \subseteq$ [2008,2012], and *t.end* <*t'.start*. In contrast, in [30], the query:

```
SELECT ?t, ?t' WHERE {
        (s,p,o) maxinterval ?t.
```

(s',p',o') maxinterval ?*t',*
FILTER (*before*(?*t*,?*t'*)}}

will return no answer. As a criticism, [30] is not able to return maximal intervals within a temporal interval of interest.

Approaches [7], [28], and [11] save some space since they either use name graphs associated with temporal intervals or associate each RDF triple with its set of validity temporal intervals.

Specialized indices for query answering are used only in [4] and [7], while the rest of the approaches use common indexes. As a final remark, we would like to state that [4] can handle some temporal constraints over RDF triples, [17] can handle anonymous timestamps, and [19] can handle anonymous temporal intervals satisfying Allen's temporal interval algebra relations.

Temporal consistency constraints are considered only in [13], which however does not answer temporal queries but only normal queries.

As a criticism to the work in [6], each RDF triple is associated with a single maximal temporal interval while an RDF triple is normally associated with multiple maximal temporal intervals.

Some of the proposed models and query languages have been implemented as stated in the main text of the paper and for some of them experimental results are provided.

In the future, extensions of the proposed temporal RDF query languages with features of SPARQL 1.1 [32], such as subqueries, and negation, will be of great importance. For example, it will be interesting to ask for events that have not occurred simultaneously before a date and their maximal temporal intervals always overlap after that date. Additionally, it will be interesting to ask for companies located in Crete that have exactly one manager at each point in time within a particular temporal interval of interest.

Future work also concerns a survey on spatial, fuzzy, provenance, and contextual RDF. Of course, aspects of contextual RDF can be time, space, trust, and authority.

REFERENCES

[1] G. Klyne and J. J. Carroll, "Resource Description Framework (RDF): Concepts and Abstract Syntax", W3C Recommendation, 10 February 2004, available at http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

[2] Patrick Hayes, "RDF Semantics", W3C Recommendation, 10 February 2004, available at http://www.w3.org/TR/2004/REC-rdf-mt-20040210/.

[3] E. Prudhommeaux and A. Seaborne, "SPARQL query language for RDF", W3C Recommendation 15 January 2008, available at http://www.w3.org/TR/rdf-sparql-query/.

[4] A. Pugliese, O. Udrea, and V.S. Subrahmanian. "Scaling RDF with Time", International World Wide Web Conference (WWW), Beijing, China, 2008, pp 605-614.

[5] B. Salzberg and V. J. Tsotras, "Comparison of access methods for time-evolving data", ACM Computing Surveys, 31(2), 1999, pp158–221.

[6] O. Udrea, D. R. Recupero, and V. S. Subrahmanian, "Annotated RDF", ACM Transactions on Computational Logic, 11(2), 2010.

[7] J. Tappolet and A. Bernstein, "Applied Temporal RDF: Efficient Temporal Querying of RDF Data with SPARQL", 6th European Semantic Web Conference (ESWC-2009), 2009, pp. 308-322.

[8] J. J. Carroll, C. Bizer, P. J. Hayes, and P. Stickler, "Named graphs", Journal of Web Semantics, 3(4), 2005, pp. 247-267

[9] A. Rodriguez, R. E. McGrath, Y. Liu, and J. D. Myers, "Semantic Management of Streaming Data", 2nd International Workshop on Semantic Sensor Networks at the International Semantic Web Conference, Washington, 2009.

[10] B. McBride and M. Butler, "Representing and Querying Historical Information in RDF with Application to E-Discovery", HP Laboratories Technical Report, HPL-2009-261, 2009.

[11] F. Grandi, "T-SPARQL: a TSQL2-like Temporal Query Language for RDF", 14th East-European Conference on Advances in Databases and Information Systems (ADBIS-2010) (Local Proceedings), 2010, pp. 21-30.

[12] R.T. Snodgrass (ed.), I. Ahn, G. Ariav, D. Batory, J. Clifford, C.E. Dyreson, R. Elmasri, F. Grandi, C.S. Jensen, W. Kafer, N. Kline, K. Kulkarni, T.Y. Cliff Leung, N. Lorentzos, R. Ramakrishnan, J.F. Roddick, A. Segev, M.D. Soo, and S.M. Sripada, "The TSQL2 Temporal Query Language", Kluwer Academic Publishers, 1995.

[13] M. Dylla, M. Sozio, and M. Theobald, "Resolving Temporal Conflicts in Inconsistent RDF Knowledge Bases", Datenbanksysteme fur Business, Technologie und Web (BTW-2011), 2011, pp. 474-493.

[14] J. Allen, "Maintaining Knowledge about Temporal Intervals", Communications of the ACM, 26(11), 1983, pp. 832-843.

[15] S. Bykau, J. Mylopoulos, F. Rizzolo, and Y. Velegrakis, "On Modeling and Querying Concept Evolution", Journal on Data Semantics, 1, 2012, pp 31-55.

[16] J. Perez, M. Arenas, and C. Gutierrez, "nSPARQL: A navigational language for RDF", Journal of Web Semantics, 8(4), 2010, pp. 255-270.

[17] C. Gutierrez, C. A. Hurtado, and A. A. Vaisman, "Introducing Time into RDF", IEEE Transactions on Knowledge and Data Engineering, 19(2), 2007, 207-218.

[18] C. Gutierrez, C. A. Hurtado, and A.O. Mendelzon, "Foundations of Semantic Web Databases", 23rd Symposium of Principles of Databases Systems (PODS-2004), 2004, pp. 95-196.

[19] C. A. Hurtado and A. Vaisman, "Reasoning with Temporal Constraints in RDF", 4th International Workshop on Principles and Practice of Semantic Web Reasoning (PPSWR-2006), 2006, pp. 164-178.

[20] M. Perry, A. P. Sheth, F. Hakimpour, and P. Jain, "Supporting Complex Thematic, Spatial and Temporal Queries over Semantic Web Data", 2nd International Conference on GeoSpatial Semantics (GeoS-2007), 2007, pp. 228-246.

[21] M. Perry and A. P. Sheth, "A Framework to Support Spatial, Temporal, and Thematic Analytics over Semantic Web Data, Knoesis Center Technical Report, KNOESIS-TR-2008-01, 2008.

[22] M. Perry, P. Jain, and A. P. Sheth, "SPARQL-ST: Extending SPARQL to Support Spatiotemporal Queries", N. Ashish and A.P. Sheth (Eds.) Geospatial Semantics and the Semantic Web - Foundations, Algorithms, and Applications, 2011, pp. 61-86.

[23] H.Krieger, "A Temporal Extension of the Hayes and ter Horst Entailment Rules for RDFS and OWL", AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning, 2011.

[24] H.Krieger, "A Temporal Extension of the Hayes/ter Horst Entailment Rules and a Detailed Comparison with W3C's N-ary Relations", Deutsches Forschungszentrum fur Kunstliche Intelligenz GmbH Technical Report, RR-11-02, 2011.

[25] H. J. ter Horst, "Completeness, decidability and complexity of entailment for RDF Schema and a semantic extension involving the OWL vocabulary, Journal of Web Semantics, 3(2-3), 2005, pp. 79-115.

[26] G. Antoniou and F. van Harmelen, A semantic web primer, MIT Press, 2004.

[27] H.U. Krieger and G.J.M. Kruijff, "Combining uncertainty and description logic rule-based reasoning in situation-aware robots" Proceedings of the AAAI 2011 Spring Symposium on Logical Formalizations of Commonsense Reasoning, 2011

[28] A. Zimmermann, N. Lopes, A. Polleres, and U. Straccia, "A General Framework for Representing, Reasoning and Querying with Annotated Semantic Web Data", Journal of Web Semantics , 11, 2012, pp. 72–95.

[29] S. Munoz, J. Perez, and C. Gutierrez, “Minimal Deductive Systems for RDF”, 4th European Semantic Web Conference (ESWC-2007), 2007, pp. 53-67.

[30] B. Motik, "Representing and Querying Validity Time in RDF and OWL: A Logic-Based Approach", 9th International Semantic Web Conference (ISWC-2010), 2010, pp. 550-565.

[31] B. Motik, B. Cuenca Grau, I. Horrocks, Z. Wu, A. Fokoue, and C. Lutz, “OWL 2 Web Ontology Language Profiles”, W3C Recommendation 27 October 2009, available at http://www.w3.org/TR/owl2-profiles/.

[32] Steve Harris and Andy Seaborne, “SPARQL 1.1 Query Language”, W3C Working Draft 24 July 2012, available at http://www.w3.org/TR/sparql11-query/

AUTHORS PROFILE

Anastasia Analyti earned a B.Sc. degree in Mathematics from University of Athens, Greece and a M.Sc. and Ph.D. degree in Computer Science from Michigan State University, USA. She worked as a visiting professor at the Department of Computer Science, University of Crete, and at the Department of Electronic and Computer Engineering, Technical University of Crete. Since 1995, she is a principal researcher at the Information Systems Laboratory of the Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH-ICS). Her current interests include reasoning on the Semantic Web, modular web rule bases, non-monotonic-reasoning, faceted metadata and semantics, conceptual modelling, contextual organization of information, information integration and retrieval systems for the web, interoperability of heterogeneous and distributed information bases, and biomedical information systems. She has participated in several research projects and has published over 55 papers in refereed scientific journals and conferences.

Ioannis Pachoulakis received a B.Sc. in Physics (1988) at the University of Crete, Greece, and a Ph.D. in Astrophysics (1996) and an M.Sc. in Engineering (1998), both from the University of Pennsylvania in the U.S.A. Since 2001, he serves as an Assistant Professor at the Department of Applied Informatics and Multimedia at TEI of Crete with mainstream interests in realistic multimedia applications, virtual reality and multimedia applications for science.

An RGB Image Encryption Supported by Wavelet-based Lossless Compression

Ch. Samson[1]
Dept. of Information Technology, SNIST,
Hyderabad, India,

V. U. K. Sastry[2]
Dept. of Computer Science & Engineering., SNIST,
Hyderabad, India,

Abstract— **In this paper we have proposed a method for an RGB image encryption supported by lifting scheme based lossless compression. Firstly we have compressed the input color image using a 2-D integer wavelet transform. Then we have applied lossless predictive coding to achieve additional compression. The compressed image is encrypted by using Secure Advanced Hill Cipher (SAHC) involving a pair of involutory matrices, a function called Mix() and an operation called XOR. Decryption followed by reconstruction shows that there is no difference between the output image and the input image. The proposed method can be used for efficient and secure transmission of image data.**

Keywords- Image compression; Wavelet Transform; image encryption; Lifting Wavelet ; Secure Advanced Hill Cipher.

I. INTRODUCTION

When network bandwidth and storage space are limited, an image to be transmitted has to be compressed. It is necessary to protect confidential image data during transmission from unauthorized access. Compression [1] reduces the storage space required to represent a given quantity of information. Image compression [1] is of two types: lossy and lossless (error- free). Lossless compression schemes are reversible so that the original data can be reconstructed exactly, while lossy schemes accept some loss of data in order to achieve higher compression. Lossless compression can be used for text, medical images and legal documents etc. whereas lossy compression is used for natural images, speech signals etc. Cryptography plays an important role in information security. Encryption [2] is the process of converting information into an unintelligible form. Image encryption has applications in Internet communication, multimedia systems, medical imaging, telemedicine, military communication, etc.

Wavelet Transform has emerged as a powerful mathematical tool in many areas of science and engineering. The power of Wavelets comes from the use of multiresolution analysis. In a recent investigation [3], we have studied the encryption of an image supported by lossy compression by using multilevel Wavelet Transform.

The study of integer wavelets based on lifting scheme [4] has gained considerable impetus in the recent years. Unlike classical wavelets which are obtained by translations and dilations of one function in the frequency domain, the wavelets governed by the lifting scheme are obtained by a new approach based on spatial domain. Many researchers [5-9] have dealt with image compression using lifting based wavelet transform.

In one of the recent investigations, Bibhudendra et al. [10] have proposed an advanced Hill cipher algorithm which uses an Involutory key matrix for image encryption. We have enhanced the advanced Hill cipher by introducing a pair of involutory matrices, a function called Mix( ) and XOR operation, and we have called this cipher as Secure Advanced Hill Cipher (SAHC).

In the present paper, our objective is to develop a method for an RGB image encryption using lifting scheme based on lossless compression. Firstly, we compress the input image using lifting wavelet transform and then encrypt the compressed image applying Secure Advanced Hill Cipher.

In what follows we present the plan of the paper. In section 2, we explain the proposed method. Section 3 describes the process of image compression using lifting wavelet transform. We present an approach for SAHC based image encryption in section 4. Section 5 deals with computations that are carried out in this analysis and draw conclusions.

II. PROPOSED METHOD

Encryption following compression leads to a faster and secure transmission of image data across a channel. So image is compressed prior to encryption. Perfect reconstruction is possible with Lifting Wavelet Transform. A digital color image is represented in terms of three color components, namely, Red, Green and Blue (RGB). Each component is like gray scale image. So the three components of an RGB image can be coded separately and concatenated at the end. The Schematic diagram of the proposed method is shown in Figure 1.

The proposed method is developed by the following steps.

1. Lifting scheme based Transform coding: Choose an integer wavelet and number of lifting steps (levels) N. Perform the transform coding of the image at level N.

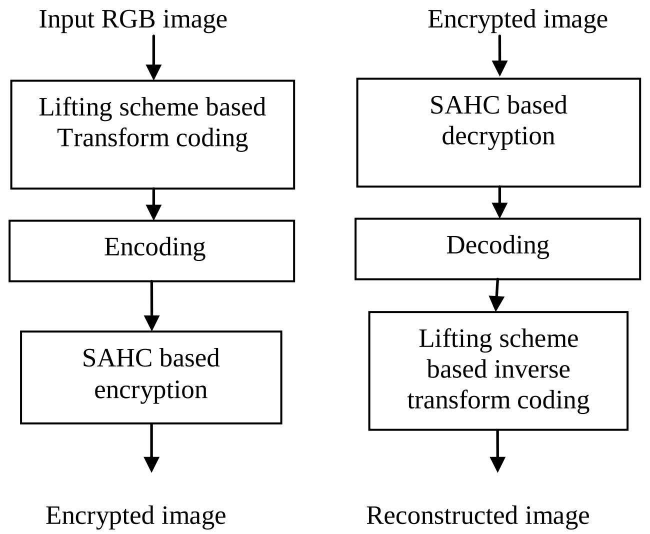
2. **Encoding**: Use lossless predictive coding to achieve additional compression.

3. Encryption: Encrypt the encoded image using Secure Advanced Hill Cipher.

4. Decryption: Get encoded image by performing decryption using Secure Advanced Hill Cipher.

5. **Decoding:** Use lossless predictive decoding to get the transform coded image.

6. Reconstruction: Use Lifting scheme based inverse transform coding to get reconstructed original input image.



(a) Process of coding (b) Process of Decoding

Figure 1. Schematic diagram of the proposed method

III. LIFTING SCHEME FOR IMAGE COMPRESSION

Lifting scheme is an alternative approach to Discrete Wavelet Transform. It was proposed by Sweldens [5]. Lifting is a way of describing and calculating wavelets. It calculates wavelets in place, which means that it takes no extra memory to do the transform. Every wavelet can be written in lifting form. The input signal is first split into even and odd indexed samples. The samples are correlated, so that it is possible to predict odd samples from even samples as given below.

$$odd_{new} = odd_{old} + \alpha\ (even_{left} + even_{right})$$

where α is the predict step coefficient. The difference between the actual odd samples and the prediction becomes the wavelet coefficients. The operation of obtaining the differences from the prediction is called the lifting step. The update step follows the prediction step, where the even values are updated from the input even samples and the updated odd samples.

$$even_{new} = even_{old} + \beta\ (odd_{left} + odd_{right})$$

where β is the update step coefficient. They become the scaling coefficients which will be passed on to the next stage of transform. This is called the second lifting step. This lifting scheme, called forward lifting scheme, is shown in Figure 2.

The basic idea of the reverse process of the above lifting scheme is displayed in Figure 3.

The lifting scheme provides integer coefficients and so it is exactly reversible. The total number of coefficients before and after the transform remains the same.

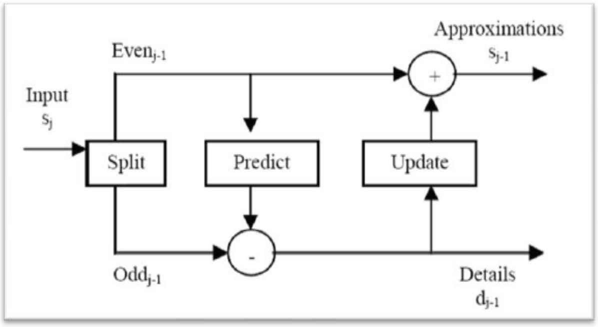


Figure 2. Forward Lifting Scheme.

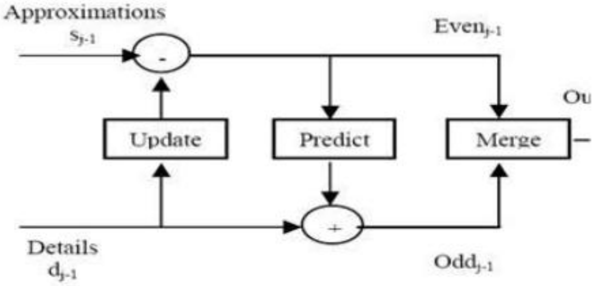


Figure 3. Reverse Lifting Scheme

The inverse transform gets back the original signal by exactly reversing the operations of the forward transform with a merge operation in place of a split operation. The number of samples in the input signal must be a power of two, and these samples are reduced by half in each succeeding step until the last step which produces one sample.

To achieve additional compression, lossless predictive coding [1] is applied to each sub band using different values of predictor coefficients alpha and beta, giving an encoded image as output. The reverse process is applied to the encoded image to get back the transformed image. Then on applying inverse transform coding on the transformed image, we get back the reconstructed image. The reconstructed image which we have obtained in our analysis is an exact replica of the original input image.

Let us now consider an RGB image given in Figure 4. and focus our attention on one of the component images. On adopting the process of compression, described above, we get the corresponding compressed image. The same procedure is applied on the remaining component images to obtain the corresponding compressed images.

IV. SAHC BASED IMAGE ENCRYPRTION

In the advanced Hill cipher, the basic equations governing encryption and the decryption are given by

$$C = AP \bmod N,$$
$$P = AC \bmod N.$$

respectively. Here A is an involutory matrix which includes the key matrix. As A is an involutory matrix, we have $A^{-1} = A$, where A^{-1} is the modular arithmetic inverse of A. Thus in the case of this cipher, we need not compute the modular arithmetic inverse of A separately, once A is known to us. This is the advantage which is being achieved in this cipher.

Image encryption using Secure Advance Hill Cipher includes a pair of involutory matrices A and B, as multiplicands of the input matrix P, a function called Mix( ) and an operation called XOR . The function Mix( ) is used for mixing the binary bits to create confusion and diffusion thoroughly. The values of d and e are integers required in the development of the involutory matrices A and B. In our analysis, they are taken as 5 and 7 respectively. The value of N is taken as 256.

ALGORITHM FOR ENCRYPTION

```
1.  Read P, K, L, d,e,r, N
2.  A = Involute (K, d)
    and
    B = Involute (L,e)
3.  Construct NT and ST
4.  for i = 1 to r
    P = (APB) mod  N
    P = Mix (P)
    P = P ⊕ ST
    end
5.  C = P
6.  Write C.
```

ALGORITHM FOR DECRYPTION

```
1.  Read C, K, L, d,e,r, N
2.  A = Involute (K, d)
    and
    B = Involute (L,e)
3.  Construct NT and ST
7.  for i = 1 to r

    C = XOR (C,ST)
    C = Imix (C)
    C = (ACB) mod  N
   end
7.  P = C
8.  Write P.
```

Here K and L are the key matrices, NT is the number table containing the numbers 0 to 255, ST is the substitution table, and r denotes the number of rounds taken as 16. The details of the advanced Hill cipher can be found in [11].

On adopting the SAHC based encryption algorithm discussed in this section, on each one of the compressed component images separately, we get the encrypted image corresponding to each one of the component images. On combining all these encrypted component images in an appropriate manner, we get the encrypted form of the color image presented in Figure 7. On using SAHC decryption algorithm, we get back the encoded image.

V. COMPUTATIONS AND CONCLUSIONS

In this paper we have implemented an RGB image encryption supported by lifting scheme based lossless compression using MATLAB [12]. In this analysis, we have considered lifting wavelet based on Haar transform. The input image and its corresponding transform coded image, encoded image, encrypted image, decrypted image, decoded image and reconstructed image are shown in Figures 4 to 10 respectively. It is interesting to note that the reconstructed image is exactly identical to the original input image.

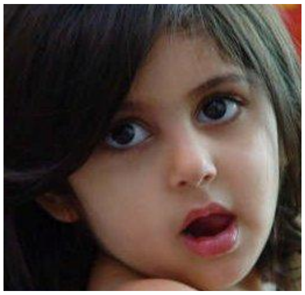


Figure 4. Input RGB image of a baby

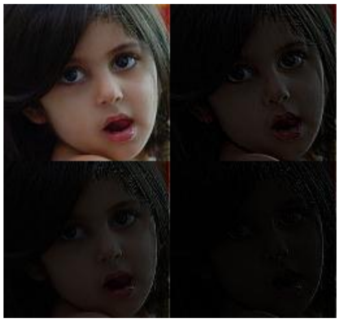


Figure 5. Image obtained after transform coding.

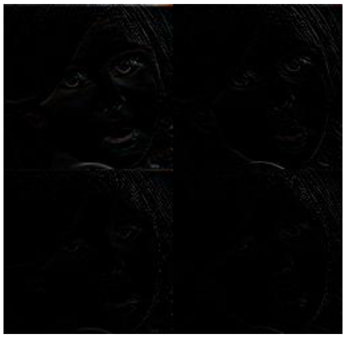


Figure 6. Encoded image.

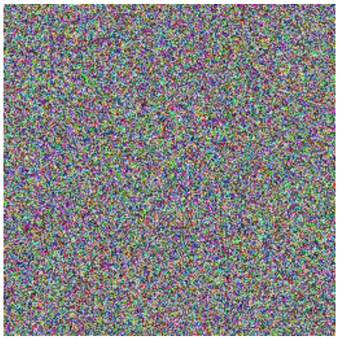


Figure 7. Encrypted image.

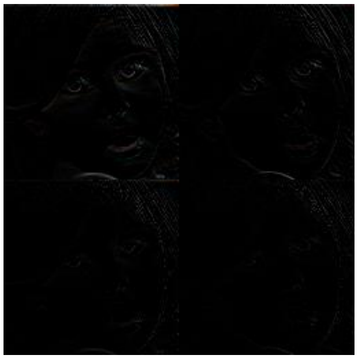


Figure 8. Decrypted image.

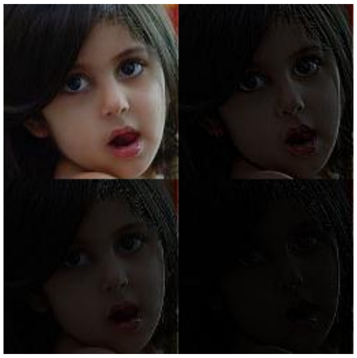


Figure 9. Decoded image.

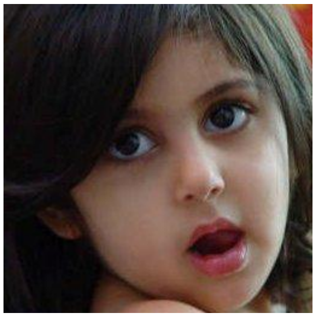


Figure 10. Reconstructed image

The same experiment is carried out with another color image which is given in Figure 11.

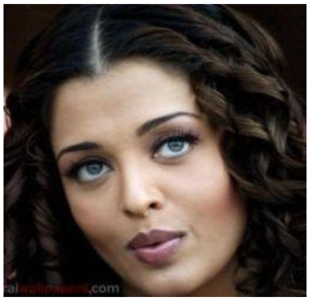


Figure 11. An RGB color image.

Thus we get the following images at various stages of the process.

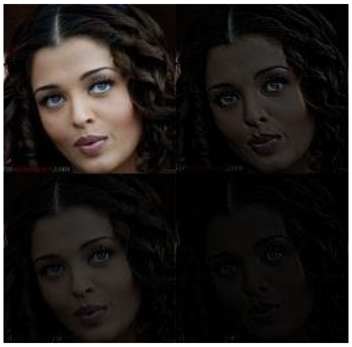


Figure 12. Image obtained after lifting based transform coding

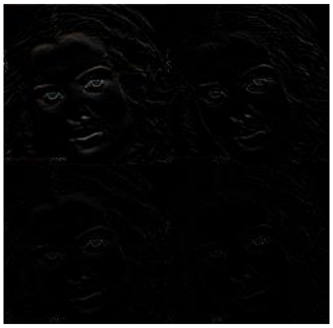


Figure 13. Encoded image.

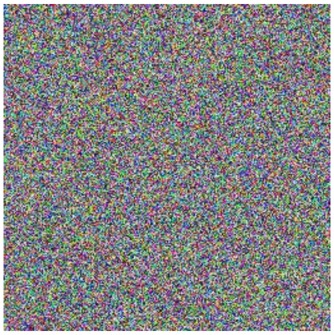


Figure 14. Encrypted image.

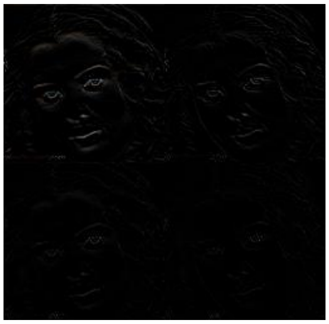


Figure 15. Decrypted image.

From the above analysis, we conclude that the encryption supported by compression is an interesting one and it can be used for the transmission color images more effectively in a secured manner.

REFERENCES

[1] Rafael C. Gonzalez & Richard E. Woods,— Digital Image processing, 2ndEdition Pearson Education 2004.

[2] William Stallings, Cryptography and Network Security, Principles and Practice, Third edition, Pearson, 2003.

[3] Ch.Samson,V. U. K. Sastry," A novel method for image encryption supported by compression using multilevel Wavelet Transform", International Journal of Advanced Computer Science and Applications,Vol. 3. No. 8,August 2012.

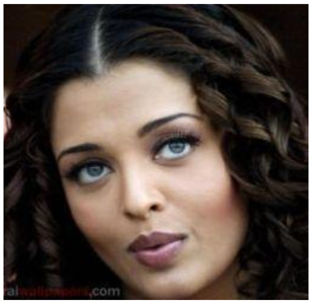


Figure 16. Decoded image.

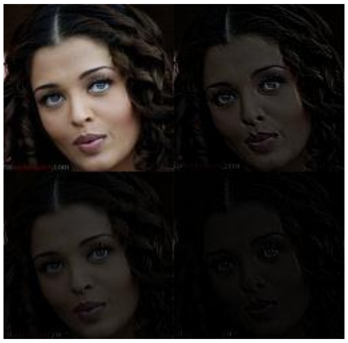


Figure 17. Reconstructed image

[4] K.P. Soman, K.I. Ramachandran, Insight into Wavelets from theory to practice, Second edition, PHI, 2006.

[5] W. Sweldens. "The Lifting Scheme: A New Philosophy in Biorthogonal Wavelet Constructions."*Proc. SPIE*, vol. 2569, pp. 68-79, 1995.

[6] C. Lian, K. Chen, H. Chen, and L. Chen, "Lifting Based Discrete Wavelet Transform Architecture for JPEG2000,*" IEEE Int. Symp. Circuits and Systems*, vol. 2, pp. 445-448, May 2001.

[7] Pei-Yin Chen, "VLSI implementation for one-dimensional multilevel lifting-based wavelet transform," IEEE Trans. Computers, Vol. 53, pp.386-398, April 2004.

[8] H. Liao, M. K. Mandal, and B.F. Cockburn, "Efficient architectures for 1-D and 2-D liftingbased wavelet transforms" IEEE Trans. Signal Processing, Vol. 52, pp. 1315-1326, May 2004.

[9] Dr.B Eswara Reddy and K Venkata Narayana,' A Lossless Image Compression Using Traditional and Lifting based Wavelets', Signal & Image Processing : An International Journal (SIPIJ) Vol.3, No.2, April 2012.

[10] Bibhudendra Acharya, Girija Sankar Rath, Sarat Kumar Patra, Saroj Kumar Panigrahy. 2007. Novel Methods of Generating Self-Invertible Matrix for Hill Cipher Algorithm, International Journal of Security, Vol 1, Issue 1, 2007, pp. 14-21..

[11] V. U. K. Sastry, Ch.Samson," Cryptography of a Gray Level Image and a Color Image Using Modern Advanced Hill Cipher Including a Pair of Involutory Matrices as Multiplicands and Involving a Set of Functions", International Journal of Engineering Science and Technology,Vol. 4 No. 7 July 2012.

[12] Alasdair Mcandrew, —Digital Image processing with MatLab, Cengage learning 2004.

AUTHORS PROFILE

Dr. V. U. K. Sastry is presently working as Professor in the Dept. of Computer Science and Engineering (CSE), Director (SCSI), Dean (R & D), SreeNidhi Institute of Science and Technology (SNIST), Hyderabad, India. He was Formerly Professor in IIT, Kharagpur, India and worked in IIT, Kharagpur during 1963 – 1998. He guided 12 PhDs, and published more than 80 research papers in various international journals. He received the best Engineering College Faculty Award in Computer Science and Engineering for the year 2008 from the Indian Society for Technical Education (AP Chapter) and Cognizant- Sreenidhi Best faculty award for the year 2012. His research interests are Network Security & Cryptography, Image Processing, Data Mining and Genetic Algorithms.

Mr. Ch. Samson obtained his Diploma from Govt. Polytechnic, Hyderabad in 1994, B. E. from Osmania University in 1998 and M. E from SRTM University in 2000. Presently he is pursuing Ph.D. from JNTUH, Hyderabad since 2009. He published 10 research papers in various international journals and two papers in conferences. He is currently working as Associate Professor and Associate Head in the Dept. of Information Technology (IT), SNIST since June 2005. His research interests are Image Processing, Image Cryptography and Network Security.

Feature Subsumption for Sentiment Classification of Dynamic Data in Social Networks using SCDDF

Jayanag. B[1], Vineela. K[2], Dr. Vasavi. S[3]
Department of Computer Science and Engineering, V. R. Siddhartha Engineering College
Vijayawada, India.

Abstract- **The analysis of opinions till now is done mostly on static data rather than on the dynamic data. Opinions may vary in time. Earlier methods concentrated on opinions expressed in an individual site. But on a given concept opinions may vary from site to site. Also the past works did not consider the opinions at aggregate level.**

This paper proposes a novel method for Sentiment Classification that uses Dynamic Data Features (SCDDF). Experiments were conducted on various product reviews collected from different sites using QTP. Opinions were aggregated using Bayesian networks and Natural Language Processing techniques. Bulk amount of dynamic data is considered rather than the static one. Our method takes as input a collection of comments from the social networks and outputs ranks to the comments within each site and finally classifies all comments irrespective of the site it belongs to. Thus the user is presented with overall evaluation of the product and its features.

Keywords- Sentiment classification, Natural language processing (NLP); opinions; features; Quick Test Professional (QTP); feature identification; sentiment prediction; summary generation.

I. Introduction

The present opinion mining is done statically only for a small set of data and the dependencies in the opinions are not considered for summarization. An architecture that could automatically process the comments, generate a generalized result out of the list of comments posted about a product by considering the dependencies could be useful to give a brief synopsis of the product. This becomes a real-life application, a completely automated solution that extracts the comments posted in a social network and categorizes them based on most prominent ranks. Thus it helps the user to know about the pros and cons of a product and its features based on the existing user's feedback with little effort.

The proposed architecture is based on opinion mining, a sub discipline within data mining and computational linguistics, refers to the computational techniques for extracting, classifying, understanding, and assessing the opinions expressed in various online news sources, social media comments, and other user-generated content.

Many numbers of sites provide different comments on the products but viewing all of them become rather difficult so we evaluate them based on the ranks and present a generalized result. The opinions posted by various users in social networks are extracted, the comments are evaluated by dividing them into tokens and using the natural language processing techniques like POS (Parts Of Speech) tagging. Meanings are analyzed by using the web dictionary WordNet [7, 21]. The dependencies in the opinions are analyzed using the Bayesian Networks and the sentiment is predicted for those corresponding words. And finally based on the predicted word counts ranks are given to these sentiments and are summarized. System gives the cumulative rank and displays the string corresponding to it.

Section II presents related work on the present study. Section III presents our proposed system. Experimental results are given in section IV. Conclusion and Future Work are given in Section V and VI.

II. Related Work

Previous works concentrated on opinions in individual sites and also limited the data set to a single line comment or static data or a limit on number of characters. Those current studies are mainly focused on mining opinions in reviews and/or classify reviews as to only positive or negative based on the sentiments of the reviewers but not on relative degree of positive or negativeness. Detailed study on previous works can be found in [13].

Abbasi et al. [1] considered web forms, blogs and articles and used WordNet score but haven't considered the word dependencies. Ahmed Abbasi [2], worked on feature selection methods and considered Intelligent Feature Selection (IFS) approach that uses syntactic and semantic information to refine larger input features, but these formation modules need to be expounded on, and real-world knowledge bases could be considered.

Cardie et al. [3], concentrated opinion-oriented information extraction. They created opinion-oriented "scenario templates" for summary representations of the opinions expressed in a document, or a set of documents to perform question answering. They did not identify product features and user opinions on these features to automatically produce a summary.

Dave et al. [4], worked on semantic classification of reviews as positive or negative ones using the available corpus from web sites, where each review already had a class e.g., binary ratings or thumbs-up and thumbs-downs. Sentiment classifiers are build around them. However, the performance was limited because a sentence contains much less information than a review.

Gary Beverungen et al. [8], considered twitter posts and summarized them using clustering. Here the data set is limited as the twitter posts considered are not more than 140 characters. Hsinchu Chen et al. [9], considered only Wal-Mart data set statically and categorized the data as direct and indirect opinions.

Minqing Hu et al. [10], considered opinions posted by customers, identified the features and gave the sentiment without considering the dependencies in the opinions. Morinaga et al. [11], compared reviews of different products of one category to find about the target product. However, it does not summarize reviews, and it does not mine product features on which the reviewers have expressed their opinions.

B. Liu et al. [12] handbook categorized the Information into two types: facts and opinions. The features are classified as explicit features and implicit features. But the dependencies are not considered here.

In [15] research work, they improved the performance of calculations and classifications using linguistic rules and constraints. Here supervised and unsupervised learning techniques are used. Feature selection methods, Information Gain (IG) and Mutual Information (MI), were applied and compared. They have compared their work with Ding et al. [6] but the results shows that there is a fall in precision and recall rates which clearly state that these methods are not that accurate.

In [18], it takes one comment at a time, the dependencies in the text are not considered and also techniques used for sentiment classification are not mentioned

In [20], NLTK 2.0.1rc1 powered text classification process is done. When the text is entered it expresses whether the text is positive negative or neutral sentiment. It takes one comment at a time, but here the results are not so accurate.

Thus the existing works are limited to a particular site or a static data set. And the opinions are just classified as positive opinions and negative opinions without considering the dependencies. Naïve Bayes classifier is used for sentiment classification.

But the dependencies that exist within words used in the comments are not considered. Section III presents our proposed system for sentiment classification.

III. PROPOSED SYSTEM

The proposed system is a unique system which takes the data dynamically, classifies, ranks are given. These ranks may vary with in time and comments posted. Comments considered here are about mobile phones, cameras and laptops. Using this system the user can know the pro's and con's about a product.

Figure 1 presents the SCDDF architecture of our proposed system. The full length description of proposed system can be found in [13].

A. Preprocessing

Firstly comments are collected dynamically from the sites using web crawler [14] QTP. Then the data set collected is tokenized [17]. Stop words like "a", "this", "is", etc are removed and dependency words like "not", "no" are considered.

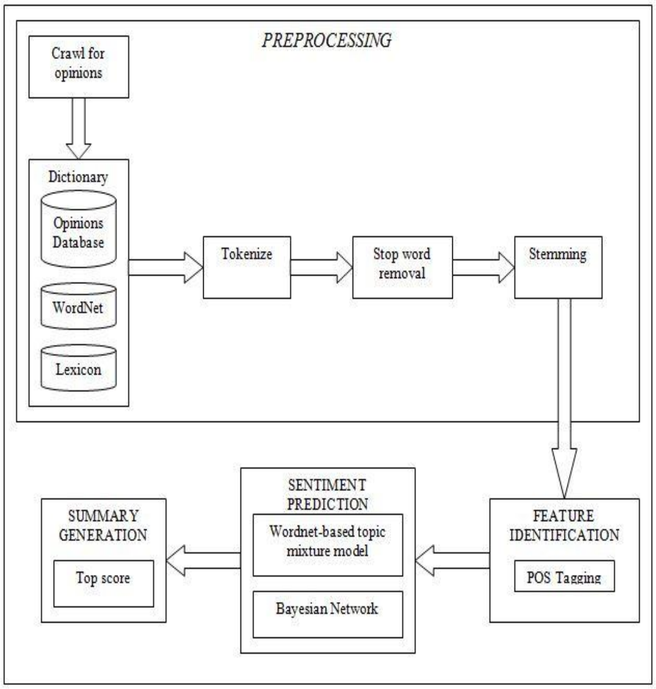


Fig. 1 Sentiment Classification for Dynamic Data Features (SCDDF).

At the end of preprocessing stemming is done. Stemming is the process where the words suffixes are removed. Porter stemmer [19] is applied for stemming. It is a 6 steps algorithm, where in each step the words are trimmed and the size of the data set will be reduced in each step.

B. Feature Identification:

In this step features are identified for the comments collected. Features like "battery", "touch" etc are identified in this step. For this process POS (Parts Of Speech) –tagging is adverbs are identified for feature identification.

Feature Identification step:

$P(O,T)=\P_i P(t_{i-1}\text{->}t_i)p(w_i|t_i)$;

where,

P(O): Opinions

P(t):Tags

P(O, T): Opinions with tags

P(w): probability of getting a word from word net

C. Sentiment Prediction:

In this step the sentiments for the comments i.e. positive and negative comments are predicted using wordnet [7, 21] and dependencies are resolved using the Bayesian network.

Example dependency comment:

This is not a great mobile.

Here "not" is a strong dependency word. Most of the works were dependencies are not considered says that the comment is a positive one as there is a positive word in it. But in actual sense it is a negative comment.

$$P(C \mid A \wedge B) = P(C \mid B) \qquad (1)$$

So by applying Bayesian networks these dependencies are resolved.

D. Summary Generation:

At last considering the scores obtained from sentiment prediction level the results generated are shown using statistical summary report. Statistical summary report consists of comments, features extracted for each comment, positive or negative score assigned along with the positive and negative label and rank of the product. Thus the user can evaluate the odds and outs of the product.

II. RESULTS

This section presents the experimented results of our proposed work SCDDF. To evaluate the performance of sentiment classification, we adopted four indexes that are generally used in text categorization:[5] Recall, Precision, F-measure and Accuracy. Performance is measured using the following metrics.

Experimental results shows that our method has produced an accuracy of 0.9 after preprocessing, 0.91 after feature identification and 0.918 after sentiment prediction for the product mobile. This shows that after each level the results are more refined to get accurate results.

Precision (P) = #Correct / #Guessed

Recall (R) = #Correct / #Relevant

Accuracy (A) = #Correct / # Total posts ; and

F-measure (F) = 2*Precision*Recall/ (Precision + Recall)

Table 1 presents the results of SCDDF when evaluated on sample data set.

size	Product	Pre-processing				Feature Identification				Sentiment Prediction			
		P	R	A	F	P	R	A	F	P	R	A	F
105	Mobile	0.91	0.92	0.905	0.915	0.914	0.93	0.91	0.922	0.921	0.935	0.918	0.928
60	Camera	0.917	0.922	0.91	0.92	0.924	0.931	0.92	0.927	0.931	0.94	0.932	0.935
60	Laptop	0.93	0.932	0.912	0.931	0.932	0.934	0.93	0.933	0.94	0.941	0.934	0.941

Table 1: The results of SCDDF when evaluated on sample data set.

Precision

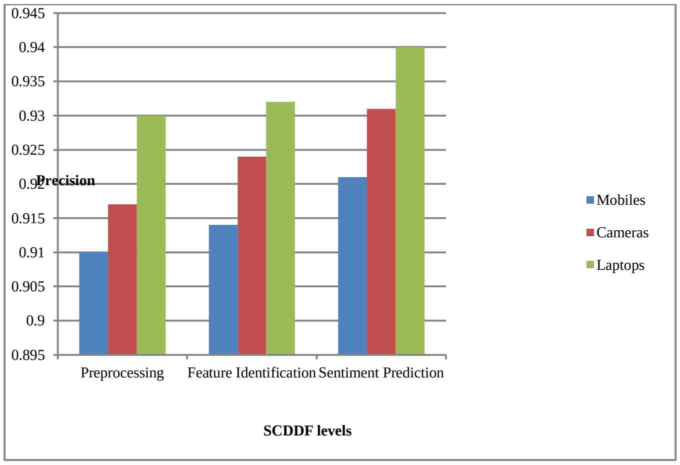


Fig 3: Precision obtained at each level.

Accuracy

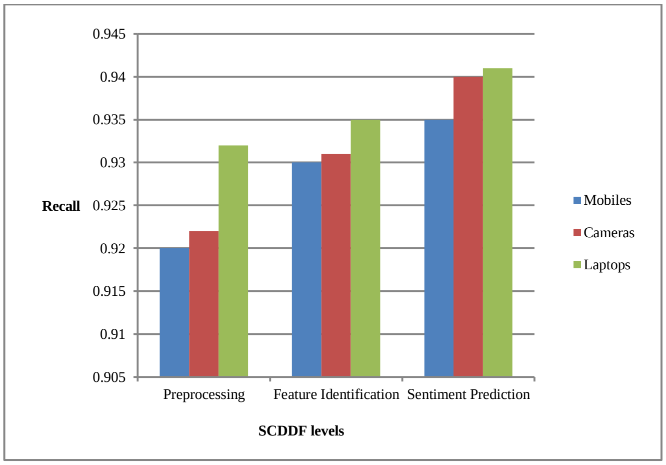


Fig 4: Recall obtained at each level.

Recall

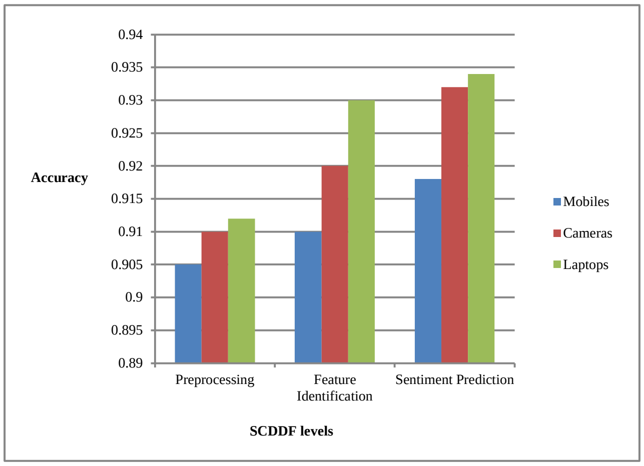


Fig 5: Accuracy obtained at each level.

F-measure

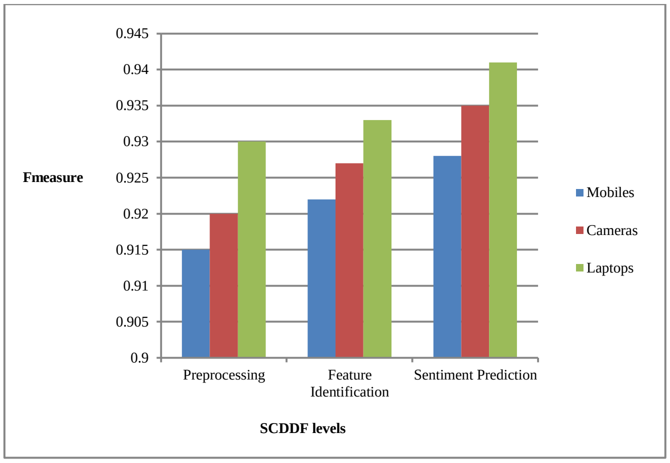


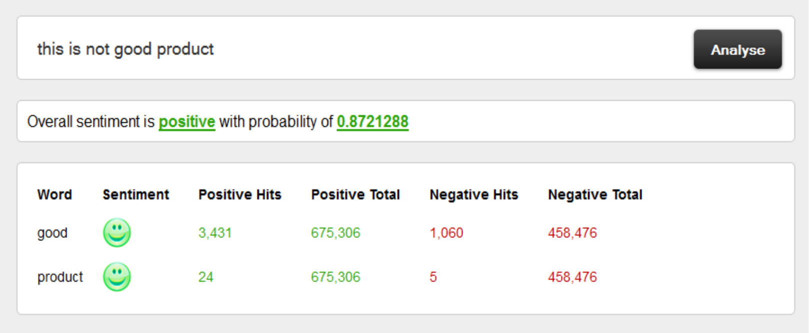
Fig 6: F-measure obtained at each level.

Similarly for camera and laptops data set, accuracy has been increased within each step. Figures 3, 4, 5,6 presents the comparison of the results with respect to each of the performance measure.

Table 2 presents the comparison of our method and online web tools. Figures 7,8,9 presents the snapshot of the execution of online tools.

Sample input comments	Sentiment analyser [18]	Nltk [20]	SCDDF [13]
fall in love on this phone! elegant design &amp; ideal specs. but i&#39;m gonna buy the international version &#39;cos the brand logo is on top, on the bellow! :D	Overall sentiment is positive with probability of 0.985837	The text is neutral.	pos 0.8901
it really sucks that the T-Mobile Version is coming out in 6 days and will have the Ics straight out of the box and that's not fair. I think they should wait just like the rest of us at the back of the line and let the originals get the up-dates first.	Overall sentiment is negative with probability of 0.1854791	The text is neg.	neg 0.3011
Its not that good compared to iphone	Overall sentiment is positive with probability of 0.7626124	The text is neg.	neg 0.2425
Its good compared to iphone	Overall sentiment is positive with probability of 0.7626124	The text is neg.	pos 0.7575

Table 2: Comparison of SCDDF with existing online tools.

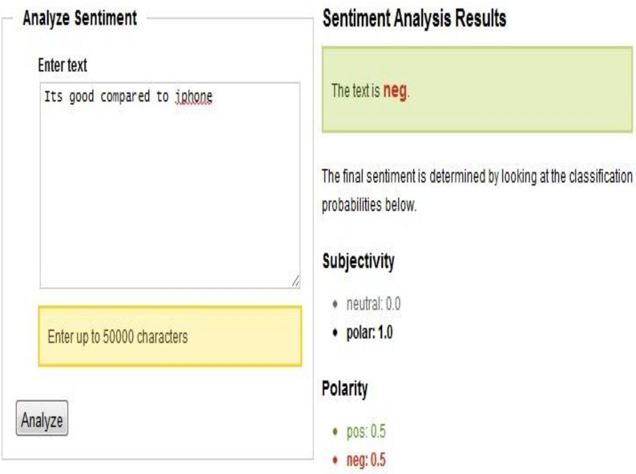


this is not good product

Overall sentiment is positive with probability of 0.8721288

Word	Sentiment	Positive Hits	Positive Total	Negative Hits	Negative Total
good	☺	3,431	675,306	1,060	458,476
product	☺	24	675,306	5	458,476

Fig 7: Snapshot of execution of [18]



Its good compared to iphone

Enter up to 50000 characters

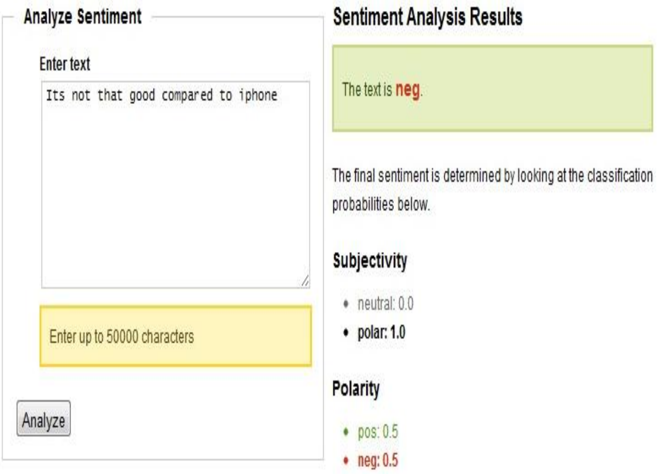
The text is neg.

The final sentiment is determined by looking at the classification probabilities below.

- neutral: 0.0
- polar: 1.0

- pos: 0.5
- neg: 0.5

Fig 8: Snapshot of execution of [20]



Its not that good compared to iphone

Enter up to 50000 characters

The text is neg.

The final sentiment is determined by looking at the classification probabilities below.

- neutral: 0.0
- polar: 1.0

- pos: 0.5
- neg: 0.5

Fig 9: Snapshot of execution of [20]

The above results shows that our proposed method performs in the same way as existing tools in some cases where as performs better in other cases. The results clearly show that the existing tools neglected the dependencies within the comments. And also one comment at a time are analyzed. But our method takes dynamic data considering the dependencies using the Bayesian networks.

Table 3 Compares our work with existing works [6, 15].

Table 3 shows that there is a clear fall in Yanyan Meng values compared with Ding et al. methods. Our method SCDDF has increase in values when compared with [6,15]. In [15] they just identified the product features using techniques like document vector, sentence vector, intensification and sentence relation. These methods are useful to find the polarities and features. Ding et al. [6] applied the rule-based sentiment analysis technique which just says about the opinion orientations and product features. Both [6,15] neglected the dependencies in the words.

IV. Conclusion

Feature subsumption for sentiment classification in social networks using natural language processing solves the problems in opinion mining and provides a novel approach for sentiment classification. It is a novel community-based evaluation that successfully captures the peculiarities of social networks.

However, the success of such an initiative eventually depends on the cooperation of the companies and institutions owning social network data, and on the agreement of enough organizations to participate in such a project.

V. Future Work

Our future work concentrates on classifying the sentiments of messages posted within the social networks such as Facebook, Twitter. This is required as in the recent times government is planning to involve a 3rd person to analyze the comments posted over such networks. Even though this is violation to human right but protects the society without leading to unwanted situations. But basic human rights should not be destructed; in this situation without causing harm to anyone our method can find sensitivity of the messages posted in the network.

References

[1] Abbasi et al., “Affect Analysis of Web Forums and Blogs using Correlation Ensembles,”IEEE Trans. Knowledge and Data Eng., vol. 20, no. 9, 2008, pp. 1168–1180.

[2] Ahmed Abbasi, “Intelligent Feature Selection for Opinion Classification”, University of Wisconsin-Milwaukee, - IEEE 2010.

[3] Cardie, C., Wiebe, J., Wilson, T. and Litman, D. 2003. Combining Low-Level and Summary Representations of Opinions for Multi-Perspective Question Answering. AAAI Spring Symposium on New Directions in Question Answering. 2003

[4] Dave, K., Lawrence, S., and Pennock, D., 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. WWW’03.

[5] David D. Lewis. Evaluating and optimizing autonomous text classification systems. In SIGIR '95, pages 246--254, New York, NY, USA, 1995. ACM Press.

[6] Ding, X., Liu, B. and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining, In Proceedings of the Conference on Web Search and Web Data Mining (WSDM).

[7] Fellbaum, C. WordNet: an Electronic Lexical Database, MIT Press 1998.

[8] Gary Beverungen and Jugal Kalita, “Evaluating methods for summarizing Twitter Posts”, WSDM’11, February 9-12,2011.

[9] Hsinchu Chen and David Zimbra, “AI and Opinion Mining”, University of Arizona, - IEEE 2010

[10] Minqing Hu and Bing Liu, “Mining Opinion Features in Customer Reviews”, American Association for Artificial Intelligence, 2004.

[11] Morinaga, S., Ya Yamanishi, K., Tateishi, K, and Fukushima, T. 2002. Mining Product Reputations on the Web. KDD’02.

[12] B. Liu, “Sentiment Analysis and Subjectivity,” Handbook of Natural Language Processing, 2nd ed., N. Indurkhya and F.J. Damerau, eds., Chapman & Hall, 2010, pp. 627–666.

[13] B. Jayanag et al., “A Study on Feature Subsumption for sentiment classification in Social Networks using Natural Language Processing Techniques”, Communicated to IJCA.

[14] Jeff Heaton, Programming Spiders, Bots, and Aggregators in Java, Publisher: Sybex, February 2002, ISBN: 0782140408

[15] Yanyan Meng Sentiment analysis: A study on product features, University of Nebraska.

[16] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[17] http://nltk.googlecode.com/svn/trunk/doc/api/nltk.tokenize-module.html

[18] http://sentiment.brandlisten.com/

[19] http://tartarus.org/martin/PorterStemmer/

[20] http://text-processing.com/demo/sentiment/

[21] http://wordnet.princeton.edu

Data set [16]	Yanyan Meng methods [15]						Ding et al. methods [6]			SCDDF [13]		
	intensification			sentence relation								
	P	R	F	P	R	F	P	R	F	P	R	F
Apex	0.66	0.63	0.64	0.63	0.65	0.64	0.89	0.88	0.89	0.91	0.9	0.91
CanG3	0.53	0.74	0.61	0.64	0.76	0.69	0.93	0.92	0.93	0.93	0.92	0.93
Nikcool	0.61	0.76	0.64	0.64	0.75	0.67	0.96	0.96	0.96	0.96	0.96	0.96
Nomp3	0.58	0.65	0.6	0.576	0.64	0.6	0.87	0.86	0.87	0.9	0.89	0.895
No6610	0.66	0.79	0.72	0.68	0.82	0.74	0.95	0.95	0.95	0.952	0.95	0.95

Table : Comparison SCDDF with [6, 15]

Application of Relevance Vector Machines in Real Time Intrusion Detection

Naveen N.C,
Associate Professor, Department of ISE,
R V College of Engineering, Bangalore
and Research Scholar, Dept of CSE,
SRM University, Chennai, India

Dr Natarajan.S,
Professor,
Department of ISE P E S I T,
Bangalore, India

Dr Srinivasan.R,
Professor Emeritus,
Department of Computer Science and Engineering,
M S R I T, Bangalore, India

Abstract— **In the recent years, there has been a growing interest in the development of change detection techniques for the analysis of Intrusion Detection. This interest stems from the wide range of applications in which change detection methods can be used. Detecting the changes by observing data collected at different times is one of the most important applications of network security because they can provide analysis of short interval on global scale. Research in exploring change detection techniques for medium/high network data can be found for the new generation of very high resolution data. The advent of these technologies has greatly increased the ability to monitor and resolve the details of changes and makes it possible to analyze. At the same time, they present a new challenge over other technologies in that a relatively large amount of data must be analyzed and corrected for registration and classification errors to identify frequently changing trend. In this research paper an approach for Intrusion Detection System (IDS) which embeds a Change Detection Algorithm with Relevance Vector Machine (RVM) is proposed. IDS are considered as a complex task that handles a huge amount of network related data with different parameters. Current research work has proved that kernel learning based methods are very effective in addressing these problems. In contrast to Support Vector Machines (SVM), the RVM provides a probabilistic output while preserving the accuracy. The focus of this paper is to model RVM that can work with large network data set in a real environment and develop RVM classifier for IDS. The new model consists of Change Point (CP) and RVM which is competitive in processing time and improve the classification performance compared to other known classification model like SVM. The goal is to make the system simple but efficient in detecting network intrusion in an actual real time environment. Results show that the model learns more effectively, automatically adjust to the changes and adjust the threshold while minimizing the false alarm rate with timely detection.**

Keywords- Intrusion Detection; Change Point Detection; Relevance Vector Machine; Outlier Detection.

I. INTRODUCTION

Reasonable level of security is provided by static defense mechanisms such as firewalls and software updates. Dynamic mechanisms can also be used to achieve security such as IDS and Network Analyzers (NA). The main difference between IDS and NA is that IDS aims to achieve the specific goal of detecting attacks whereas NA aims to determine the changing trends in network of computers [1] [18]. Earlier work emphasized that data can be obtained by three ways using real traffic, sanitized traffic and simulated traffic. But in real time fast response with reduced false positives to external events within an extremely short time is demanded and expected [19]. Therefore, an alternative algorithm to implement real time learning is imperative for critical applications for fast changing environments. Even for offline applications, speed is still a need, and a real time learning algorithm that reduces training time and human effort to nearly zero would always be of considerable value. Mining data in real time is still a big challenge [21].

IDS involve automatic identification of unusual activity by collecting data, and comparing it with reference data. An assumption of IDS is that a network’s normal behavior is distinct from abnormal or intrusive behavior, which can be a result of various attack/s. In this research work, flow analysis is used for network traffic analysis which searches for behavioral characteristics in a flow. There are various characteristics such as transferred bytes, packets, flow length, inter-arrival times, inter-packet gaps, etc that are monitored and computed. Data that is collected from flows can be used on high-speed networks as there is no deep packet inspection.

In this paper a hybrid approach for improving the performance of detection algorithm by building more intelligence to the system is proposed. In this direction CP detection is considered for discovering change points if properties of network behavior change. CP is the change in characteristics that occur very fast with respect to the sampling period of the measurements, if not instantaneously. The detection of changes refers to tools that help to decide whether such a change has occurred in the characteristics or not. Outlier Detection is a major step in Data Mining (DM) problem which discovers abnormal or deviating data points with respect to distribution in data [20]. Outliers are often considered as an error or noise although they may carry very important information.

A real time detection system is one in which network intrusion detection happens while an attack is occurring. A real time IDS captures the present network traffic data which is on line data. Bayesian learning algorithms, like RVM allow the user to specify a probability distribution over possible parameter values from the learned classifier. This will provide one solution to the over fitting problem as the algorithm can use prior distribution to regularize the classifier.

II. Related Work

Research shows that many Machine Learning (ML) techniques can be used for data classification. It is presented that popular supervised learning techniques gives high detection accuracy for IDS. A wide range of real world applications are discussed in the community of Statistical Analysis and DM [3] [13]. Statistical techniques usually assume an underlying distribution of data and require the elimination of data instances containing noise. Statistical methods though computationally intense can be applied to analyze the data [4]. Statistical methods are widely used to build behavior based IDS. The behavior of the system is measured by a number of variables sampled over time such as the resource usage duration, the number of processors, memory disk resources consumed during that session etc. The model keeps averages of all the variables and detects whether thresholds are exceeded based on the standard deviation of the variable. Very few on line (real time) network IDS approaches are proposed until now. Liao and Vemuri [5] develop a real time IDS using Self Organizing Maps (SOM) and preprocess their dataset with 10 features for each data record containing information of 50 packets. M. Al-Subaie [6] uses Hidden Markov Models over Neural Networks in anomaly intrusion detection to classify normal network activity and attack using a large training dataset. The approach was evaluated by analyzing how it affected the classification results. Ben Amor [7] designs a real time IDS using Naïve Bayes and Decision Trees and the results show that the Naive Bayes gives higher detection speed and detection rate than the Decision Trees. Authors in [8] [14] propose a hybrid intelligent systems using Decision Trees (DT), SVM and Fuzzy SVM for anomaly detection (unknown or new attacks). The results show that the hybrid DT–SVM approach improves the performance for all the classes when compared to a SVM approach.

SVM proposed by (Burges 1998, Cortes and Vapnik 1995) is a supervised learning algorithm that is used increasingly in IDS. The classification performance of SVM model is better than the classification methods, such as ANN [9]. The benefit of SVMs is that they learn very effectively with high dimensional data. Rung Ching Chen [10] uses Rough Set Theory (RST) and SVM to detect intrusions. Initially, RST is used to preprocess the data and reduce the dimensions. Later, the features selected by RST are sent to SVM model to learn and test. This method proves to be effective and also decreased the space density of data. The SVM is one of the most successful classification algorithms in the DM area [17].

RVM, proposed by Tipping [11] is a sparse machine learning algorithm that is similar to the SVM in many respects. It is capable of delivering a fully probabilistic output and it is proved to have nearly identical performance to, if not better than, that of SVM in several benchmarks. Di He [12] proposes an IDS approach based on the RVM where a Chebyshev chaotic map is introduced as the inner training noise signal. The result shows that the approach can reach higher detection probabilities under different kinds of intrusions and the computational complexity reduces efficiently. Li Rui [16] improves the generalization performance of RVM by an incremental relevance vector machine algorithm and the results are better than RVM and SVM. This guarantees the reliability of using RVM based approach for designing IDS. RVM has a better generalization performance than SVM due to the less support vectors.

III. Methodology

Over 90% of Internet traffic uses the Transmission Control Protocol (TCP). Because of its widespread use and its impressive growth, the research focuses on the detection of anomalous behavior within TCP traffic. Exploring the TCP packet attributes would enable a classifier to identify normal and abnormal activity on a packet-by-packet basis. From these attributes, a decision tree is built which will enable to identify and classify different attacks and violations. The process of building a classifier model using RVM is depicted in Fig. 1.

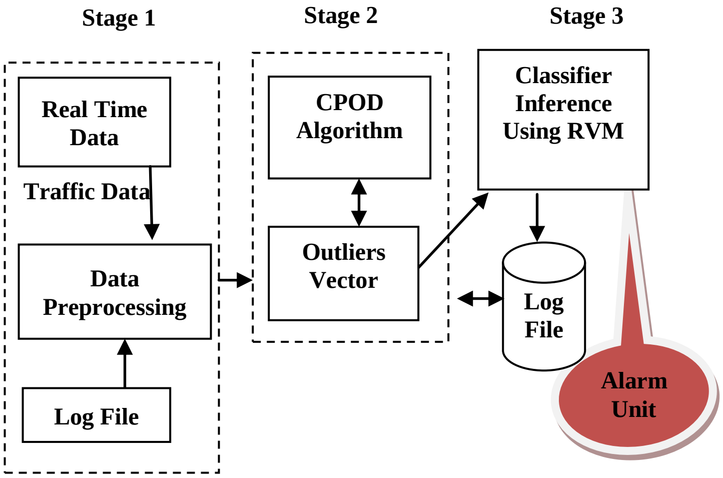


Figure 1. Architecture of the model

A. Dataset Description

The first stage of the implementation involves in training the system. For the present problem data is collected from the campus network to measure the accuracy and attacks. Data such collected is preprocessed and used to detect change point in network performance characteristics. Traffic in the network results in continuous change as the user's login and make use of Internet. For capturing the packets in real time JPCAP and WINPCAP tool is used to collect the information that is being transmitted. JPCAP provides facilities to capture and save raw packets live. It can automatically identify packet types and can generate corresponding Java objects for Ethernet, IPv4, IPv6, ARP/RARP, TCP, UDP, and ICMPv4 packets. Packets can also be filtered according to the user requirement. JPCAP is developed on LIBPCAP / WINPCAP, which is implemented in C and Java and is the industry-standard tool for link-layer network access. In Windows environment WINPCAP allows applications to capture and transmit network packets bypassing the protocol stack. The network data is collected from the interface which is capable of capturing information flowing within the local network. For example, anomalies can be detected on a single machine, a group of network a switch or a router. For the current research work the TCP/IP packet is collected in real time from the research lab network and dumped for further process. The Data Preprocessing phase handles the conversion of raw packet or connection data into a format that algorithms can utilize and store the results in the knowledge base. Rather than operating on a raw network dump file, the algorithm uses summary information to perform the analysis. Data is preprocessed to generate summary lines about each connection found in the dump file. The resulting summary file is then parsed and processed by the algorithm to give a count to each data/each time point, with a higher score indicating a high possibility of being an outlier/a change point.

The Log File stores the data as rules produced by the detection algorithm for further mining process. It may also hold information for the preprocessor, such as patterns for recognizing attacks and conversion templates. This Training Data is responsible for generating the initial rule sets that needs to be used for deviation analysis. It can be triggered automatically based on time or the amount of pre-processed data available.

The proposed Outlier Detection Algorithm examines the network data and creates a description of differences and stores in the outlier vectors for further reference. If a deviation is detected it signals the alarm unit. A strategy for invoking the deviation analyzer is by querying periodically the outlier vectors for the new profiles. Also the profiler may signal when a new profile is added and the Alarm Unit is responsible for informing the administrator when the deviation analyzer reports unusual behavior in the network stream. This can be in the form of SMS, e-mails, console alerts, log entries etc.

In the data preprocessing step as shown in Figure 1, packets are captured using JPCAP library and information is extracted that includes IP header, TCP header, UDP header, and ICMP header from each packet. After that, the packet information is partitioned and formed into a record by aggregating information every 30 minutes. Each record consists of data features considered as the key signature features representing the main characteristics of network data and activities.

IV. Proposed Algorithm

A. *Change Point Outlier Detection (CPOD):*

To illustrate the problem, network data collected is observed with threshold values in the college local area network at regular intervals of time window for a certain period of time. Most of the threshold variations may be statistically regular, but once a while there may be an outlier point i.e. a marked deviation from the previous data. In general detecting such outliers is important because they may be caused by an anomaly within the network or from the external environment.

B. *CP Distribution*

Since the CP can occur randomly and at different times the current model assumes that the number of CP m is fixed and treated as an unknown random variable.

If there are n observations which are denoted by $o_1, o_2, \ldots,$ o_n, and a set of CP denoted by $0 < c_1 < c_2 < \cdots < c_m < n$, where the number of CP m is unknown it is assumed that the CP occur at discrete time, $1, \ldots, n - 1$.From this a prior distribution of the CP is considered.

C. *Sampling stage for CP*

1. At time n the algorithm starts
2. N data collected is sampled for time t without attack
3. Within this data sample if any CP is detected it is saved as a training data with parameter W_i
4. These values are updated to the training database

This algorithm has a computational cost of O(n). The following are the requirements of CPOD algorithm. The statistics should be on line meaning an outlier has to be detected as it appears. A change point has to be detected within some constant number of observations after the change happens. Specific assumptions regarding distributions are not done and hence the detection can be adaptive to a non-stationary time series and robust to a wide variety of distributions.

D. *The CPOD Algorithm*

We denote a data sequence as $\{x_i : i = 1, 2, \ldots N\}$, i is the time variable, p denotes the number of data points to be observed before initiating analysis. Cx_i denotes the current data point in the time series being considered for analysis. t and s denotes the thresholds for change point and outlier detection respectively. Threshold value is the mean magnitude of fluctuation allowed in the data points within which they won’t be classified. Window size w, v denotes the number of data points to consider for computing the median and meaning respectively. The algorithm chooses median over a small window, as it is less sensitive to outliers and helps in localizing the deviation. $w < v$ in all cases, $(w/v) < 0.5$ is found optimal. We maintain a vector to signify the classification state of each data point. States could be {0,1,2,3} where ‘0’ means neither outlier nor change point, ‘1’ means outlier, ‘2’ means outlier with high probability that previous point was the change point, and ‘3’ means the change point done from previous two observations.

If the thresholds are well tuned, CPOD can detect maximum outliers and change points in a time series.

CPOD Algorithm (Input : p,s,t,w,v)

Step 1: The iteration for all data points is done after initializing i=p+1

$$\forall\ (\mathbf{i > p | p < w < v < (N - i)})$$

Step 2: The median and mean over the windows v, w is computed respectively as in (1) and (2)

$$C\,\tilde{x}_i = \tilde{x}_{i-1}^{\,i-w} \qquad (1)$$

Mean of values in window ’v’

$$C\,\bar{x}_i = \bar{x}_{i-1}^{\,i-v} \qquad (2)$$

Step 3: Score1 is the ratio of absolute difference between the current data point from the median to the mean amplified by the threshold and calculated as

$$\text{Score}_{1i} = (|\text{Cx}_i - \text{C}\,\tilde{x}_i\,| * t) / \text{C}\,\bar{x}_i \text{ and} \qquad (3)$$

Score 2 is the normalized ratio of two distance magnitudes i.e. the median from mean and the mean from the current data point and calculated as

$$\text{Score}_{2i} = (|\text{C}\,\tilde{x}_i - \text{C}\,\bar{x}_i\,|) / (|\text{C}\,\bar{x}_i - \text{C}\,\tilde{x}_i\,|) * 100 \quad (4)$$

The median is over the short term window w and mean over the longer term window v. By taking ratio, makes the score robust to fluctuations that may happen in the mean over a long term. We heuristically found w/v < 0.5 to be the best choice. The resolution of detection is dependent on the threshold.

If Score1 > Score2, we classify the point as an outlier. Score2 is used as a data-dependent cutoff to classify Score1 as outlier or not. Score1 is sensitive to mean and to the deviation of current data point from median. Score2 is sensitive to deviation of current point from mean and to deviation of mean from median. In window v, Score1 will be greater than Score2, with the presence of outliers or with data points subsequent to a change point.

Step 4: A stronger possibility of current data point being an outlier or CP is indicated if Score1 is higher than Score2. To classify the current data point as CP an additional check is made to find if the point lies beyond a certain band around the median represented as LL_i and UL_i. The classification state is then saved in vector V.

$$\text{Score}_{1i}\begin{cases} > Score_{2i} \wedge (Cx_i < LL_i \vee Cx_i > UL_i) : V_i = 1 \\ \leq Score_{2i} \vee (LL_i > Cx_i > UL_i) : V_i = 0 \end{cases} \qquad (5)$$

Step 5: State information in vector V is used to classify outlier and CP. If the current point has a higher Score1 as indicated in V_i the past three states are considered for classification. Vector V_{si} is used to express the sum state of V_i,V_{i-1} and V_{i-2}. V_{si} stores state of the current data point with respect to past two data points. If the value of V_{si} is 3 it indicates that outliers were detected in the past and current point could be the change point. We test the previous states to make sure there was no change point detected in the past two data points. If detected then the CP is inferred as an outlier. Similarly if the value of V_{si} is 1 then it is possible that current point is an outlier as shown in (6) and (7).

$$\mathbf{V}_i \begin{cases} \mathbf{= 1 \ : \ V_{si} = (V_{\mathit{i}} + V_{\mathit{i-1}} + V_{\mathit{i-2}})} \\ \qquad\qquad\qquad \mathbf{= 0 \ : \ V_{si} = 0} \end{cases} \qquad \mathbf{(6)}$$

$$\mathbf{V_{si}} \begin{cases} \mathbf{= \ 3 \ \wedge (V_{s_{\mathit{i-1}}} = 3 \ \vee \ V_{s_{\mathit{i-2}}} = 3) \ \ : \ V_{s_{\mathit{i}}} = 1} \\ \mathbf{= \ 1 \ \wedge (V_{s_{\mathit{i-1}}} = 1 \ \vee \ V_{s_{\mathit{i-2}}} = 1) \ \ : \ V_{s_{\mathit{i}}} = V_{s_{\mathit{i}}} + V_{s_{\mathit{i}}} - 1} \end{cases} \qquad \mathbf{(7)}$$

Finally Cx_i is classified depending on the values of V_{si}. If the state of current point is 0, then there is no significant deviation. If the current point is 1 or 2 and the current data point deviates more than s% threshold it is inferred as an outlier and signifies a higher possibility of Cx_i - 1 being a CP. If the state of current point is 3, with accuracy it is inferred that two points prior to current one is the CP.

$$\mathbf{V_{si}} \begin{cases} \mathbf{= \ 0 : No\ change\ ,\ adapt\ LL,\ UL\ to\ C}\,\tilde{x}_i & \mathbf{(8)} \\ \mathbf{= \ (1 \vee 2) \wedge (\ Cx_{\mathit{i}} > (Cx_{\mathit{i}} + s\ \%\ Cx_{\mathit{i}}\,)\,) \ : Outlier} \\ \mathbf{> \ 2 : Change\ point,\ adapt\ LL,\ UL\ to\ change} \end{cases}$$

Step 6: The classification of outliers and change points as signified in the V_{si} vector is reported. The scores and state elements of vector V, V_s for past data points N, N-1 and N-2 are persisted. Persistence of median and mean over the window sizes while classifying current data point enables online implementation.

Table I List of Network Dataset Features Collected

1. appName	10. direction
2. totalSourceBytes	11. sourceTCPFlagsDescription
3. totalDestinationBytes	12. destinationTCPFlagsDescription
4. totalDestinationPackets	13. source
5. totalSourcePackets	14. protocolName
6. sourcePayloadAsBase64	15. sourcePort
7. sourcePayloadAsUTF	16. destination
8. destinationPayloadAsBase6	17. destinationPort
9. destinationPayloadAsUTF	18. startDateTime
	19. stopDateTime

RVM is currently of much interest in the research community as they provide a number of advantages. RVM is based on a Bayesian formulation of a linear model with an appropriate prior that results in a sparse data representation. As a result, they can generalize well and provide inferences at very low computational cost. Many applications like object detection and classification, target detection in images, classification of micro calcifications from mammograms etc are developed. RVM produces a function which is comprised of a set of kernel functions also known as basis functions and a set of weights. This function represents a model for the system presented to the learning process from a set of training data set. The kernels and weights calculated by the learning process and the model function defined by the weighted sum of kernels are fixed. From this set of training vectors the RVM selects a sparse subset of input vectors which are deemed to be relevant by the probabilistic learning scheme. This is used for building a function that estimates the output of the system from the inputs. These relevant vectors are used to form the basis functions and comprise the model function.

In the classification phase each of the network data selected from the feature selection phase is classified as normal data or attack data. This phase consists of two main dataset which are used for training and testing. The log file contains four weeks of training data and one week of testing data. A total of 19 features are captured as listed in Table I. During the first phase training is performed using RVM with a set of network records with known answer classes. Based on the training the IDS model can classify the data in each record into normal network activity or main attack types. Then the model is tested with new or untrained dataset where each record was captured in a real time environment in the college research lab.

For an input vector x, an RVM classifier models the Probability distribution of its class labeled C ε (1, +1} using logistic regression as

$$p\left((C=1\middle|x)=\frac{1}{1+\exp(-fRVM(x))}\right) \qquad \textbf{(9)}$$

where $f_{RVM}(x)$ the classifier function is given by,

$$f_{RVM}(x) = \sum_{i=1}^{N} \alpha_i K(x,x_i) \qquad \textbf{(10)}$$

where K(.,.) is a kernel function, and x_i, i = 1,2,...,N, are training samples. The parameters α_i, i 1, 2, ..., N, in $f_{RVM}(x)$ are determined using Bayesian estimation, introducing a sparse prior on α_i The parameters α_i are assumed to be statistically independent obeying a zero-mean Gaussian distribution with variance λ_i^{-1}, used to force them to be highly concentrated around zero, leading to very few nonzero terms in $f_{RVM}(x)$.

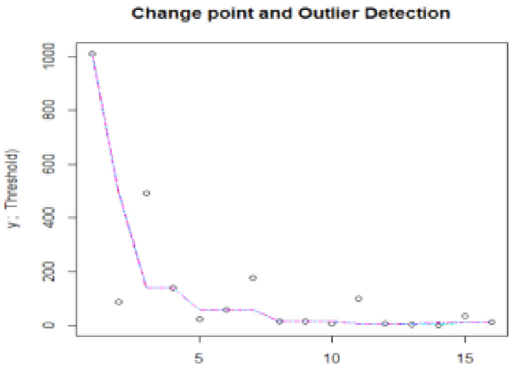
V. Experimental Results

Sample statistics of IP addresses and their count observed in our research laboratory for a time period of 30 minutes specified as window 'w_1', 'w_2', 'w_3'.

Table II Sample statistics of IP addresses

Source IP Address	***Time Window 'w_1'***	***Time Window 'w_2'***	***Time Window 'w_3'***
172.16.30.28	1010	1011	2000
172.16.30.91	86	86	90
172.16.30.75	415	492	512
172.16.30.108	140	140	140
172.16.30.70	24	24	24
172.16.30.92	58	58	58
172.16.30.68	175	175	179
172.16.30.69	14	14	14
172.16.30.96	14	14	14
172.16.30.35	7	7	7
172.16.30.95	100	100	100
172.16.30.101	11	12	12

For the real time statistics the data was collected for one week and the graph is as shown in Figure 2 and 3.



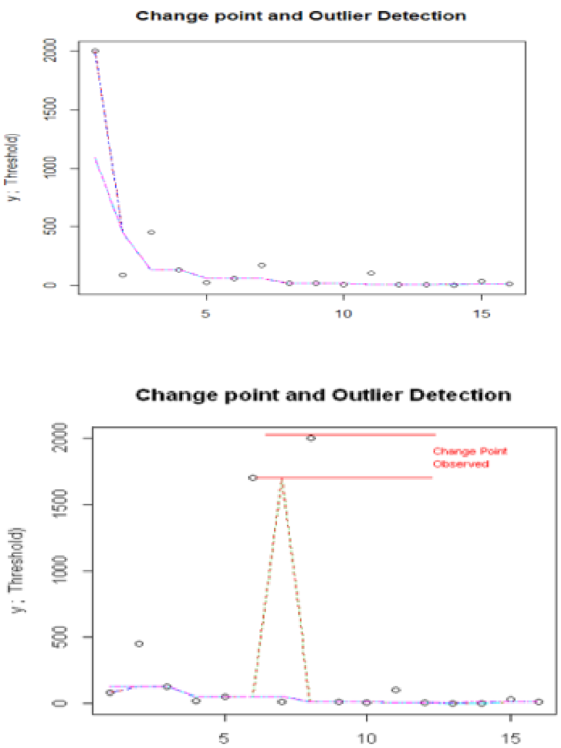
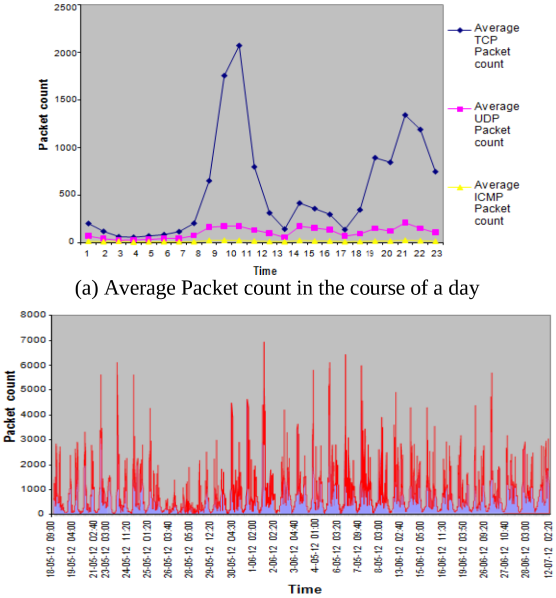


Figure 2. Plot of real time data and detection of change point and outliers



(a) Average Packet count in the course of a day

(b) Traffic Statistics in the course of a month

Figure 3. Plot of real time data collected for 2 weeks

Table III Comparison between RVM and SVM models

Model	***Number of training data***	***Number of vectors***	***Testing Performance***
SVM	100	25	0.71
	500	129	0.72
	1000	230	0.81
	5000	540	0.82

RVM	100	17	0.76
	500	109	0.81
	1000	170	0.86
	5000	240	0.88

Table III shows the comparison between the SVM and RVM models. The value of testing performance from RVM model is effectively same as that of SVM with lesser support vectors. The performance of the RVM is also better than the SVM.

VI. Conclusion

In this research work a solution for classifying outliers and change points from real time data is proposed which is addressed in two parts: scoring and classification. Scores are computed that reflect outliers and incrementally discover to keep the state of outliers in data series. The algorithm is characterized in its property to address outliers and change points at the same time. This enables to deal with frequent and fast changes in the source. The current implementation and usage indicates the success of the algorithm. This gives a unifying view of outlier detection and change point detection in real time network data.

Usage of RVM shows a competitive accuracy maintaining its sparseness ability. Experimental results show that the RVM model achieves essentially the same performance with a much sparser model as a previously developed SVM model. The much reduced computational complexity in RVM makes it more feasible for real time processing while designing IDS. The proposed method is competitive with respect to processing time and allows the use of selected training data set. The result shows an improvement in RVM classification performance. In this work, the design and successful implementation of a system with outlier detection was done.

References

[1] Olin Hyde, Machine Learning For Cyber Security at Network Speed & Scale, 1st Public Edition: October 11, 2011

[2] Iftikhar Ahmad, Azween Abdullah and Abdullah Alghamdi, Towards the Selection of Best Neural Network System for Intrusion Detection, International Journal of the Physical Sciences Vol. 5(12), October, 2010 , pp. 1830-1839

[3] Fabio Pacifici, Change Detection Algorithms: State of the Art, v1.2, Earth Observation Laboratory, Tor Vergata University, Rome, Italy, Feb 28, 2007

[4] G. Mohammed Nazer, A. Arul Lawrence Selvakumar, Current Intrusion Detection Techniques in Information, European Journal of Scientific Research, EuroJournals Publishing, Inc. 2011, pp. 611-624

[5] Liao Y, Vemuri VR. Use of K-nearest Neighbor Classifier for Intrusion Detection. Computers & Security 2002;21: 439–48.

[6] M. Al-Subaie and M. Zulkernine. Efficacy of Hidden Markov Models Over Neural Networks in Anomaly Intrusion Detection. In 30th Annual International Computer Software and Applications Conference (COMPSAC'06), pages 325–332, 2006.

[7] N. Ben Amor, S. Benferhat and Z. Elouedi. Naive Bayes vs Decision Trees in Intrusion Detection Systems. In SAC '04: Proceedings of the 2004 ACM symposium on Applied computing, pages 420–424, New York, NY, USA, 2004. ACM. ISBN 1-58113-812-1.

[8] Sandhya Peddabachigari,Ajith Abraham, Crina Grosanc,Johnson Thomas, Oklahoma State University, Modeling Intrusion Detection System Using Hybrid Intelligent Systems, Journal of Network and Computer Applications, Elsevier Ltd, 2005

[9] Zhiqiang ZHANG, Jianzhong CUI, Network Intrusion Detection Based on Robust Wavelet RVM Algorithm, Journal of Information & Computational Science, 2011, pp. 2983–2989

[10] Rung-Ching Chen, Kai-Fan Cheng, Chia-Fen Hsieh, Using Rough Set and Support Vector Machine for Network Intrusion Detect**,** International Journal of Network Security & Its Applications (IJNSA), Vol 1, No 1, April 2009

[11] M. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine", Journal of Machine Learning Research, 2001, pp 211 - 214.

[12] Di He, Shanghai Jiao Tong University , Improving the Computer Network Intrusion Detection Performance Using the Relevance Vector Machine with Chebyshev Chaotic Map, IEEE, 2011

[13] Reza Sadoddin, Ali A. Ghorbani, "An Incremental Frequent Structure Mining Framework for Real-Time Alert Correlation", Computers & Security, 2009, pp 153 – 173

[14] Shaohua Teng, Hongle Du, Naiqi Wu, Wei Zhang, Jiangyi Su, "A Cooperative Network Intrusion Detection Based on Fuzzy SVMs", Journal of Networks, Vol. 5, No. 4, April 2010

[15] Phurivit Sangkatsanee, Naruemon Wattanapongsakorn, Chalermpol Charnsripinyo, "Practical Real-Time Intrusion Detection Using Machine Learning Approaches", Computer Communications , 2011, pp 2227–2235

[16] Li Rui, "Computer Network Attack Evaluation Based on Incremental Relevance Vector Machine Algorithm", Journal of Convergence Information Technology, JCIT, Volume7, Number1, January 2012

[17] Javier M. Moguerza and Alberto Munoz, "Support Vector Machines with Applications", Statistical Science, 2006, Vol. 21, No. 3, pp 322 – 336

[18] Chenfeng Vincent Zhou, Christopher Leckie, Shanika Karunasekera, "A Survey of Coordinated Attacks and Collaborative Intrusion Detection", Computers & Security, 2010, 124 – 140

[19] Georgios P. Spathoulas, Sokratis K. Katsikas "Reducing False Positives in Intrusion Detection Systems", Computers & Security, 2010, pp 35 – 44

[20] Anna Koufakou, Michael Georgiopoulos, "A Fast Outlier Detection Strategy for Distributed High-Dimensional Data Sets with Mixed Attributes", Data Mining Knowledge Discovery, 2010, pp 259–289

[21] Wu, S., & Yen, E., "Data mining-based intrusion detectors", Expert Systems with Applications, 2009, pp 5605–5612.

Convenience and Medical Patient Database Benefits and Elasticity for Accessibility Therapy in Different Locations

Bambang Eka Purnama
Faculty of Informatics Engineering
Surakarta University, Surakarta, Indonesia

Sri Hartati
Faculty of Mathematics and Natural Science
University of Gadjah Mada Yogyakarta, Indonesia

Abstract—**When a patient comes to a hospital, clinic, physician practices or other clinics, the enrollment section will ask whether the patient in question had never come or not. If the patient in question said he had never come then the officer will ask you Medication Patient Identification Card (KiB), which will be used to search for patient records in question. In the conventional health care, then the officer will use a tracer to locate patient records at the storage warehouse in the form of stacks of paper. If a patient at a hospital is still a bit it will not be problematic, but if the patient sudha achieve large-scale number in the hundreds of thousands or even millions it will certainly cause problems.**

Database records are kept in hospital untapped to the maximum to be exchanged at another hospital when the patient arrives at another hospital for further treatment or research purposes. This study aims to produce a computerized model of inter Medical Information Systems Hospital. Facilitate the benefits of this research is in the medical records of patients get information, patient history properly stored in computerized medical records, patient data search can be found quicker resulting in faster unhandled The expected outcome of this research is rapidly tertanganinya patients coming to a clinic and when the patient comes to the clinic to another place then the patient's medical resume database and the analysis can be found immediately.

Keywords- Patient Medical Record.

I. Background Problem

Medical Record management activities will produce data and information in the form of indicators to be used as the evaluation of hospital services. Medical record service system goal is to provide information to facilitate management of the service to patients and facilitate managerial decision-making by the provider of clinical and administrative health care facilities. Therefore we need good data management RM start of input, process and output.

But the RM data management activities are currently running there are still some problems that the patient data input, written by officers in TPPRJ not complete, the process (data management still done conventionally) and output (reports / information only in the form of the ratio of old and new patient visits , the ratio of patient visits and specialists poly) so that the evaluation activities undertaken by service managers in particular to determine the productivity of outpatient services to be obstructed.

Hospital Information System should be able to contribute to all activities of the hospital management. Management information system of a hospital not only serve the statistical data requirements alone but should be able to generate useful information for medical decision-making process. In addition to the medical record contains information about all patients who had been treated, it can also be used as a reference when the patient was treated again. Health workers will be difficult to take action if not yet know the history or the history of the patient's disease before the action is recorded in the previous medical record file. One other important matters contained in the medical record file is the availability and completeness of its contents when needed.

As an initial illustration, there is a case study conducted by Clarke toxicology section, hospital, Edinburgh, UK that is usually used for inpatient services. The main function of the toxicology section is to provide medical care. In practice there is often confusion between the team of physicians and delivery of health services to patients. A patient should be placed in service where and treated by a doctor who? The problem is long and difficult quest to collect patient data is fragmented. Telephone and the conversation became a fact of discussion and exchange of information to get the conclusion of patient care. But after using EMR, EMR servants saw enough to get the patient's medical summary and decided conclusion. Prihartono (2008)

Electronic Medical Records are required to provide web-based system, easy to use and requires no investment in costly infrastructure and resources. The system is also required to have electronic prescriptions, receiving lab results electronically, using a structured data and nomenclature are given by the SNOMED (Systematized Nomenclature of Medicine), NDC (National Drug Code), or other data for documentation and have the ability to generate clinical data , administrative and demographic reports. Gates and Roeder (2011)

Electronic Medical Record is widely used in various hospitals in different parts of the world to replace or complement the medical record file form. Since the development of e-Health, EMR is at the heart of information from hospital information systems. Prihartono (2008) Electronic Medical Records (EMRs) is a computerized medical information system that collects, stores and displays patient information. It is a means to create a legible and terorganisirnya records and access to clinical information about patients. Furthermore EMRs are intended to replace the existing system (often paper-based) medical records which are familiar to practitioners. Patient records have been kept in paper form in a long time, they had consumed the greater space and delaying access to medical care becomes less efficient. In contrast, EMRs storing individual patient clinical information electronically and allows instant availability of this information to all providers in the chain of health and assist in providing a coherent and consistent care. Although expectations are high and interest in EMRs across the world, their overall adoption rate is relatively low and facing some problems. For example, employment is deemed contrary to the traditional style of a doctor, they need a greater ability in handling the computer and install a system that absorbs sufficient financial resources. Boonstra and Broekhuis (2010)

With the EHR enables the implementation of an increasingly complex cross-communication among health professionals with various parties who are both providing care for patients in health care facilities, EHR can also be used as one important input in assessing the success of health programs at institutions of existing services. (Minister of Health RI, 2005).

A clinical information system is a collection of various information technology applications that provide a centralized repository of information related to patient care across distributed locations. This repository is the patient's disease history and interactions with coding knowledge provider that can help doctors decide on the patient's condition, treatment options and medical procedures. The Repository also encodes status decisions, actions taken to decision-making and relevant information that can help in performing the act. The database can also store information about patients, including genetic, environmental and social context. Sittig et al (2002)

Leading health organizations have emphasized the importance of integrating information technology into the healthcare system to improve provider practices, improve the quality of patient care and reduce medical errors. One of the problems that interfere with the spread of technology into health care is how to combine the practical, clinical information systems can be used in the work environment of providers. Alexander (2008)

Electronic Medical Record technology allows medical providers to store and exchange medical information using a computer. Although the technology has been available since the 1970s, only 50% of hospitals are adopting the basic EMR system in 2005. Slow diffusion of EHR that has attracted attention, since the adoption of EMR could reduce U.S. $ 1,900,000,000,000 annual U.S. health care bill through increased efficiency and comfort. Although some hospitals adopted EMR necessary for the transfer of electronic information, but also must cooperate and coordinate cross hospital. In 2006 an eHealth Initiative survey (Covich Bordenick, Marchibroda and Welebob (2006)) identified more than 165 active Health Information Exchange initiatives in the U.S., 45 are being implemented and 26 are fully operational. Miller and Tucker (2007)

Data warehouse modeling method has been used in the industry standard for years for decision support in various fields. Data warehouse design for healthcare industry outside well understood and has been widely discussed. Healthcare is still far behind, in the field of data warehouse management, decision support and the need to move forward in this direction. Parmanto et al (2005)

With the awareness of medical errors and increase the focus on improving the quality of patient care, President George W. Bush (American President Currently it is) called for electronic health records for all Americans by 2014. Latest figures estimate the adoption of EHR in the outpatient environment to 13% for the base system and 4% only for a fully functional EHR system. It also includes penalties for providers who fail to adopt. Morton et al (2009)

Program to introduce an electronic medical record that enables the sharing of health information between sites is being conducted in many developed countries including Australia, Canada, Denmark, England, Finland, France, New Zealand and the United States. The information is uploaded records the patient's identity, CHI number (unique patient identifier in Scotland), and the prescription drug reactions or allergies. All patients have the information that is uploaded to a central database 2 times a day unless they have actively opted out of the system. Health professionals who want to access information in the ECS is expected to obtain consent of the patient at the time of contact (unless the patient is unconscious). Johnstone and McCartney (2010)

Health Level Seven (HL7) is a standard for electronic exchange of patient medical record information is supported by the National Committee on Vital and Health Statistics (NCVHS). HL7 standards developed by bodies accredited by the American National Standards Institute. This is a message standard that allows software applications to exchange information across platforms in a way to protect the meaning of the information submitted. Gudea (2005)

II. Medical Record

In the explanation of Article 46 paragraph (1) Medical Practice Act, which is a medical record is a file containing records and documents about the identity of thepatient, examination, treatment, action, and other services provided to patients. In the Minister of Health of the Medical Record Number 749a / Menkes / Per / XII / 1989 explained that the medical record is a file containing records and documents about the identity of the patient, examination, treatment, action, and other services to patients in healthcare facilities.

III. Electronic Medical Record

Electronic Medical Record (EMR) is the lifetime of the patient medical records in electronic format and can be accessed by computer from a network with the main purpose to provide or improve care and health services in an efficient and integrated. EMR become a key strategy of integrated health services at various hospitals. Prihartono (2008)

IV. Benefits Of Medical Records

a. Treatment of Patients. Medical records serve as the basis and guidance to plan and analyze the disease and plan treatment, care and treatment to be given to the patient.
b. Improving the Quality of Service. Creating Medical Record for the organization of medical practice with a clear and complete information will improve the quality of care to protect medical personnel and the achievement of optimal health.
c. Education and Research. Medical record is the chronological progression of disease information, medical services, treatments and medical procedures, useful for the development of information materials for teaching and research in the field of medical and dental professions.
d. Financing. Medical record file can be used as guidance and materials to establish the financing of health services at health facilities. Notes can be used as proof of financing to the patient.
e. Health Statistics. Medical records can be used as health statistics, especially for studying the development of public health and to determine the number of patients on specific diseases.
f. Problems of Proof Law, Discipline and Ethics. Medical record is the main written evidence, making it useful in the resolution of legal issues, discipline and ethics.

V. Study References

According Hosizah explained that implementation of the hospital Medical Records Indonesia started in 1989 in line with the Regulation of Minister of Health Affairs Medical Record, which includes the setting is still paper-based medical records (conventional). Conventional medical record is considered no longer appropriate for use in the 21st century the use of information-intensive and environment-oriented automation and health care is not solely focused on the work unit. Currently in Indonesia there were approximately 1300 hospitals and thousands of health centers (Menkes RI) that the government would need to think about the design of the parent (grand design) EHR strategically arranged by region includes eastern Indonesia, Central and West.

According Prihartono (2008) From Clarke writing a case study that the effects of EMR technology implementation is often unpredictable and can only be determined by using it. To be successful applied technology such as EMR, theory and practice must be balanced. So we need a test EMR. Janz and Brian Hennington test the adoption of EMR by physicians with the model Unified Theory of Acceptance and Use of Technology (UTAUT). By doing a literature study of the implementation of the UTAUT Model year 2000 - 2007, Hennington concluded that the factors influencing the decision-making EMR adoption: uncertainty on investment turnover EMR, EMR integration with existing business processes before, the potential of EMR to improve quality of care, convenience of use EMR, the amount of effort to change the workflow to fit the use of EMR, as well as funding availability and duration of adaptation. EMR adoption by a hospital, is also influenced by the local law of privacy and the use of EMR by the trend of other hospitals. Miller and Tucker empirically using a variation in the privacy of local law. The result indicates a positive network effect in the spread of EMR. There are five variables used to help predict the decision for pushing hospitals to adopt EMR, namely:

a. InstalledHSA: number of hospitals are adopting EMR in a year
b. HospPrivLaw: Indicators of privacy laws
c. HospPrivLaw * InstalledHSA: results of a variable time before
d. Xit: hospital characteristics and state
e. Eit: Error stochastic

All of the data to a formula derived from the data base of 2005 issued by the "Healthcare Information and Management Systems Society. HIMSS database covers most of the hospitals in the United States. The author gets the data as much as 4010 the hospital. 1937 hospitals have adopted EMR. 3988 the hospital's decision to adopt a system of "enterprise-wide EMR".

Two researchers from Malaysia, Mohd and Mohamad, forming a model of acceptance of EMR in the form of the questionnaire survey, particularly for the major hospitals in Malaysia. The core of their model is the incorporation of the Technology Acceptance Model (TAM) with User Interface Interaction Satisfaction Questionnaire (Marquis). By combining the TAM and the Marquis, as well as several other models of theories, then obtained a receipt of the appropriate models to evaluate the EMR.

According Handoyo et al (2008), are presented in this paper that in order to build theHospital Information System is the most efficient use of the Prado due Framework

According Arianto in his paper entitled Open Platform-Based Applications Programming Xml Web Services (Case Study: Collaboration Applications and Data Exchange of the Population With Medical Records). It is said that the population data can be used to mensuport patient data at a hospital. In the presented architecture is one solution for distributed computing applications in a cluster collaboration and data exchange. Architecture described in his presentation.

According Setyanto in his paper entitled Mobile Medical Records said that the medical records of mobile applications will make double the storage of medical records. Medical records will be stored on the server where the hospital treated patients and in patients of coffee in the mobile device. Double the storage is done so that when the required medical records of patients at other hospitals where new patients never treated, the new hospital can retrieve data from the patient's pertinent medical records without the need to deal with patients from the hospital. The addition of medical records at the time of treatment of transactions to be recorded back into the patient's mobile device. In this way the medical records contained in the mobile device will be the most complete data. To enable the synchronization with the data at the beginning of the hospital if the patient wants a new backup data as well as his medical records to synchronize the applications that are embedded in the mobile device will synchronize with the hospital system from which he recorded via web services service owned hospitals. This feature can be designed automatically when a signal is a data communications network.

Doctors will be helped in doing the best treatment decisions for patients. Patients will benefit from the certainty of the data unreadable medical history and basis for treatment decisions for themselves. Government and health researchers will get an abundance of data is ready if that research could be done more easily and the data is more complete. Strategic decisions taken by the government can also better because of the completeness of data. Mobile data communication service providers also benefit from the increased traffic that is not only extended service mobile banking, but possess the new land mobile medic.

According to Boonstra and Broekhuis (2010) explained that the implementation of Electronic Medical Records, Financial constraints become a major factor. Monetary aspect is an important factor for many physicians. Common questions faced by clinicians is whether the costs of implementing and running an affordable EMR system and whether they can benefit financially from it. The cost of EMR can be divided into two, namely the initial cost and ongoing costs, monitoring, upgrades, and administration costs. Their surveys, have concluded that the physician has adequate technical knowledge and skills to deal with EMRs. Meade et al observed in a context that most of the current generation of doctors in Ireland to receive their qualifications before the IT program was introduced.

According to Desroches et al in his paper entitled Use of Electronic Health Records in U.S. Hospitals. Conduct a survey using the methods surveyed all acute care hospitals that are members of the American Hospital Association to the presence of the specific functions of electronic records. Using the definition of electronic health records based on expert consensus, to determine the proportion of hospitals that have such systems in their clinical areas. Also examined the relationship of adoption of electronic health records with certain hospital characteristics and the factors that are reported to be barriers or facilitators of adoption.

From a survey based on responses from 63.1% of surveyed hospitals, only 1.5% of U.S. hospitals have comprehensive electronic records system that is in all clinical units and an additional 7.6% have basic systems in which at least one clinical unit . Computerized provider order for the drug has been applied to 17% of the hospital. Larger hospitals located in urban and teaching hospitals were more likely to have an electronic records system. Respondents cited capital requirements and high maintenance costs as the main barriers to implementation, although hospitals with electronic records systems were less likely to mention the constraints of a hospital without such a system. AK Jha concluded his research is a very low level of adoption of electronic health records in U.S. hospitals suggest that policymakers face major obstacles to the achievement of healthcare performance goals that depend on health information technology.

Raisinghani and Young put his research, entitled Personal health records: key adoption issues and implications for management. Presented that Electronic Personal Health (PHRs) have been considered as a tool to empower consumers to become active decision makers about their health, instead of leaving the decision to the provider. Paper-based health systems and fragmented is no longer suitable for the digital economy in the 21st century. Integrated health information technology system is the solution to change clinical practice to consumer centric and information. Tools such as PHRs are a means to achieve goals that provide better health, safer and more affordable for consumers. However, there has been little research done to show the real value of PHR, although widely perceived value of this technology. Although survey data indicate that there is a lack of awareness among the public, consumers accept this concept, especially when the doctor recommends it.

Zaroukian and Sierra in his paper benefiting from ambulatory EHR implementation: solidarity, six sigma, and willingness to strive. Explained that the system of electronic ambulatory health records has the potential to improve the quality of healthcare. Optimizing the value of EHR implementation requires that providers and staff to be effective and efficient EHR users so that the graph paper is no longer needed or desired. Transition from paper charts to EHR systems require changes in new learning. This case study describes how the EHR implementation of timely and routine use in a large medical clinic. Observed benefits include improved patient access, workflow efficiency, communication, use of decision support, and financial performance. These success factors and implementation strategies can help others trying to encourage greater adoption and use of EHRs.

Balfour et al in his paper entitled Health Information Technology presented the United States have been slow to use HIT. However, a variety of factors including increased government involvement, which accelerate the implementation and use of HIT. E-prescribing and EHR both electronic means to provide better coordination of care by allowing the various health care professionals to access patient medical records. Adoption of e-prescribing can reduce medication errors due to bad handwriting. Unfortunately, barriers to implementing e-prescribing and EHR is still there, including resistance to learning new technologies, the initial start-up costs, delays in seeing a return on investment, lack of standard platforms, increasing the administrative burden and incentive alignment.

Shekelle et al in the paper Costs and benefits of health information technology to take the source data from PubMed, Cochrane Controlled Clinical Trials Register and Cochrane Database purpose Effectiveness Reviews (DARE) is an electronic search for articles published since 1995. Some of the reports prepared by private industry were also reviewed. Using the method of the 855 studies screened, 256 were included in the final analysis.

The results of 256 studies, 156 concerned decision support, electronic medical records of 84 and 30 about to be computerized physician entry (categories are not mutually exclusive). 124 of the studies assessing the effects of HIT systems in outpatient settings or outpatient; 82 assessed its use in the hospital setting or hospitalization. The ability of Electronic Health Records (EHRs) to improve the quality of care in ambulatory care facilities is shown in a small series of studies conducted at four sites (three U.S. medical centers and one in the Netherlands). HIT has the potential to enable a dramatic transformation in the delivery of health care, making it safer, more effective and more efficient.

Gagnon et al, Interventions for promoting information and communication technologies adoption in healthcare professionals. Exposure produces 10 studies met the inclusion criteria.. Use of the Internet for audit and feedback, and email to the provider-patient communication, were targeted in two studies. Their conclusion is very limited evidence on effective interventions to promote the adoption of ICT by health professionals. Small effects have been reported for interventions targeting the use of electronic databases and digital libraries. Effectiveness of interventions to promote the adoption of ICT in healthcare settings is still uncertain, and further trials are designed with both needed.

Morton et al (2009) concluded EHR has been developed over nearly three decades, yet some providers to realize an integrated electronic health records. Recent figures estimate 3-8% of EHR adoption in ambulatory care settings to 13% for the base system and 4% only for a fully functional EHR system. Patients seemed to accept that their physicians have computerized records secure and confidential, but they are increasingly unhappy with the safety record is held centrally. The concern is that the data may be used inappropriately by the government or may be hijacked. The risk increases with the number of illegal access data stored in a single repository (honeypot effect).

Pusic et al (2004) presented to exploit the opportunities for this type of clinical decision support interventions then it must have an effective health information systems. While electronic health records and databases to help physicians manage information, patient-specific recommendations provided by the clinical decision support systems can do more to improve decision making and help ensure patient safety. Computer technology can help to generate suggestions for specific cases of clinical decision making. The system used is usually referred to as clinical decision support systems. computerization can help make this valuable investment with a safer, more efficient and more effective health care. It is imperative that physicians involved in the development and rigorous scientific evaluation of this system.

Skouroliakou et al (2008) describes use of computerized hospital records could potentially reduce medical errors and improve the cost effectiveness of care by revealing the relationship between the severity of illness and resource consumption in the ICU setting. The importance of computerized data management to improve the safety and efficacy in the ICU for premature neonates has been fully realized for several decades. Availability of this information is enabling physicians to minimize the errors and re-evaluate current clinical practice.

Abdrbo et al (2011) described information systems can facilitate communication between nurses, doctors and other health team members and improve patient outcomes. In addition, the use of IS will ensure the completeness of documentation of patient care, facilitating the evaluation of the results of patient care and improve patient safety.

Parmanto et al (2005) presented to achieve national interoperability and realizing benefits, physician adoption rate should be increased substantially. However, implementing the right systems the right way is essential to ensure the success of the project and protect patient safety. Nearly 75% of all the major health information technology projects fail. Understanding of the factors associated with the acceptance of physicians' will enable organizations to better assess the readiness of the system and facilitate a successful implementation.

Johnstone and McCartney (2010) presented patients seem to accept that their physicians have a safe and confidential computerized records, but they are increasingly unhappy with the holding of centralized security record. The concern is that the data may be used inappropriately by the government or may be hijacked. The risk increases with the number of illegal access data stored in a single repository (honeypot effect).

Skouroliakou et al in an article entitled Data Analysis of the Benefits of an Electronic Registry of Information in a Neonatal Intensive Care Unit in Greece explained that the electronic documentation of several procedures for neonates, such as parenteral nutrition in the ICU, has been referred to in the literature. Establishment of monitoring systems allow for the research results as well as for the management of information. Availability of this information is enabling physicians to minimize the errors and re-evaluate current clinical practice. Over the last two years of a software program that combines rapid report generation and capacity for simple statistical analysis was developed and used for collecting, storing, and analyzing the data of newborns treated in intensive care unit three levels of Lito Maternity Hospital.

Et al in his paper Abdrbo Development and Testing of a Survey Instrument to Measure Benefits of a Nursing Information System, Health Information Management concluded benefits associated with quality of care using the System Information is associated with improved accessibility, accuracy and completeness of patient information that increases the effectiveness of nursing care . Nauright and Simpson12 reported high reliability (Cronbach alpha = .94) for quality-of-care items included in the questionnaire they used in their study of 697 nurses and staff of public hospitals.

Survey conducted by Hussein et al (2011) of 80 patients as responders were randomized at Bandung in January-May 2011. About 70% of respondents (56 people) were aged 18-50 years, while 30% of respondents (24 people) were aged or under 9-17 years following table Generate

No	Question	Y	N
1	Understand the medical record in general	90%	10%
2	Knowing the long-term goal of the medical record	70%	30%
3	Knowing the content rather than medical record	80%	20%
4	Registered in more than one hospital in Bandung	99%	1%
5	Percent came to the medical is about 1 times / month or more	80%	20%
6	Been admitted to hospital	60%	40%
7	Ever have one of the actions by the medical officer	50%	50%

8	Given 60% more on medical services who had obtained	20%	80%
9	Given 60% more on drugs that have consumed the prescription / doctor action	10%	90%
10	Given 60% more hospital ever provide medical services	80%	20%

Based on the above table it can be concluded that patients rarely remember who had obtained medical services from childhood to the present, patients given the drug also rarely ever consumed from childhood to the present, when the average to obtain medical services once a month and is registered in more from one hospital. Thus the need for a complete documentation of the patient's health, correct and up to date that can provide health information to patients and medical personnel in times of need. Also needed the flexibility to read and access information by the patient's health is concerned. A total of 40 respondents had experienced any of the actions by medical personnel. The interviews explained that the medical officers often make mistakes because the action does not 'know' the patient at hand, that does not hold a valid and complete information regarding the patient's personal health information, not even knowing the patient's medical history at hand. The result is a medical tort (malpractice) so that patients suffering from trauma, fainting, itching rash, should be treated intensively, even resulting in loss of life.

Development of the National Medical Record System Web-based, so that medical record information can be accessed freely by the patient and medical personnel wherever the patient is or requires medical treatment. Web technologies into one technology that is widely used today because of the ability to meet the needs of web users are mobile .. Due to the web-based system, then both patients and medical personnel have access rights to different medical record. For example, the patient has a right to see medical records and medical history, while the doctor has the right to record into the patient's medical record is examined. The hospital is the party who first made the patient medical record (if the patient has not had a medical record), after the patient's medical record form the physician and / or other hospital staff can record the medical treatment and see pasian medical record, if needed. National web-based medical record can be accessed by the patient's mobile, so that patients can learn about personal health, knowing his medical history, and the patient can decide to choose a doctor or hospital that is suitable based on medical records. Know yourself is important, so are familiar with our own personal health, hospital and medical staff of data.

Implementation and Maintenance of Web-based Medical Record System to build Portal Center is administered by the Department of Health, Handling security using username and password plated, Maintenance of data by backing up your data based on a specific time period.

Every user has different levels of access rights to maintain the validity of the data. Normal development of databases that can make a good implementation of the system. Server support 24 hours to make the availability of patient health information for patients and physicians. Simple interface but achieve the delivery of information.

VI. DISCUSSION

Physician practice conditions, hospital treatment and Hall are still many who use handwritten notes on paper to write medical records. That is not good, but in a previous study found many weaknesses. Various studies have previously shown that a lot of paper medical records led to the wrong perception because the article misspelled the difficult doctor read by other paramedics. Because medical records are poor or non-existent making it difficult to trace the history of the patient before the doctor who caused the worng diagnoiswill allow the caused the wrong diagnosis that will allow the mall practice. Electronic Medical Record technology allows medical providers to storeand mempertukaran medical information using a computer instead of paper. Different metadata systems in hospitals and other clinics should be made uniform by the authorities to enable the synchronization of data and can be utilized to the maximum.

Web portal development database of medical records are appropriately built in Indonesia which was shaded by the relevant departments of the ministries of health should require hospitals to build a database that is connected to the portal. Access password is not enough to use, it must use additional authentication, can be a barcode or fingerprint. Thus there is no one else can enter into the system unless the patient is concerned even if the user id and password not remembered.

VII. CONCLUSION

a. Indonesia has not been widely applied in electronic medical records so that there are many drawbacks

b. Medical Record Information system is built in standard among hospitals and clinics will provide benefits of convenience and carrying medical resume so that patients can be misdiagnosed and press the number mall practice

VIII. ADVICE

Leadership of hospitals and clinics need to think logically and humbled to be willing to apply the patient's medical record system

REFFERENCE

[1] Amalia R. Miller_ and Catherine E. Tucker, (2007), Privacy Protection and Technology Diffusion: The Case of Electronic Medical Records, MIT Sloan School of Business, MIT, Cambridge

[2] Gagnon MP, Légaré F, Labrecque M, Frémont P, Pluye P, Gagnon J, Car J, Pagliari C, Desmartis M, Turcot L, Gravel K (2009) Interventions for promoting information and communication technologies adoption in healthcare professionals. Faculty of Nursing, Université Laval - Centre hospitalier universitaire de Q Canada

[3] Shekelle PG, Morton SC, Keeler EB. (2006), Costs and benefits of health information technology, Evid Rep Technol Assess

[4] Balfour DC 3rd, Evans S, Januska J, Lee HY, Lewis SJ, Nolan SR, Noga M, Stemple C, Thapar K, (2009), Health information technology, Sharp Rees-Stealy Medical Group, San Diego, California, J Manag Care Pharm.

[5] Zaroukian MH, Sierra A. (2006), Benefiting from ambulatory EHR implementation: solidarity, six sigma, and willingness to strive. J Healthc Inf Manag,.20(1):53-60.

[6] Raisinghani MS, Young E. Personal health records: key adoption issues and implications for management. School of Management, P.O. Box 425738, CFO 405, Denton, TX 76204-5738, USA.

[7] Jha AK, Desroches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, Shields A, Rosenbaum S, Blumenthal D, (2009), Use of Electronic Health Records in U.S. Hospitals. Citations selected from a literature search of the following database sources: PubMed, Web of Science, and CINHAL

[8] Albert Boonstra, Manda Broekhuis, (2010), Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions, Boonstra and Broekhuis BMC Health Services Research

[9] K.M. Clarke, M. Rouncefield, M.J. Hartswood, R.N. Procter. (2001), The Electronic Medical Record And Everyday Medical Work. Health Informatics Journal, Vol. 7, No. 3-4, p. 168-170, SAGE

[10] Iwan Prihartono, (2008), Adopsi Rekam Medis Elektronik, Fakultas Ilmu Komputer, Universitas Indonesia,

[11] Edwin Yaqub, MSc, and Andre Barroso, PhD, (2010), Distributed Guidelines (DiG): A Software Framework for Extending Automated Health Decision Support to the General Population_2, Health Information Management, 1-15

[12] By Dean F Sittig, PhD; Brian L Hazlehurst, PhD; Ted Palen, MD, PhD; John Hsu, MD; Holly Jimison, PhD; Mark C Hornbrook, PhD, (2002), A Clinical Information System Research Landscape, The Permanente Journal, Volume 6 No. 2

[13] Gregory L. Alexander, PhD, RN, (2008), A Descriptive Analysis of a Nursing Home Clinical Information System with Decision Support, Perspectives in Health Information Management 5; 12

[14] Bambang Parmanto PhD, Matthew Scotch, and Sjarif Ahmad, (2005), A Framework for Designing a Healthcare Outcome Data Warehouse, the American Health Information Management Association

[15] Mary E. Morton, PhD, rhia, dan Susan Wiedenbeck, PhD, (2009), A Framework for Predicting EHR Adoption Attitudes: A Physician Survey, Health Information Management 6

[16] Chris Johnstone, MBChB, and Gerry McCartney, MBChB, (2010), A Patient Survey Assessing the Awareness and Acceptability of the Emergency Care Summary and its Consent Model, Health Information Management

[17] Ronaldo Parente, PhD, Ned Kock, PhD, and John Sonsini, (2004), An Analysis of the Implementation and Impact of Speech-Recognition Technology in the Healthcare Sector, Health Information Management 1;5

[18] Martin Pusic, MD, Dr J. Mark Ansermino, FFA, MMed, MSc, FRCPC, (2004), Clinical decision support systems, BCMJ, Vol. 46, No. 5

[19] Robert J. Campbell, EdD, (2005), Consumer Informatics: Elderly Persons and the Internet, Health Information Management

[20] Maria Skouroliakou, PhD; George Soloupis, MSc; Antonis Gounaris,MD; Antonia Charitou, MD; Petros apasarantopoulos; Sophia L. arkantonis, PhD; Christina Golna, LLM, MSc; and Kyriakos Souliotis, PhD, (2008), Data Analysis of the Benefits of an Electronic Registry of Information in a Neonatal Intensive Care Unit in Greece, Health

[21] Sorin Gudea, (2005), Data, Information, Knowledge: A Healthcare Enterprise Case Study, Health Information Management

[22] Amany A. Abdrbo, BSN, MSN, PhD; Jaclene A. Zauszniewski, PhD, RN-BC, FAAN; Christine A. Hudak, RN, PhD, CPHIMS; and Mary K. (2011), Anthony, PhD, RN, CS, Development and Testing of a Survey Instrument to Measure Benefits of a Nursing Information System, Health Information Management 1-15.

[23] Hatmoko Tri Arianto, Pemprograman Aplikasi Platform Terbuka Berbasis Xml Web (Studi Kasus: Kolaborasi Aplikasi Dan Pertukaran Data Kependudukan Dengan Catatan Medis), Fakultas Teknik, Universitas Diponegoro

[24] Hosizah, Electronic Health Record (Ehr) Atau Rekam Kesehatan Elektronik: Change In The Him Department,

[25] Eko Handoyo, Agung Budi Prasetijo, Fuad Noor Syamhariyanto, (2008), Aplikasi Sistem Informasi Rumah Sakit Berbasis Web Pada Sub-Sistem Farmasi Menggunakan Framework Prado, Universitas Diponegoro, Semarang

[26] Inne Gartina Husein , Anwar Hasdian Lubis (2011), Usulan Pengembangam Kebutuhan Medicls Record Berbasis Teknologi Web, Prosiding Conferensi Nasional ICTM Politeknik Telkom Bandun.

A computational linguistic approach to natural language processing with applications to garden path sentences analysis

DU Jia-li
School of Foreign Languages/School of Literature
Ludong University/Communication University of China
Yantai/ Beijing, China

YU Ping-fang
School of Liberal Arts/Institute of Linguistics
Ludong University/Chinese Academy of Social Sciences
Yantai/ Beijing, China

***Abstract*— This paper discusses the computational parsing of GP sentences. By an approach of combining computational linguistic methods, e.g. CFG, ATN and BNF, we analyze the various syntactic structures of pre-grammatical, common, ambiguous and GP sentences. The evidence shows both ambiguous and GP sentences have lexical or syntactic crossings. Any choice of the crossing in ambiguous sentences can bring a full-parsed structure. In GP sentences, the probability-based choice is the cognitive prototype of parsing. Once the part-parsed priority structure is replaced by the full-parsed structure of low probability, the distinctive feature of backtracking appears. The computational analysis supports Pritchett's idea on processing breakdown of GP sentences.**

***Keywords*- Natural language processing; computational linguistics; context free grammar; Backus–Naur Form; garden path sentences.**

I. INTRODUCTION

The advent of the World Wide Web has greatly increased demand for natural language processing (NLP). NLP relates to human-computer interaction, discusses linguistic coverage issues, and explores the development of natural language widgets and their integration into multi user interfaces[1].The development of language technology has been facilitated by two technical breakthroughs: the first emphasizes empirical approaches and the second highlights networked machines [2].Natural language and databases are core components of information systems, and NLP techniques may substantially enhance most phases of query processing, natural language understanding and the information system [3-5].

By means of developed or used methods, metrics and measures, NLP has accelerated scientific advancement in human language such as machine translation[6-7], automated extraction systems from free-texts[8], the semantics-originated Generalized Upper Model of a linguistic ontology [9], artificial grammar learning (AGL) system[10], NIMFA[11], etc. Understanding natural language involves context-sensitive discrimination among word senses, and a growing awareness is created to develop an indexed domain-independent knowledge base that contains linguistic knowledge [12-17].

There are a lot of helpful NLP models for linguistic research focusing on various application areas, e.g. Zhou & Hripcsak' medical NLP model and Plant& Murrell's dialogue system.

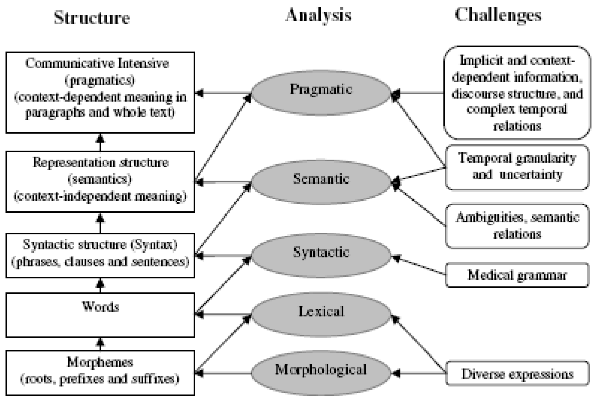

Structure
Analysis
Challenges
Communicative Intensive (pragmatics) (context-dependent meaning in paragraphs and whole text)
Representation structure (semantics) (context-independent meaning)
Syntactic structure (Syntax) (phrases, clauses and sentences)
Words
Morphemes (roots, prefixes and suffixes)
Pragmatic
Semantic
Syntactic
Lexical
Morphological
Implicit and context-dependent information, discourse structure, and complex temporal relations
Temporal granularity and uncertainty
Ambiguities, semantic relations
Medical grammar
Diverse expressions

Figure 1 Zhou & Hripcsak' Medical NLP Model

Zhou & Hripcsak' medical NLP model comprises three parts, i.e. "structure", "analysis" and "challenges". "Analysis" consists in morphological, lexical, syntactic, semantic and pragmatic parts. Morphology and lexical analysis determine the sequences of morphemes used to create words. Syntax emphasizes the structure of phrases and sentences to combine multiple words.

Semantics highlights the formation of the meaning or interpretation of the words. Pragmatics concerns the situation of how context affects the interpretation of the sentences and of how sentences combine to form discourse. [18]

Plant& Murrell's Dialogue NLP System discusses the importance of Backus–Naur Form (BNF). This system analyzes the possibility for any user who understands formal grammars to replace or upgrade the system or to produce all possible parses of the input query without requiring any programming.

In the model, BNF is extended with simple semantic tags. The matching agent searches through a knowledge base of scripts and selects the most closely matching one. In this model, BNF is very helpful and useful for system to analyze natural language. [19]

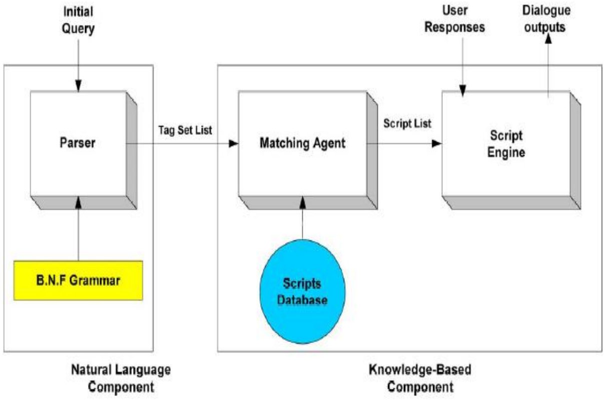


Figure 2 Plant& Murrell’s Dialogue NLP System

The computational analysis of Garden Path (GP) sentences is one of the important branches of NLP for these sentences are hard for machine to translate if there is no linguistic knowledge to support.

GP sentences are grammatically correct and its interpretation consists of two procedures: the prototype understanding and the backtracking parsing. At the first time, readers most likely interpret GP sentences incorrectly by means of cognitive prototype. With the advancement of understanding, readers are lured into a parse that turns out to be a dead end. With the help of special word or phrase, they find that the syntactic structure which is being built up is different from the structure which has been created, namely it is a wrong path down which they have been led. Thus they have to return and reinterpret, which is called backtracking. "Garden path" here means "to be led down the garden path", meaning "to be misled". Originally, this phenomenon is analyzed by the psycholinguists to illustrate the fact that human beings process language one word at a time when reading. Now, GP phenomenon attracts a lot of interest of scholars from perspectives of syntax[20-24], semantics[25-28], pragmatics[29-30], psychology[31-34], computer and cognitive science[35-38].

In this paper, Context Free Grammar (CFG) and BNF will be used to discuss the automatic parsing of GP sentences. Meanwhile, the pre-grammatical sentences, common sentences and ambiguous sentences will be analyzed from the perspective of computational linguistics as the comparison and contrast to GP sentences.

II. The NLP-Based Analyses Of Non-Gp Sentences

Non-GP sentences in this paper include the pre-grammatical sentences, common sentences and ambiguous sentences, all of which are shown how different they are from GP sentences.

A. *Analysis of Pre-Grammatical Sentences*

A pre-grammatical sentence is incorrect in grammar even though we can guess the meaning by the separated words or phrases. According to CFG, this kind of sentence fails to be parsed successfully.

Example 1: *The new singers the song.

G={Vn, Vt, S, P}

Vn={Det, Adj, N, NP, S, VP, V}

Vt={the, new, singers, song}

S=S

P:

1. S→NP VP
2. NP→Det N
3. NP→Det Adj N
4. VP→V NP
5. Det→{the}
6. N→{singers, song}
7. Adj→{new}
8. V→{?}

The new singers the song

Det new singers the song (5)

a. Det Adj singers the song (7)
b. Det Adj N the song (6)
c. NP the song (3)
d. NP Det song (5)
e. NP Det N (6)
f. NP NP (2)
g. FAIL

From the parsing of Example 1, we can see the whole structure of sentence is [The new singers]NP+[the song]NP，namely the absence of V is the reason why it fails to be parsed successfully.

In a pre-grammatical sentence, the syntactic structure is not correct and the relationships among the parts are isolated even though sometimes the possible meaning of the sentence can be inferred from the evidence. For example, in the programming rules of (8), we can enter a lot of related verbs to rewrite example 1, e.g. V→{hear/play/write/sing/ record}. Thus the pre-grammatical sentence can be created into a common one.

B. *Analysis of Common Sentences*

A common sentence is grammatically acceptable and both CFG and BNF can parse it smoothly and successfully. If “record(verb)” is added into example 1, the formed sentence is a common one.

Example 2：The new singers record the song.

G={Vn, Vt, S, P}

Vn={Det, Adj, N, NP, V, VP, S}

Vt={the, new, singers, record, song}

S=S

P:

1. S→NP VP
2. NP→Det N
3. NP→Det Adj N
4. VP→V NP
5. Det→{the}
6. N→{singers, song}
7. Adj→{new}
8. V→{record}

The new singers record the song

a. Det new singers record the song (5)

b. Det Adj singers record the song (7)

c. Det Adj N record the song (6)

d. NP record the song (3)

e. NP V the song (8)

f. NP V Det song (5)

g. NP V Det N (6)

h. NP V NP (2)

i. NP VP (4)

j. S (1)

k. SUCCESS

The syntactic structure of example 2 is [The new singers] NP+[record]V+[the song]NP, and the whole parsing is smooth.

Backus-Naur Form (BNF) is another useful formal language to describe the parsing of NLP. The details of BNF definition are as follows.

```
syntax ::=
    rule ::= identifier "::=" expression
    expression ::= term { "|" term }
    term ::= factor
    factor ::= identifier |
    quoted_symbol |
    "(" expression ")" |
    "[" expression "]" |
    "{" expression "}"
    identifier ::= letter { letter | digit }
quoted_symbol ::= """" """"
```

Thus we can use BNF to define Augmented Transition Network (ATN) which will be introduced to analyze the related sentences in this paper.

```
<ATN>::=<State Arc>{<State Arc>}
<State Arc>::=<State><Arc>{<Arc>}
<Arc>::=CAT<Category><Preaction>
    |PUSH<State>< Preaction >
    |TST<Node>< Preaction >
    |POP<Expression><Test>
<Preaction>::=<Test>{<Action>}<Terminal Action>
<Action>::=SETR<Register><Expression>
    | SENDR<Register><Expression>
    | LIFTR<Register><Expression>
<Terminal Action>::= TO<State>[<Form>]
    | JUMP<State>[<Form>]
<Expression>::=GETR<Register>|*
    | GETF<Feature>
    |APPEND<Register><Expression>
    |BUILD<Fragment>{<Register>}
```

In the semantic network, some nodes are associated with lexicon entries. In order to analyze example 2 clearly and concisely, we find a detailed description of lexicon is necessary besides the grammatical analysis. “CTGY” means category; “PRES”, present; “NUM”, number; “SING”, singular.

(The((CTGY. DET)))

(New((CTGY. ADJ)))

(Singers((CTGY.N) (NUM. PLURAL)))

(Record((CTGY.V)(PAST.RECORDED)(PASTP. RECORDED)))

(Record((CTGY. V) (TENSE.PRES))

(Song((CTGY. N) (NUM. SING)))

Based on the evidence discussed above, we can create an augmented transition network to analyze example 2.

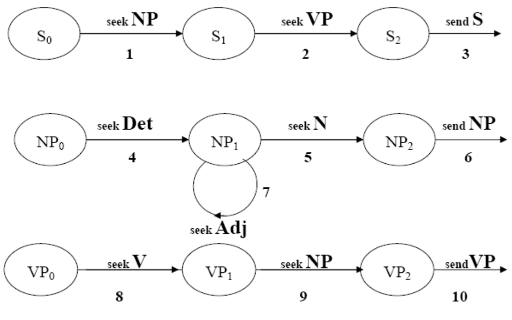


Figure 3 ATN of Example 2

The ATN in Fig. 3 shows the details of parsing of example 2, which belongs to the category of common sentence. There is no backtracking or ambiguity existing in the procedure shown below.

1. System tries to seek NP in arc 1 and then PUSH NP <The new singers> to NP subnet;

2. NP subnet begins to parse NP <The new singers>. In arc 4, Det <the> is set in register.

3. In arc 7, Adj <new> is analyzed and the result is set in register.

4. In arc 5, N<singers> is interpreted.

5. In arc 6, the result of parsing in NP subnet is popped to general net in arc 1.

6. Again in arc 1, the popped result is set in register.

7. In arc 2, system starts to seek VP<record the song> and PUSH to VP subnet.

8. VP subnet begins to parse VP<record the song>. In arc 8, V<record> is set in register.

9. In arc 9, VP subnet begins to interpret NP <the song>. There is no related rule to support the procedure in this VP subnet and as a result, the sub-sub-net of NP is activated again. NP <the song> is pushed to NP subnet.

10. NP sub-sub-net begins to parse NP<the song>. In arc 4, Det <the> is set in register again.

11. In arc 5, N <song> is parsed.

12. In arc 6, the result of parsing in NP sub-sub-net is popped to VP subnet.

13. In arc 9, NP<the song> is set in register.

14. In arc 10, VP<record the song> is popped.

15. In arc 2, the parsing result of VP subnet is set.

All the parsing results of subnets and sub-subnets show that S<the new singers record the song> is grammatically and semantically acceptable and reasonable. The information is set in register. System returns "SUCCESS" and parsing is over.

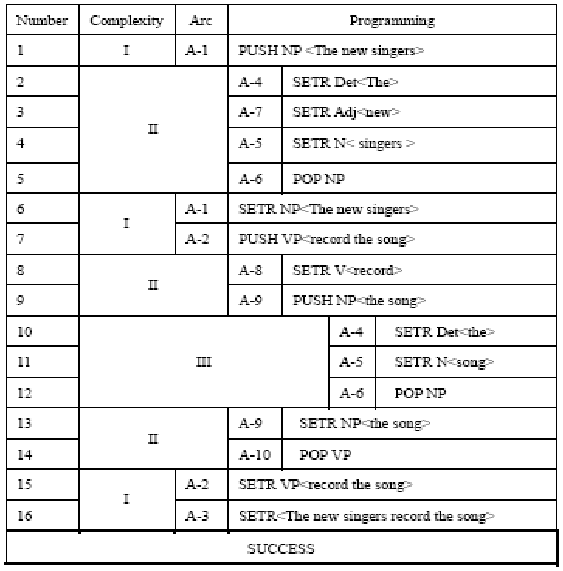
The algorithm of parsing discussed above can be found in Table 1, in which "Number" means the steps of parsing; "Complexity", the hierarchical levels of net; "Arc" or "A-?", the respective numbers shown in Fig. 3; "Programming", the BNF description.

C. Analysis of Ambiguous Sentences

An ambiguous sentence has more than one possible meaning, any of which can convey and carry the similar, different and even opposite information.

Example 3 : The detective hit the criminal with an umbrella.

The example above brings syntactic ambiguity for the different syntactic structures convey different meanings.



Number	Complexity	Arc	Programming
1	I	A-1	PUSH NP <The new singers>
2	II	A-4	SETR Det<The>
3		A-7	SETR Adj<new>
4		A-5	SETR N< singers >
5		A-6	POP NP
6	I	A-1	SETR NP<The new singers>
7		A-2	PUSH VP<record the song>
8	II	A-8	SETR V<record>
9		A-9	PUSH NP<the song>
10	III	A-4	SETR Det<the>
11		A-5	SETR N<song>
12		A-6	POP NP
13	II	A-9	SETR NP<the song>
14		A-10	POP VP
15	I	A-2	SETR VP<record the song>
16		A-3	SETR<The new singers record the song>
SUCCESS			

Table 1 Parsing Algorithm of Example 2

In example 3, two meanings are carried. The first is the detective using an umbrella hit the criminal, while the other is the detective hit the criminal who is carrying an umbrella.

G={Vn, Vt, S, P}

Vn={Det, N, NP, V, VP, S, Prep, PP}

Vt={the, detective, hit, criminal, with, an, umbrella}

S=S

P:

1. S→NP VP
2. NP→NP PP
3. NP→Det N
4. PP→Prep NP
5. VP→VP PP
6. VP→V NP
7. PP→Prep NP
8. Det→{the, an}
9. N→{detective, criminal, umbrella}
10. Prep→{with}
11. V→{hit}

(The((CTGY. DET)))

(Detective((CTGY.N) (NUM. SING)))

(Hit((CTGY. V) (PAST. HIT) (PASTP. HIT)))

(Hit((CTGY. V) (ROOT. HIT) (TENSE.PAST)))

(Hit((CTGY. V) (ROOT. HIT) (TENSE.PASTP)))

(Criminal ((CTGY.N) (NUM. SING)))

(With((CTGY.PREP)))

(An((CTGY. DET)))

(Umbrella((CTGY.N) (NUM. SING)))

The detective hit the criminal with an umbrella.

a. Det detective hit the criminal with an umbrella (8)

b. Det N hit the criminal with an umbrella (9)

c. NP hit the criminal with an umbrella (3)

d. NP V the criminal with an umbrella (11)

e. NP V Det criminal with an umbrella (8)

f. NP V Det N with an umbrella (9)

g. NP V NP with an umbrella (3)

h. NP VP with an umbrella (6)

i. NP VP Prep an umbrella (10)

j. NP VP Prep Det umbrella (8)

k. NP VP Prep Det N (9)

l. NP VP Prep NP (3)

m. NP VP PP (4)

n. NP VP (5)

o. S (1)

p. SUCCESS

Based on the parsing above, we can find the first exact meaning of example 3 is “The detective using an umbrella hit the criminal”. Another parsing which means “The detective hit the criminal who is carrying an umbrella” is shown as follows.

The detective hit the criminal with an umbrella

a. Det detective hit the criminal with an umbrella (8)

b. Det N hit the criminal with an umbrella (9)

c. NP hit the criminal with an umbrella (3)

d. NP V the criminal with an umbrella (11)

e. NP V Det criminal with an umbrella (8)

f. NP V Det N with an umbrella (9)

g. NP V NP with an umbrella (3)

h. NP V NP Prep an umbrella (10)

i. NP V NP Prep Det umbrella (8)

j. NP V NP Prep Det N (9)

k. NP V NP Prep NP (3)

l. NP V NP PP (4)

m. NP V NP (2)

n. NP VP (6)

o. S (1)

p. SUCCESS

In ATN created by means of example 3, three subnets are involved, i.e. NP subnet, VP subnet and PP subnet. S net is the general net. The reason why the different meanings of example 3 can be expressed lies in the attached structures of PP subnet. When PP subnet is attached to VP subnet, namely VP→VP PP is activated, the parsing result is “The detective using an umbrella hit the criminal”. When PP subnet serves NP subnet, i.e. NP→NP PP, the interpretation is “The detective hit the criminal who is carrying an umbrella”.

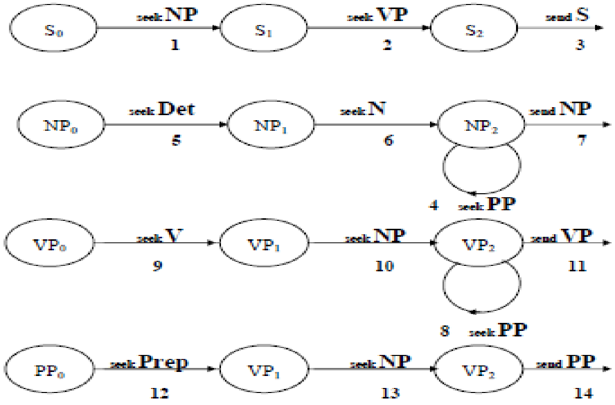


Figure 4 ATN of Example 3

From the Fig. 4, we can notice the difference of PP subnet which can be attached to NP subnet in arc 4 or to VP subnet in arc 8.

The parsing algorithm of example 3 in “VP→VP PP” includes 24 steps and highest level of syntactic structure is “IV”.

1 In arc1, S-net seeks NP<the detective>. NP subnet used to parse noun phrase is activated.

2 In arc 5, NP subnet finds Det<the>.

3 In arc 6, N<detective> is parsed and set in register.

4 In arc 7, the parsing result is popped up to arc 1 where it is pushed.

5 In arc 1, NP<the detective> is set in register.

6 In arc 2, S-net seeks VP and the other part of VP<hit the criminal with an umbrella> is pushed down to VP subnet.

7 In arc 9, VP subnet seeks V<hit> firstly.

8 In arc 10, subnet seeks NP, and NP<the criminal> is pushed again to NP subnet to interpret.

9 In arc 5, NP subnet finds Det<the>.

10 In arc 6, NP subnet seeks N<criminal>.

11 In arc 7, the result of parsing of NP<the criminal> is popped up to arc 10 where it is pushed down.

12 In arc 10, NP<the criminal> is set in register.

13 In arc 8, VP subnet seeks PP<with an umbrella> and PP subnet is activated.

14 In arc 12, PP subnet finds Prep<with>.

15 In arc 13, PP subnet tries to parse NP <an umbrella> and for the third time, NP subnet is provided for the parsing.

16 In arc 5, NP subnet searches for Det<an>.

17 In arc 6, N<umbrella> is parsed in NP subnet.

18 In arc 7, the result is popped back to arc 13.

19 In arc 13, NP <an umbrella> is set in register.

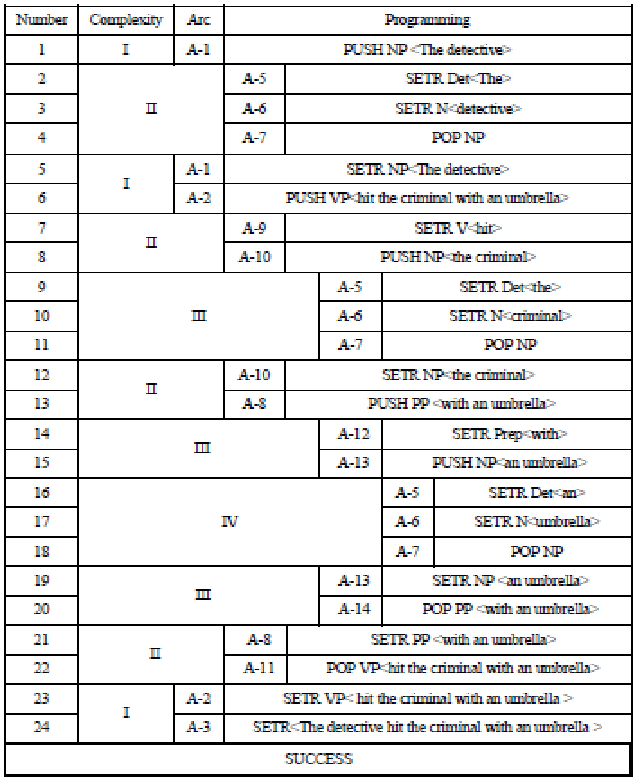
20 In arc 14, PP<with an umbrella> is parsed successfully and it is popped up to arc 8.

21 In arc 8, PP<with an umbrella> is set in register.

22 In arc 11, the parsing of VP<hit the criminal with an umbrella> is finished and system has the result popped up to arc 2.

23 In arc 2, VP<hit the criminal with an umbrella> is set in register.

24 In arc 3, S<the detective hit the criminal with an umbrella> is parsed completely. System returns "SUCCESS" and parsing is over.



Number	Complexity	Arc	Programming
1	I	A-1	PUSH NP <The detective>
2	II	A-5	SETR Det<The>
3		A-6	SETR N<detective>
4		A-7	POP NP
5	I	A-1	SETR NP<The detective>
6		A-2	PUSH VP<hit the criminal with an umbrella>
7	II	A-9	SETR V<hit>
8		A-10	PUSH NP<the criminal>
9	III	A-5	SETR Det<the>
10		A-6	SETR N<criminal>
11		A-7	POP NP
12	II	A-10	SETR NP<the criminal>
13		A-8	PUSH PP <with an umbrella>
14	III	A-12	SETR Prep<with>
15		A-13	PUSH NP<an umbrella>
16	IV	A-5	SETR Det<an>
17		A-6	SETR N<umbrella>
18		A-7	POP NP
19	III	A-13	SETR NP <an umbrella>
20		A-14	POP PP <with an umbrella>
21	II	A-8	SETR PP <with an umbrella>
22		A-11	POP VP<hit the criminal with an umbrella>
23	I	A-2	SETR VP< hit the criminal with an umbrella >
24		A-3	SETR<The detective hit the criminal with an umbrella >
SUCCESS			

Table 2 Parsing Algorithm of Example 3 in "VP→VP PP"

The parsing algorithm of example 3 in "NP→NP PP" also has 24 steps and highest level of syntactic structure is "V", which means this parsing needs more cognitive or system burden to parse.

From Step 1 to Step 7, system parses example 3 along the same path in which both NP<the detective> and V<hit> are interpreted successfully without the existence of ambiguity. The same algorithm can be seen in both Table 2 and Table 3.

From Step 8, the difference appears. For the sake of clear and concise explanation, we start the algorithm used in Table 3 from step 8.

8 In arc 10, VP subnet seeks NP. Different from Step 8 in Table 2 where NP<the criminal> is pushed down to NP subnet, NP<the criminal with an umbrella> in Table 3 is pushed down, which means <with an umbrella> is just a modifier for <the criminal>.

9 In arc 5, NP subnet finds Det<the>.

10 In arc 6, NP subnet seeks N<criminal>.

11 In arc 4, NP subnet seeks PP<with an umbrella>, which will be pushed down to PP subnet.

12 In arc 12, PP subnet finds Prep<with>.

13 In arc 13, PP subnet seeks NP. NP<an umbrella> is pushed down to NP subnet again.

14 In arc 5, NP subnet seeks Det<an>.

15 In arc 6, N<umbrella> is parsed in NP subnet.

16 In arc 7, the result of parsing NP<an umbrella> is popped back to arc 13.

17 In arc 13, NP<an umbrella> is set in register.

18 In arc 14, the parsing result of PP<with an umbrella> is popped back to arc 4.

19 In arc 4, PP<with an umbrella> is set in register.

20 In arc 7, the parsing result of NP<the criminal with an umbrella> is popped back to arc 10.

21 In arc 10, NP<the criminal with an umbrella> is set in register.

22 In arc 11, the result of parsing VP<hit the criminal with an umbrella> is popped up to arc 2.

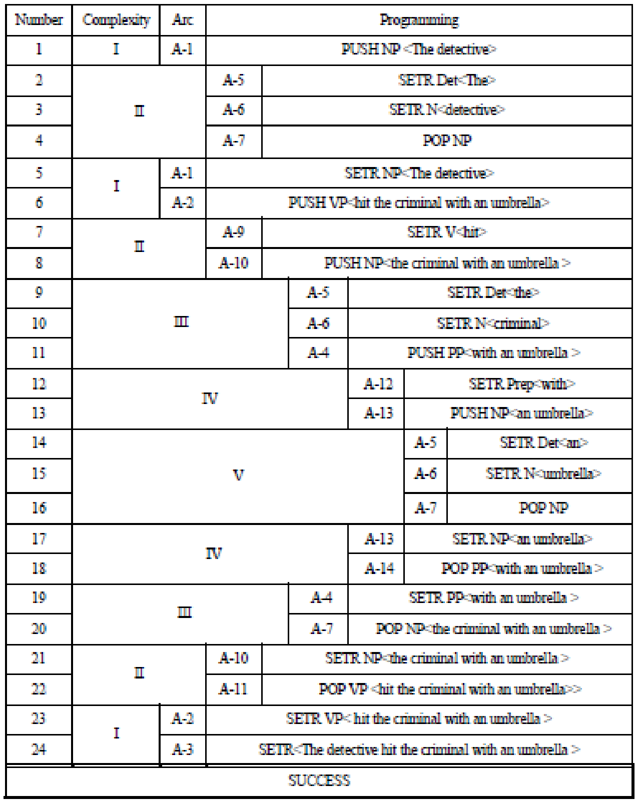
23 In arc 2, VP<hit the criminal with an umbrella> is set in register.

24 In arc 3, S<the detective hit the criminal with an umbrella> is parsed smoothly. System returns "SUCCESS" and parsing is over.

The difference between Table 2 and Table 3 shows that "VP→VP PP" parsing is easier than "NP→NP PP" parsing since the first is less complex than the second. This provides the evidence that there is a default parsing even though more than one interpretation is involved in an ambiguous sentence.

In example 3, "VP→VP PP" algorithm in which the sentence is parsed into "The detective using an umbrella hit the criminal" is the default interpretation.

Besides syntactic ambiguity shown in example 3, the existence of homographs is another important model to produce multi-meaning.



Number	Complexity	Arc	Programming
1	I	A-1	PUSH NP <The detective>
2	II	A-5	SETR Det<The>
3		A-6	SETR N<detective>
4		A-7	POP NP
5	I	A-1	SETR NP<The detective>
6		A-2	PUSH VP<hit the criminal with an umbrella>
7	II	A-9	SETR V<hit>
8		A-10	PUSH NP<the criminal with an umbrella >
9	III	A-5	SETR Det<the>
10		A-6	SETR N<criminal>
11		A-4	PUSH PP<with an umbrella >
12	IV	A-12	SETR Prep<with>
13		A-13	PUSH NP<an umbrella>
14	V	A-5	SETR Det<an>
15		A-6	SETR N<umbrella>
16		A-7	POP NP
17	IV	A-13	SETR NP<an umbrella>
18		A-14	POP PP<with an umbrella >
19	III	A-4	SETR PP<with an umbrella >
20		A-7	POP NP<the criminal with an umbrella >
21	II	A-10	SETR NP<the criminal with an umbrella >
22		A-11	POP VP <hit the criminal with an umbrella>>
23	I	A-2	SETR VP< hit the criminal with an umbrella >
24		A-3	SETR<The detective hit the criminal with an umbrella >
SUCCESS			

Table 3 Parsing Algorithm of Example 3 in "NP→NP PP"

Example 4: Failing student looked hard.

In example 4, both "failing" and "hard" have two meanings, namely, "failing(adj or Grd)" and "hard(adj or adv)". The semantic network of lexicon conveys the information.

(Failing((CTGY. GRD)))

(Failing((CTGY. ADJ)))

(Student((CTGY.N) (NUM. SING)))

(Look((CTGY.V)(PAST.LOOKED) (PASTP.LOOKED)))

(Looked((CTGY. V) (ROOT.LOOK) (TENSE. PAST)))

(Hard((CTGY. ADJ)))

(Hard((CTGY. ADV)))

From the lexicon, we can see the difference of homographs, which lead to four ambiguous sentences.

G={Vn, Vt, S, P}

Vn={N, NP, V, VP, S, Adv, Adj, Grd}

Vt={failing, student, looked, hard}

S=S

P:

1. S→NP VP
2. NP→Adj N
3. NP→Grd N
4. VP→V Adj
5. VP→V Adv
6. Adj→{failing, hard}
7. Grd→{failing}
8. N→{student}
9. V→{looked}
10. Adv→{hard}

Failing student looked hard (Grd+Adj)

a. Grd student looked hard (7)

b. Grd N looked hard (8)

c. NP looked hard (3)

d. NP V hard (9)

e. NP V Adj (6)

f. NP VP (4)

g. S (1)

h. SUCCESS

Failing student looked hard (Adj+Adj)

a. Adj student looked hard (6)

b. Adj N looked hard (8)

c. NP looked hard (2)

d. NP V hard (9)

e. NP V Adj (6)

f. NP VP (4)

g. S (1)

h. SUCCESS

Failing student looked hard (Grd+Adv)

a. Grd student looked hard (7)

b. Grd N looked hard (8)

c. NP looked hard (3)

d. NP V hard (9)

e. NP V Adv (10)

f. NP VP (5)

g. S (1)

h. SUCCESS

Failing student looked hard (Adj+Adv)

a. Adj student looked hard (6)

b. Adj N looked hard (8)

c. NP looked hard (2)

d. NP V hard (9)

e. NP V Adv (10)

f. NP VP (5)

g. S (1)

h. SUCCESS

According to the ambiguous interpretations of example 4, a special ATN used to analyze the sentence can be shown below.

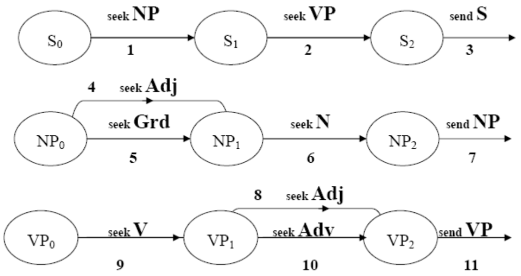
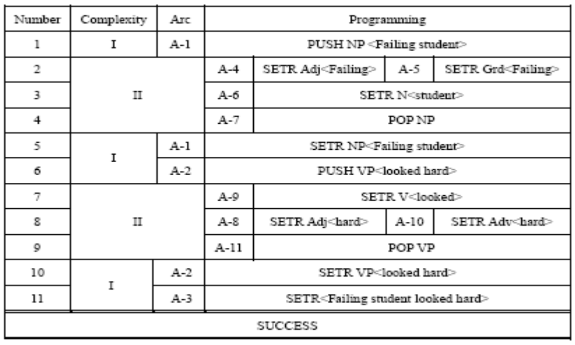


Figure 5 ATN of Example 4

In Fig. 5, we can see both NP subnet and VP subnet have bi-arcs which act as the same function of grammar. For example, arc 4 and arc 5 before NP1 exist in the same syntactic position and have the same function. Meanwhile, arc 8 and arc 10 before VP2 perform similar grammatical function in VP subnet. The BNF of example 4 is provided as follows.



Number	Complexity	Arc	Programming		
1	I	A-1	PUSH NP <Failing student>		
2	II	A-4	SETR Adj<Failing>	A-5	SETR Grd<Failing>
3		A-6	SETR N<student>		
4		A-7	POP NP		
5	I	A-1	SETR NP<Failing student>		
6		A-2	PUSH VP<looked hard>		
7	II	A-9	SETR V<looked>		
8		A-8	SETR Adj<hard>	A-10	SETR Adv<hard>
9		A-11	POP VP		
10	I	A-2	SETR VP<looked hard>		
11		A-3	SETR<Failing student looked hard>		
SUCCESS					

Table 4 Parsing Algorithm of Example 4

The whole BNF-based algorithm of example 4 is shown in Table 4, by which four interpretations discussed above can be parsed.

1 In arc 1, S-net needs NP and system pushes <failing student> to NP subnet.

2 In arc 4 and arc 5, NP subnet can parse <failing> as Adj or Grd. Both are correct and this is the first ambiguity. The parsing results are saved respectively.

3 In arc 6, N <student> is set in register.

4 In arc 7, NP<failing student> is parsed successfully (either Adj+N or Grd+N) and the result is popped back to arc 1 which needs the parsing result of NP<failing student>.

5 In arc 1, NP<failing student> is set in register

6 In arc 2, S-net seeks VP and <looked hard> is pushed down to the VP subnet.

7 In arc 9, V<looked> is found.

8 In arc 8 and arc 10, Adj<hard> or Adv<hard> is analyzed smoothly. This is the second ambiguity after the first one in arc 4 and arc 5.

9 In arc 11, the result of parsing (either V+Adj or V+Adv) is popped up to arc 2 where VP <looked hard> is pushed down.

10 In arc 2, the parsing result of VP <looked hard> is set in register.

11 In arc 3, S< failing student looked hard > is parsed successfully and smoothly, including four results of parsing, i.e. Adj+N+V+Adj, Adj+N+V+Adv, Grd+N+V+Adj, Grd+N+V+Adv. System returns "SUCCESS" and parsing is over.

From the discussion above, we can know a pre-grammatical sentence (e.g. example 1) is not good enough to meet the requirements of syntax for it fails to consist in the necessary components. A common sentence (e. g. example 2) is the essential part of natural language, and the exact expression is the core of the sentence. An ambiguous sentence comprises ambiguous structures (e.g. example 3) or ambiguous words (e.g. example 4), and any ambiguous interpretation is acceptable and understandable even though sometimes the parsing has different complexity.

III. The NLP-Focused Analyses Of Gp Sentences

The parsing of a GP sentence includes two procedures, i.e. the prototype understanding and the backtracking parsing. The prototype understanding refers to the default parsing of cognition according to decoder's knowledge database. The backtracking parsing means the original processing breaks down and the decoder has to re-understand the GP sentence when the new information used to decode the sentence is provided linearly. Therefore, processing breakdown is the distinctive feature of the parsing of GP sentence.

Example 5: The opposite number about 5000.

The sentence is a GP one which contains the prototype understanding and backtracking parsing. The decoding experiences the breakdown of cognition.

G={Vn, Vt, S, P}

Vn={Det, Adj, N, LinkV, Adv, Num, NP, VP, S, NumP}

Vt={the, opposite, number, about, 5000}

S=S

P:

1. S→NP VP

2. NP→Det Adj

3. NP→Det Adj N

4. NumP→Adv Num

5. VP→LinkV NumP
6. Det→{the}
7. N→{number}
8. Adv→{about}
9. LinkV→{number}
10. Adj→{opposite}
11. Num→{5000}

(The((CTGY. DET)))

(Opposite((CTGY. ADJ)))

(Number((CTGY.LINKV)(PAST.NUMBERED)(PASTP. NUMBERED)))

(Number((CTGY. LINKV)(TENSE. PRES))

(Number((CTGY. N) (NUM. SING)))

(About((CTGY.ADV)))

(5000((CTGY. NUM)))

The opposite number about 5000

a. Det opposite number about 5000 (6)

b. Det Adj number about 5000 (10)

c. Det Adj N about 5000 (7)

d. NP about 5000 (3)

e. NP Adv 5000 (8)

f. NP Adv Num (11)

g. NP NumP (4)

h. FAIL and backtrack to another path:

i. Det Adj number about 5000 (10)

j. NP number about 5000 (2)

k. NP LinkV about 5000 (9)

l. NP LinkV Adv 5000 (8)

m. NP LinkV Adv Num (11)

n. NP LinkV NumP (4)

o. NP VP (5)

p. S (1)

q. SUCCESS

From the lexicon analysis of example 5, we can notice the significant difference between "number (noun)" and "number (linking verb)".

According to the interpretation in LDOCE, "number (noun)" can mean "a word or sign that represents an amount or a quantity" just in the sentence of "Five was her lucky number"; or "a set of numbers used to name or recognize someone or something" in the sentence of "He refused to swap it with opposite number Willie Carne after the game because he had promised it to the Mirror."

Besides the noun function, "number" can be parsed as "lingking verb". For example, in the sentence of "The men on strike now number 5% of the workforce", "number" is interpreted as "if people or things number a particular amount, that is how many there are."

Based on the discussion above, ATN of example 5 can be created.

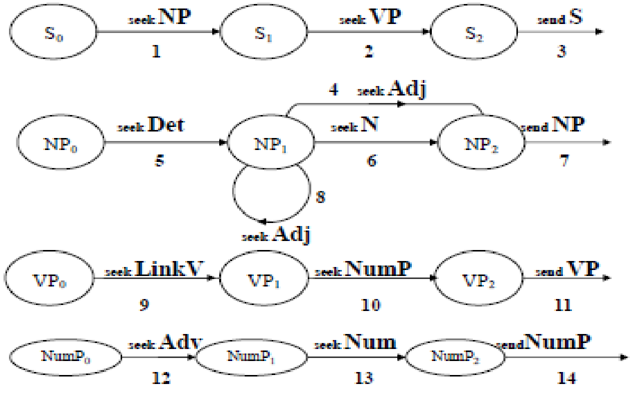


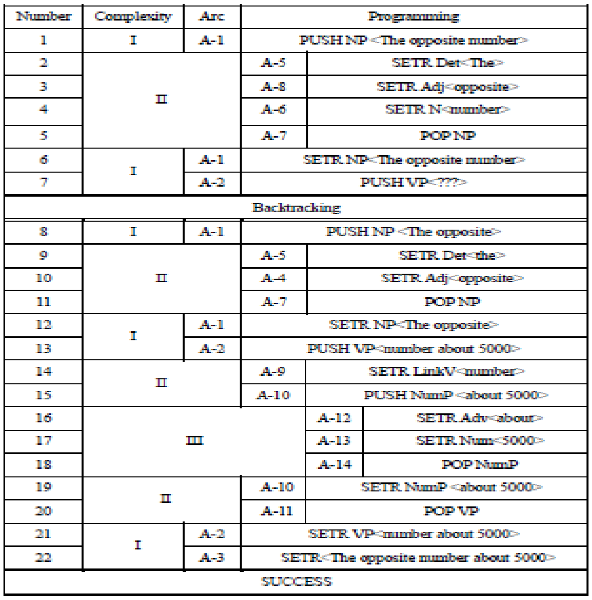
Figure 6 ATN of Example 5

In Fig. 6, the core of the parsing lies in NP subnet in which both "NP→Det Adj" and "NP→Det Adj N" are accepted. In cognitive system, "number (noun)" functions in order of priority while "number (lingking verb)"has a notably low probability. The difference of cognition can be shown in the ERP experiments and the psychological results develop the prototype ideas.[39-41]

The BNF-based algorithm of example 5 includes 22 steps during the parsing, which can be shown in Table 5.

1. In arc 1, S net firstly seeks NP. System pushes down to NP subnet. According to the cognitive knowledge of decoder, "number(noun)"in <the opposite number>" is firstly parsed.
2. In arc 5, Det<the> is set in register.
3. In arc 8, Adj<opposite> is interpreted successfully.
4. In arc 6, N<number> is set in register.
5. In arc 7, parsing result of NP<the opposite number> is popped up to arc 1 in S network where it is pushed down.
6. In arc 1, NP<the opposite number> is set in register.
7. In arc 2, S network seeks VP and tries to push down to VP subnet. But the left components<about 5000>fail to find V according to lexicon analysis. System returns "FAIL" and backtracks to the original path in arc 1 where another parsing can be chosen besides the original one. In example 5, the cognitive crossing lies in the difference of "number(noun)" and "number(linking verb)".
8. In arc 1, system seeks NP and <the opposite> instead of the original <the opposite number> is pushed down to NP subnet.

9. In arc 5, Det<the> is set in register.
10. In arc 4, Adj<opposite> is parsed.
11. In arc 7, NP<the opposite> is parsed successfully and sent back to arc 1.
12. In arc 1, the parsing result of NP<the opposite> is set in register.
13. In arc 2, VP<number about 5000> is pushed down to VP subnet.
14. In arc 9, <number> is interpreted as a linking verb according to (Number((CTGY. LINKV))), and the result of parsing is set in register.
15. In arc 10, VP subnet seeks NumP<about 5000>. NumP subnet is activated.
16. In arc 12, the interpretation of Adv<about> is set in register.
17. In arc 13, the number <5000> is parsed.
18. In arc 14, NumP<about 5000> is popped up to arc 10.
19. In arc 10, the result of parsing NumP<about 5000> is set in register.
20. In arc 11, after parsing VP<number about 5000> successfully and smoothly, system returns to arc 2.
21. In arc 2, VP<number about 5000> is set in register.
22. In arc 3, both NP<the opposite> and VP<number about 5000> are set in register and the whole parsing of S<The opposite number about 5000> is completed. System returns "SUCCESS" and parsing is over.



Number	Complexity	Arc	Programming
1	I	A-1	PUSH NP <The opposite number>
2	II	A-5	SETR Det<The>
3		A-8	SETR Adj<opposite>
4		A-6	SETR N<number>
5		A-7	POP NP
6	I	A-1	SETR NP<The opposite number>
7		A-2	PUSH VP<???>
Backtracking			
8	I	A-1	PUSH NP <The opposite>
9	II	A-5	SETR Det<the>
10		A-4	SETR Adj<opposite>
11		A-7	POP NP
12	I	A-1	SETR NP<The opposite>
13		A-2	PUSH VP<number about 5000>
14	II	A-9	SETR LinkV<number>
15		A-10	PUSH NumP <about 5000>
16	III	A-12	SETR Adv<about>
17		A-13	SETR Num<5000>
18		A-14	POP NumP
19	II	A-10	SETR NumP <about 5000>
20		A-11	POP VP
21	I	A-2	SETR VP<number about 5000>
22		A-3	SETR<The opposite number about 5000>
SUCCESS			

Table 5 Parsing Algorithm of Example 5

From the algorithm in Table 5, we can see the distinctive feature of parsing is the existence of "backtracking", at which breakdown happens and system has to return to the original crossing to find another road out. This optional procedure needs the help of lexical, semantic, grammatical and cognitive knowledge.

Example 6: The new record the song.

G={Vn, Vt, S, P}

Vn={Det, Adj, N, V, NP, VP, S}

Vt={the, new, record, song}

S=S

P:

1. S→NP VP
2. NP→Det Adj
3. NP→Det Adj N
4. NP→Det N
5. VP→V NP
6. Det→{the}
7. N→{record, song}
8. V→{record}
9. Adj→{new}

(The((CTGY. DET)))

(New((CTGY. ADJ)))

(Record((CTGY.V)(PAST.RECORDED)(PASTP. RECORDED)))

(Record((CTGY.V)(ROOT.RECORD)(TENSE. PRES)))

(Record((CTGY. N) (NUM. SING)))

(Song((CTGY.N) (NUM. SING)))

The new record the song

a.	Det new record the song	(6)
b.	Det Adj record the song	(9)
c.	Det Adj N the song	(7)
d.	NP the song	(3)
e.	NP Det song	(6)
f.	NP Det N	(7)
g.	NP NP	(4)
h.	FAIL and backtrack to another path:	
i.	Det Adj record the song	(9)
j.	NP record the song	(2)
k.	NP V the song	(8)
l.	NP V Det song	(6)
m.	NP V Det N	(7)
n.	NP V NP	(4)
o.	NP VP	(5)

p. S (1)

q. SUCCESS

From the parsing above, we can know example 6 is another GP sentence since there is breakdown in the processing. In example 6, “record(verb)” and “record(noun)” can be chosen randomly. However, NP<the new record> has a high probability of parsing. This is the reason why the priority parsing selects “record(noun)” rather than “record(verb)”. The process of choosing can be shown in ATN networks.

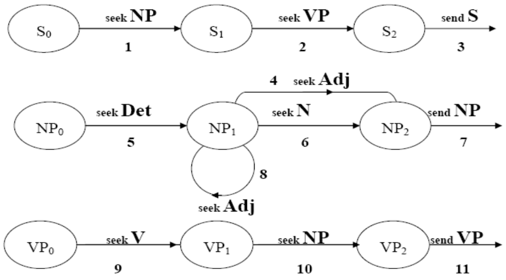


Figure 7 ATN of Example 6

In Fig. 7, NP subnet structure is the obvious reason why the GP phenomenon appears. Both NP→Det Adj and NP→ Det Adj N are reasonable and acceptable when “the new record” is parsed. Generally speaking, Adj is used to modify the Noun, the model of NP→Det Adj N is the prototype of parsing, and system interprets example 6 by means of this programming rule rather than NP→Det Adj. After completing the NP subnet parsing of <the new record>, system returns to S network to seek VP. However, the left phrase <the song> has no VP factor according to the lexicon knowledge, and system stops, backtracks and transfers to another programming rule, i.e. NP→Det Adj. Cognitive breakdown happens. The whole processing algorithm of example 6 is shown in Table 6.

1. In arc 1, S network needs NP information. The prototype of NP<the new record> has higher probability than NP<the new> in decoder’s cognition, and NP<the new record> is pushed down to NP subnet.
2. In arc 5, system finds Det<the>.
3. In arc 8, Adj<new> is found.
4. In arc 6, N<record> is matched.
5. In arc 7, system finishes the parsing NP<the new record> and returns to arc 1.
6. In arc 1, the parsing result of NP<the new record> is saved.
7. In arc 2, system seeks VP information. However, no related lexicon knowledge is provided in (The((CTGY. DET))) and (Song((CTGY.N) (NUM. SING))). System fails and backtracks to arc 1 to find another programming rule of NP→Det Adj instead of NP→Det Adj N.
8. In arc 1, NP<the new> is chosen as a new alternative. NP subnet is activated once more.
9. In arc 5, Det<the> is set in register.

Number	Complexity	Arc	Programming
1	I	A-1	PUSH NP <The new record>
2	II	A-5	SETR Det<The>
3		A-8	SETR Adj<new>
4		A-6	SETR N<record>
5		A-7	POP NP
6	I	A-1	SETR NP<The new record>
7		A-2	PUSH VP<???>
Backtracking			
8	I	A-1	PUSH NP <The new>
9	II	A-5	SETR Det<the>
10		A-4	SETR Adj<new>
11		A-7	POP NP
12	I	A-1	SETR NP<The new>
13		A-2	PUSH VP<record the song>
14	II	A-9	SETR V<record>
15		A-10	PUSH NP <the song>
16	III	A-5	SETR Det<the>
17		A-6	SETR N<song>
18		A-7	POP NP
19	II	A-10	SETR NP <the song>
20		A-11	POP VP
21	I	A-2	SETR VP<record the song>
22		A-3	SETR<The new record the song>
SUCCESS			

Table 6 Parsing Algorithm of Example 6

10. In arc 4, Adj<new> is interpreted successfully.
11. In arc 7, NP<the new> is parsed completely and the result is popped back to arc 1.
12. In arc 1, the popped result of NP is set in register.
13. In arc 2, system seeks VP and <record the song> is pushed down to VP subnet.
14. In arc 9, VP subnet is activated and the knowledge of (Record((CTGY.V)(ROOT.RECORD)(TENSE. PRES))) helps system regard <record> as verb.
15. In arc 10, VP subnet seeks NP. The NP<the song> is pushed down to NP subnet.
16. In arc 5, NP subnet is activated again. Det<the> is set in register.
17. In arc 6, N<song> is set in register.
18. In arc 7, NP<the song> is parsed completely and the result is popped up to arc 10 where it is pushed down.
19. In arc 10, the parsing result of NP<the song> is set in register.
20. In arc 11, VP<record the song> is parsed successfully and popped up to arc 2.
21. In arc 2, the parsing result of VP<record the song> is set in register.

22. In arc 3, system finishes the parsing of NP<the new> and VP<record the song>. S<the new record the song> is saved. System returns “SUCCESS” and parsing is over.

From the discussion about example 5 and example 6, we can find both of them have the distinctive feature of “backtracking”. The fact that high probability parsing in GP sentences has to be replaced by the low probability interpretation is the fundamental distinction from pre-grammatical sentences, common sentences and ambiguous sentences. Processing breaks down when system backtracks to find new path out.

Based on the analyses of computational linguistics shown above, we can see more likeness and unlikeness exist between the ambiguous sentences and GP sentences. An effective and systematic attempt at comparison and contrast may contribute to our understanding of the special phenomenon.

IV. The Comparison And Contrast Of Ambiguous Sentences And Gp Sentences

Ambiguous sentences and GP sentences have close similarities and significant differences in many aspects, e.g. lexicon knowledge, syntactic structures and decoding procedures.

A. The Similarity and Difference in Lexicon Knowledge

The lexicon knowledge is the basic information for system to parse and a detailed analysis of related category is essential and necessary. Let’s firstly compare the similarity and contrast the difference among example 3, example 4, example 5 and example 6, which are shown as follows.

In example 3, the lexicon analysis includes Det<the, an>, N<detective, criminal>, Prep<with> and V<hit>. Since the singular noun N<detective> needs present verb <hits> or past verb <hit> to cooperate, example 3 must be a past tense rather than a present tense for there is no <hits> provided in the sentence. Example 3 is a structure-based ambiguous sentence and lexicon knowledge helps few for reducing ambiguities.

(The((CTGY. DET)))

(Detective((CTGY.N) (NUM. SING)))

(Hit((CTGY. V) (PAST. HIT) (PASTP. HIT)))

(Hit((CTGY. V) (ROOT. HIT) (TENSE.PAST)))

(Hit((CTGY. V) (ROOT. HIT) (TENSE.PASTP)))

(Criminal ((CTGY.N) (NUM. SING)))

(With((CTGY.PREP)))

(An((CTGY. DET)))

(Umbrella((CTGY.N) (NUM. SING)))

In example 4, lexicon knowledge contains the analyses of Grd<failing>, Adj<failing, hard>, N<student>, V<looked>, Adv<hard>. The homonyms of <failing> and <hard> bring the double ambiguities in the sentence, which results in four different meanings. The whole ambiguity lies in the lexical multi-meaning. Therefore, example 4 is the model of lexical ambiguity.

(Failing((CTGY. GRD)))

(Failing((CTGY. ADJ)))

(Student((CTGY.N) (NUM. SING)))

(Look ((CTGY.V)(PAST.LOOKED) (PASTP.LOOKED)))

(Looked((CTGY. V) (ROOT.LOOK) (TENSE. PAST)))

(Hard((CTGY. ADJ)))

(Hard((CTGY. ADV)))

In example 5, the lexical database comprises Det<the>, Adj<opposite>, LinkV<number>, N<number>, Adv<about>, and Number<5000>. The homonym <number> has two grammatical functions, i.e. linking verb and noun.

The different choices result in different sentences. According to the probability, NP<the opposite number> is the prototype parsing, and correspondingly, N<number> is adopted firstly even though this path is considered to be a dead end finally. Generally speaking, the lexical crossing leads to the processing breakdown of GP sentence.

(The((CTGY. DET)))

(Opposite((CTGY. ADJ)))

(Number((CTGY.LINKV)(PAST.NUMBERED)(PASTP. NUMBERED)))

(Number((CTGY. LINKV)(TENSE. PRES))

(Number((CTGY. N) (NUM. SING)))

(About((CTGY.ADV)))

(5000((CTGY. NUM)))

In example 6, a lot of lexicons are analyzed, i.e. Det<the>, Adj<new>, V<record> and N<record, song>. The meaning of <record> diverges markedly when N<record> is replaced by V<record> to meet the requirements of syntax. This sentence is another example in which processing breakdown is a direct consequence of lexical divergence.

(The((CTGY. DET)))

(New((CTGY. ADJ)))

(Record((CTGY.V)(PAST.RECORDED)(PASTP. RECORDED)))

(Record((CTGY.V)(ROOT.RECORD)(TENSE. PRES)))

(Record((CTGY. N) (NUM. SING)))

(Song((CTGY.N) (NUM. SING)))

From the discussion above, we can see the existence of homonyms is an obvious reason which brings ambiguous phenomenon and GP effect, just as in example 4, example 5 and example 6.

However, this is not the only reason for the appearance of ambiguity or GP phenomenon. Sometimes, the divergence of syntactic structures also leads to ambiguity or GP effect.

B. The Similarity and Difference in Syntactic Structures

Stanford parser is a very useful parser which is created by means of both highly optimized PCFG (probabilistic context free grammar), lexicalized dependency parsers and lexicalized PCFG. “Probabilistic parsers use knowledge of language gained from hand-parsed sentences to try to produce the most likely analysis of new sentences.” The Stanford parser can be used to parse example3, example 4, example 5 and example 6 on line. The results of syntactic structures are provided as follows.

In example 3, the tags include <the/DT>, <detective/NN>, <hit/VBD>, <criminal/NN>, <with/IN>, <an/DT>, and <umbrella/NN>. The parsing structure is a full parsed one in which <the detective> is parsed as NP; <hit the criminal with an umbrella>, VP; <the criminal>, sub-net’s NP; <with an umbrella>, sub-net’s PP parsed as a modifier for <hit>; <with>, sub-net’s PP; <an umbrella>, sub-sub-net’s NP. The hierarchical structure is similar to the complexity in Table 2. Stanford parser provides one of the two interpretations, namely model of “VP→VP PP” rather than the model of “NP →NP PP” since the former has higher probability than the latter from the perspective of statistics. In other words, “VP→ VP PP” is the prototype parsing for its simpler syntactic structure.

```
(ROOT
 (S                                          (I)
   (NP (DT the) (NN detective))                        (II)
   (VP (VBD hit)                             (II)
     (NP (DT the) (NN criminal))                       (III)
     (PP (IN with)                           (III)
       (NP (DT an) (NN umbrella))))          (IV)
    (. .)))
```

In example 4, the tags are <Failing/NN>, <student/NN>, <looked/VBD> and <hard/JJ>. This is another whole parsed structure in which all the components are interpreted successfully. The word of <failing> is considered Noun (i.e. Grd); <hard>, JJ (i.e. Adj). The parsed syntactic structure is similar to “Grd+Adj” which is the highest probability in statistics of parsing database among four ambiguous models. The hierarchical level is II shown in Table 4.

```
(ROOT
 (S                                          (I)
   (NP (NN Failing) (NN student))            (II)
   (VP (VBD looked)(ADJP (JJ hard)))         (II)
    (. .)))
```

In example 5, tags are < the/DT >, < opposite/JJ >, <number/NN >, < about/RB> and <5000/CD >. According to Stanford parser, this is a part-parsed sentence since the final result is NP rather than S, which shows the prototype of NP<the opposite number> has the higher probability than NP<the opposite>. In other words, Stanford parser only finishes the first part of the parsing before the backtracking in Table 5.

```
(ROOT
 (NP                                         (I)
   (NP (DT the) (JJ opposite) (NN number))           (II)
   (QP (RB about) (CD 5000))                 (II)
    (. .)))
```

In example 6, tags comprise <the/DT>, <new/JJ>, <record/NN >, and <song/NN>. This is another example of part-parsed structure in which only the programming rule of N→{record} is adopted while V→{record} fails to be used. That means NP<the new record> has stronger statistical probability than NP<the new >. Stanford parser only parses the steps from 1-7 in Table 6 and then system gives the final result is NP instead of S, which ignores the left parsing steps after the backtracking.

```
(ROOT
 (NP                                         (I)
   (NP (DT the) (JJ new) (NN record))        (II)
   (NP (DT the) (NN song))                   (II)
    (. .)))
```

From the discussion about syntactic structures, we can see both ambiguous sentences and GP sentences can have more than one syntactic structure. According to PCFG, the strongest probability parsing is the final result in Stanford parser. If another more complex structure is adopted, cognitive burden of decoders will be lifted and increased. Once this happens, another ambiguous sentence will be provided by means of the ambiguous syntactic structure besides the original one. On the contrast, if probability-based parsing returns the final result of a GP sentence as a part-parsed structure, the rule-based programming will be activated and a full-parsed new structure can be obtained only if the processing breakdown can be overcome.

During the re-parsing procedures, an ambiguous structure can bring different full-parsed results, while a GP sentence breaks down firstly for its part-parsed structure and then moves on to another full-parsed path. An ambiguous structure leads to multi-results, all of which are reasonable and acceptable while a GP sentence structure only brings one full-interpreted result besides the processing breakdown.

V. Conclusion

By comparing programming procedures, lexicon knowledge, parsing algorithms and syntactic structures between pre-grammatical sentences, common sentences, ambiguous sentences and GP sentences, we conclude that the formal methods of computational linguistics, e.g. CFG, BNF, and ATN, are useful for computational parsing. Pre-grammatical sentences have part-parsed structure and system returns the final result to be Phrases rather than S. Common sentences are normal in grammar and semantics, and there is no lexical or syntactic crossing for parsing. Ambiguous sentences have ambiguity created by ambiguous structures or lexicons, both of which can bring full-parsed results. GP sentences comprise part-parsed structure built by the high statistical probability method, and full-parsed structures created by rule-based method. When the parsing shifts from part-parsed structure to the full-parsed one, processing breakdown of GP sentences occurs. This paper supports the idea raised by Pritchett [42]that processing breakdown is a distinctive feature in the parsing of a GP sentence.

ACKNOWLEDGMENT

This research is supported in part by grants of YB115-29 and YB115-41 from "the Eleventh Five-year" research projects of Chinese language application, and grant 11YJA740111 from the Ministry of education and science planning project.

REFERENCES

[1] B. Manaris, "Natural language processing: A human-computer interaction perspective," Advances in Computers, vol. 47, 1998, pp. 1-66.

[2] P. Jackson and F. Schilder, "Natural language processing: Overview," Encyclopedia of Language & Linguistics, 2006, pp. 503-518.

[3] D. E. Suranjan, P. A. N. Shuh-Shen, and A. B. Whinston, "Natural language query processing in a temporal database," Data & Knowledge Engineering, vol. 1, June 1985, pp. 3-15.

[4] E. Métais, "Enhancing information systems management with natural language processing techniques," Data & Knowledge Engineering, vol. 41, June 2002, pp. 247-272.

[5] G. Neumann, "Interleaving natural language parsing and generation through uniform processing," Artificial Intelligence, vol. 99, Feb. 1998, pp. 121-163.

[6] M. Maybury, "Natural language processing: System evaluation," Encyclopedia of Language & Linguistics, 2006, pp. 518-523.

[7] C. Mellish and X. Sun, "The semantic web as a linguistic resource: Opportunities for natural language generation," Knowledge-Based Systems, vol. 19, Sep. 2006, pp. 298-303.

[8] W. W. Chapman et al, "Classifying free-text triage chief complaints into syndromic categories with natural language processing," Artificial Intelligence in Medicine, vol. 33, Jan. 2005, pp. 31-40.

[9] J. A. Bateman, J. Hois, R. Ross, and T. Tenbrink, "A linguistic ontology of space for natural language processing," Artificial Intelligence, vol. 174, Sep. 2010, pp. 1027-1071.

[10] P. F. Dominey, T. Inui, and M. Hoen, "Neural network processing of natural language: Towards a unified model of corticostriatal function in learning sentence comprehension and non-linguistic sequencing," Brain and Language, vol. 109, June 2009, pp. 80-92.

[11] M. Stanojević, N. Tomašević, and S. Vraneš, "NIMFA – natural language implicit meaning formalization and abstraction," Expert Systems with Applications, vol. 37, Dec. 2010, pp. 8172-8187.

[12] V. R. Dasigi, and J. A. Reggia, "Parsimonious covering as a method for natural language interfaces to expert systems," Artificial Intelligence in Medicine, vol. 1, 1989, pp.49-60.

[13] S. J. Conlon, J. R. Conlon, and T. L. James, "The economics of natural language interfaces: Natural language processing technology as a scarce resource," Decision Support Systems, vol. 38, Oct. 2004, pp. 141-159.

[14] S. Menchetti, F. Costa, P. Frasconi, and M. Pontil, "Wide coverage natural language processing using kernel methods and neural networks for structured data," Pattern Recognition Letters, vol. 26, Sep. 2005, pp. 1896-1906.

[15] E. Alba, G. Luque, and L. Araujo, "Natural language tagging with genetic algorithms," Information Processing Letters, vol. 100, Dec. 2006, pp. 173-182.

[16] W. M. Wang, C. F. Cheung, W. B. Lee, and S. K. Kwok, "Mining knowledge from natural language texts using fuzzy associated concept mapping," Information Processing & Management, vol. 44, September 2008, pp. 1707-1719.

[17] P. Cimiano, P. Haase, J. Heizmann, M. Mantel, and R. Studer, "Towards portable natural language interfaces to knowledge bases – The case of the ORAKEL system," Data & Knowledge Engineering, vol. 65, May 2008, pp. 325-354.

[18] L. Zhou and G. Hripcsak, "Temporal reasoning with medical data—A review with emphasis on medical natural language processing," Journal of Biomedical Informatics, vol. 40, April 2007, pp. 183-202.

[19] R. Plant and S. Murrell, "A natural language help system shell through functional programming," Knowledge-Based Systems, vol. 18, Feb. 2005, pp. 19-35.

[20] J. L. Du, P. F. Yu, "Syntax-directed machine translation of natural language: Effect of garden path phenomenon on sentence structure," International Conference on Intelligent Systems Design and Engineering Applications, 2010, pp. 535－539.

[21] J. Häussler and M. Bader, "The assembly and disassembly of determiner phrases: Minimality is needed, but not sufficient," Lingua, 119(10), 2009, pp.1560-1580.

[22] T. Malsburg and S. Vasishth, "What is the scanpath signature of syntactic reanalysis?" Journal of Memory and Language, 65(2), 2011, pp.109-127.

[23] K. R. Christensen, "Syntactic reconstruction and reanalysis, semantic dead ends, and prefrontal cortex," Brain and Cognition, 73(1), 2010, pp. 41-50.

[24] T. G. Bever, "The cognitive basis for linguistic structures," In J. R. Hayes, Ed. Cognition and the Development of Language, New York: John Wiley and Sons, 1987, pp. 279-352.

[25] Y. H. Jin, "Semantic analysis of Chinese garden-path sentences," Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, 2006, (7), pp. 33–39.

[26] K. Christianson et al, "Thematic roles assigned along the garden path linger," Cognitive Psychology, 2001, (42), pp. 368–407.

[27] M. P. Wilson and S. M. Garnsey, "Making simple sentences hard: Verb bias effects in simple direct object sentences," Journal of Memory and Language, 2009, 60(3), pp. 368-392.

[28] A. D. Endress and M. D. Hauser, "The influence of type and token frequency on the acquisition of affixation patterns: Implications for language processing," Journal of Experimental Psychology: Learning, Memory, and Cognition, vol. 37, Jan. 2011, pp. 77-95.

[29] D. J. Foss and C. M. Jenkins, "Some effects of context on the comprehension of ambiguous sentences," Journal of Verbal Learning and Verbal Behavior, 1973, (12), pp. 577.

[30] K. G. D. Bailey and F. Ferreira, "Disfluencies affect the parsing of garden-path sentences," Journal of Memory and Language, 2003, (49), pp. 183–200.

[31] M. Bader and J. Haussler, "Resolving number ambiguities during language comprehension," Journal of Memory and Language, 2009, (08).

[32] N. D. Patson et al, "Lingering misinterpretations in garden-path sentences: Evidence from a paraphrasing task," Journal of Experimental Psychology: Learning, Memory, and Cognition, 2009, 35(1), pp. 280-285.

[33] J. Feeney, D. Coley, and A. Crisp, "The relevance framework for category-based induction: Evidence from garden-path arguments," Journal of Experimental Psychology: Learning, Memory, and Cognition, vol. 36, July 2010, pp. 906-919.

[34] Y. Choi and J. C. Trueswell, "Children's (in)ability to recover from garden paths in a verb-final language: Evidence for developing control in sentence processing," Journal of Experimental Child Psychology, 2010, 106(1), pp. 41-61.

[35] P. F. Yu and J. L. Du, "Automatic analysis of textual garden path phenomenon: A computational perspective," Journal of Communication and Computer, 2008, 5 (10), pp. 58-65.

[36] B, McMurray, M. K. Tanenhaus and R. N. Aslin, "Within-category VOT affects recovery from 'lexical' garden-paths: Evidence against phoneme-level inhibition," Journal of Memory and Language, 2009, 60(1), pp. 65-

91.

[37] P. L. O'Rourke and C. V. Petten, “Morphological agreement at a distance: Dissociation between early and late components of the event-related brain potential,” Brain Research, 2011, 1392(5), pp. 62-79.

[38] J. L. Du and P. F. Yu, “Towards an algorithm-based intelligent tutoring system: computing methods in syntactic management of garden path phenomenon,” Intelligent Computing and Intelligent Systems, 2010, (10), pp. 521-525.

[39] E. Malaia, R. B. Wilbur, and C. Weber-Fox, “ERP evidence for telicity effects on syntactic processing in garden-path sentences,” Brain and Language, 2009, 108(3), pp.145-158.

[40] A. Staub, “Eye movements and processing difficulty in object relative clauses,” Cognition, 2010, 116(1), pp. 71-86.

[41] M. O. Ujunju, G. Wanyembi and F. Wabwoba. “Evaluating the role of information and communication technology (ICT) support towards processes of management in institutions of higher learning,” International Journal of Advanced Computer Science and Applications, 2012, 3(7), pp. 55-58.

[42] B. L. Pritchett, “Garden path phenomena and the grammatical basis of language processing,” Language, 1988, (64), pp. 539-576.

Effect of Driver Strength on Crosstalk in Global Interconnects

Kalpana.A.B
Assistant Professor,
Department of Electronics and Communication
Bangalore Institute of Technology
Bangalore, India

P.V.Hunagund
Professor
Department of Applied Electronics
Gulbarga University
Gulbarga, India

Abstract— **The Noise estimation and avoidance are becoming critical, in today's high performance IC design. An accurate yet efficient crosstalk noise model which contains as many driver/interconnect parameters as possible, is necessary for any sensitivity based noise avoidance approach. In this paper, we present an analysis for crosstalk noise model which incorporates all physical properties including victim and aggressor drivers, distributed RC characteristics of interconnects and coupling locations in both victim and aggressor lines. Also shown that crosstalk can be minimized by driver sizing optimization technique. These models are verified for various deep submicron technologies.**

Keywords- Coupling; crosstalk; Interconnect; noise; victim.

I. Introduction

Coupling capacitance between neighboring nets is a dominant component in today's deep submicron designs as taller and narrower lines are being laid out closer to each other [1]. This trend is causing the ratio of crosstalk capacitance to the total capacitance of a wire to increase. On top of these interconnect related trends, more aggressive and less noise immune circuit structures such as dynamic logic are being employed more commonly due to performance requirements.

As a result, a significant crosstalk noise problem exists in today's high performance designs. The net on which noise is being induced is called the *victim* net whereas the net that induces this noise is called the *aggressor* net. Crosstalk noise not only leads to modified delays [2, 3] but also to potential logic malfunctions [4, 5]. To be able to deal with the challenges brought by this recently emerging phenomenon, techniques and tools to estimate and avoid crosstalk noise problems should be incorporated into the IC design cycle from the early stages. Any such tool requires fast yet accurate crosstalk noise models both to estimate noise and also to see the effects of various interconnect and driver parameters on noise. Several papers, which propose crosstalk models, can be found in recent literature. In [6], telegraph equations are solved directly to find a set of analytical formulae for peak noise in capacitively coupled bus lines. [7] derives bounds for crosstalk noise using a lumped model but assuming a step input for aggressor driver. The peak noise expression in [7] is extended by [8, 9] to consider a saturated ramp input and a π circuit to represent the interconnect. These models fail to represent the distributed nature of the interconnect. In [10], an Elmore delay like peak noise model is obtained for general RC trees but it assumes an infinite ramp input. This assumption causes the model to significantly overestimate peak noise, especially for small aggressor slews, which is very likely to occur in today's deep submicron designs. Devgan's metric has been improved in [11]. Interconnect crosstalk can be modeled and minimized using different techniques [12, 13] It is also shown that crosstalk can be minimized by driver sizing optimization technique [14, 15].

II. Noise Avoidance Technique: Driver Sizing

A general case for two coupled lines is shown in Figure 1. Both aggressor and victim lines are divided into 3 regions: interconnect segment before coupling location, coupling location and interconnect segment after coupling location. These regions of aggressor and victim lines are represented by L_{al}, L_c, L_{ar}, L_{vl} and L_{vr} as seen in the figure 3. We propose the linear model shown in Figure 4 to compute crosstalk noise at the receiver of victim net. Victim driver is modeled by effective holding resistance R_h whereas aggressor driver is modeled by an effective Thevenin model consisting of a saturated ramp voltage source with a slew rate of t_r and the Thevennin resistance R_{th}. Other components of our model are computed based on the technology and geometrical information obtained from Figure 1. Coupling node (node 2 in aggressor net and node 5 in victim net) is defined to be the middle of coupling location for both nets, i.e. L_{al} + L_c/2 away from aggressor driver and L_{vl} + L_c/2 away from the victim driver. For the aggressor net, let the upstream and downstream resistance-capacitance at node 2 be R_{a1}-C_{au} and R_{a2}-C_{ad} respectively. Then, C_{a1} = C_{au}/2, Ca2 = (C_{au}+C_{ad})/2 and C_{a3} = C_{ad}/2+C_{la}. Similarly for the victim net, let the upstream and downstream resistance capacitance pair at

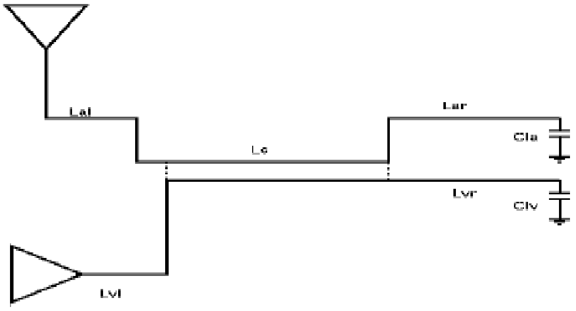


Figure 1. Linear crosstalk noise model

node 5 be R_{v1}-C_{vu} and R_{v2} -C_{vd} respectively. Then, $C_{v1} = C_{vu}/2$, $C_{v2} = (C_{vu} + C_{vd})/2$ and $C_{v3} = C_{vd}/2 + C_{lv}$.

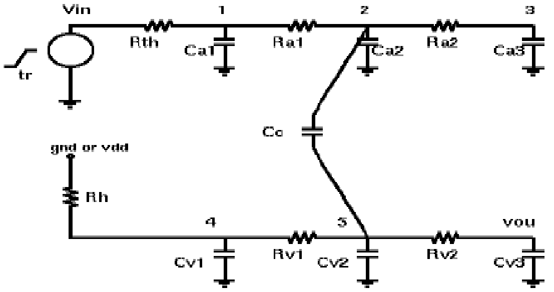


Figure 2. Linear crosstalk noise model

To simplify the analytical calculation of transfer function H(s) from V_{in} to V_{out}, we initially decouple the aggressor line from victim line (Figure 3 (a)), and compute the transfer function from V_{in} to V_2. We then apply $V_2(s)$ to the victim line as seen in Figure 3 (b). This assumption is valid when victim line is not loading aggressor line at node 2 significantly.

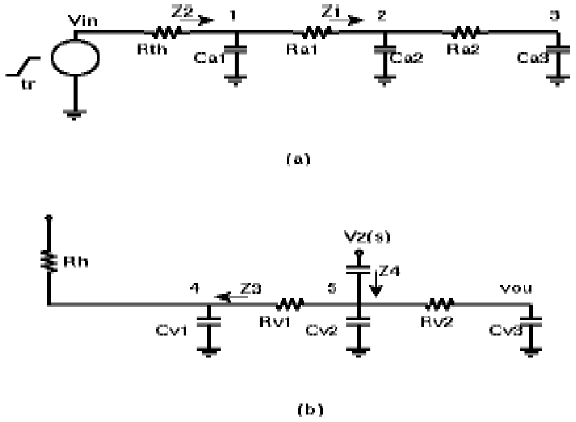


Figure 3. Decoupled model to calculate transfer Function.

We will look at driver sizing both from the point of view of victim driver sizing and aggressor driver sizing. Intuitively, if a victim driver is sized up, its effective conductance increases thus it becomes stronger to hold a net at a steady voltage (V_{dd} or ground). On the other hand, if an aggressor driver is sized down, its effective conductance decreases thus it cannot transition as fast and as a result noise amount that it can induce on a victim net decreases. Victim driver is modeled by effective holding resistance R_h whereas aggressor driver is modeled by an effective Thevenin model consisting of a saturated ramp voltage source with a slew rate of t_r and the Thevenin resistance R_{th}. Using our model, we have calculated the sensitivity of peak noise to R_h and

R_{th} which represent victim and aggressor driver sizes, respectively.

$$\frac{\delta_{vpeak}}{\delta R_h} = \frac{C_C}{t_r}\left(1 - e^{-t_r / t_v}\right) - \left(R_h + R_{v1}\right)\frac{C_C\left(C_C + C_{v1} + C_{v2} + C_{v3}\right)}{t_v^2} e^{-t_r/t_v} \qquad (15)$$

$$\frac{\delta v_{peak}}{\delta R_{th}} = \frac{-\left(R_h + R_{v1}\right)C_C\left(C_{a1} + C_{a2} + C_{a3}\right)}{t_v^2} e^{-t_r/t_v} \qquad (16)$$

Since Equation (16) is always negative, sizing down the aggressor driver (i.e., sizing up R_{th}) will always reduce peak noise. But how effective a reduction it will be, depends on the parameters of Equation (16). Increasing R_{th} will be more effective on noise reduction if the numerator of Equation (16) is greater than its denominator.

If the equation parameters are carefully observed, this mathematical condition translates to the following circuit condition. Noise reduction effect of increasing R_{th} is more, when we have a strong aggressor (strong aggressor driver, wide/short aggressor line). The effects of sizing up victim driver (i.e.sizing down R_h) is more complicated. In terms of peak noise reduction, victim driver sizing becomes a more effective noise avoidance tool as the RC time constant of victim line decreases.

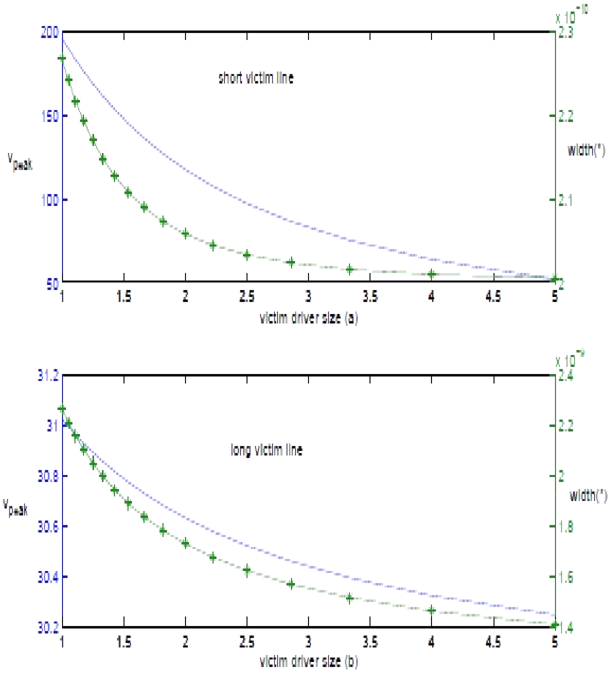


Figure 4. Sensitivity of victim driver sizing effects to victim line properties

Figure 4(a) shows the effects of victim driver sizing on a short victim line. Note that peak noise voltage is reduced by 75mV/38.5% whereas noise width is reduced by 22ps/9.6% when victim driver size is doubled. As RC time constant of victim line increases, victim driver sizing becomes less effective in terms of peak noise reduction but it is important to notice the effects on noise width.

As seen in Figure 6(b), victim driver sizing on a long victim line reduces noise width by 550ps/24% while peak noise is reduced by 0.4mV/1% when victim driver size is doubled. One other important observation about victim driver sizing is the diminishing returns effect.

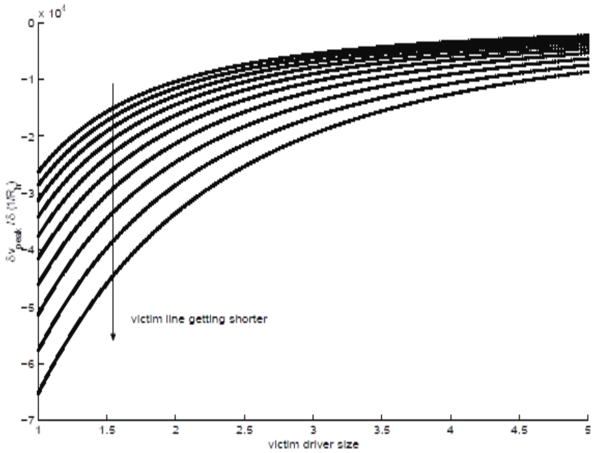


Figure 5. Diminishing returns effect in victimdriver sizing.

Figure 5 shows change in $\delta v_{peak}/\delta(1/R_h)$ as victim driver is sized up, for a range of victim line lengths. As can be seen, the effect of driver sizing diminishes as victim driver is sized up. A driver sizing tool should take this effect into account to be able to steer away from non-optimal sizes and to make sure that the area trade-off is worthwhile.

III. Results

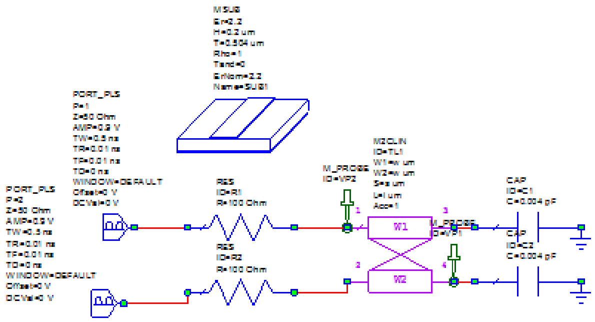


Figure 6. Experimental circuit using AWR

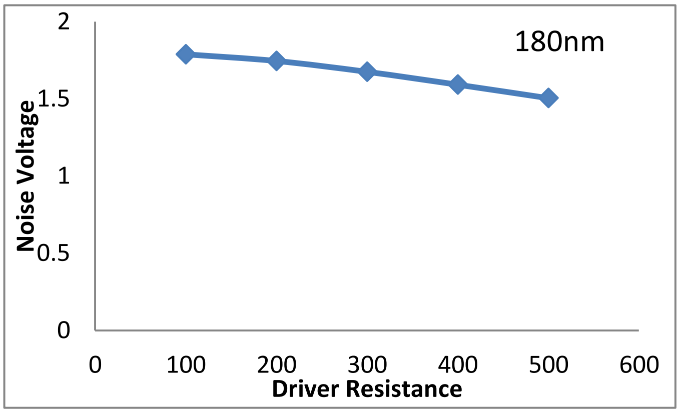


Figure 7. Noise voltage with change in driver resistance for 180 nm

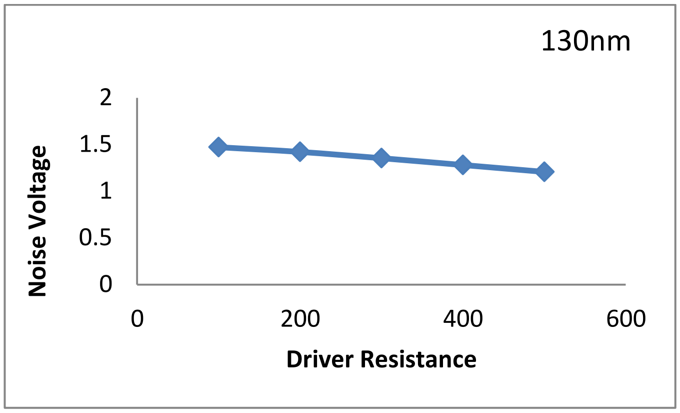


Figure 8. Noise voltage with change in driver resistance for 130 nm

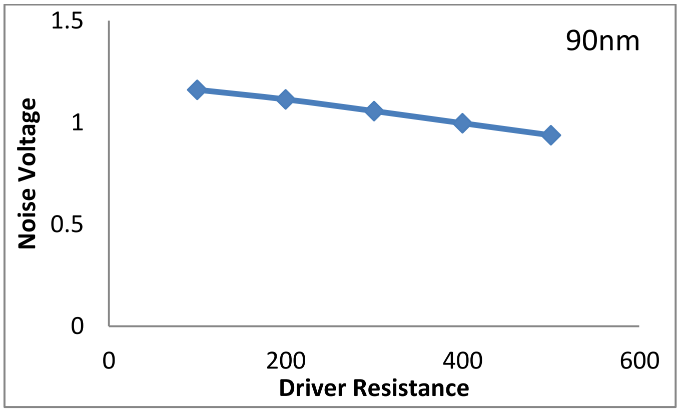


Figure 9 Noise voltage with change in driver resistance for 90 nm

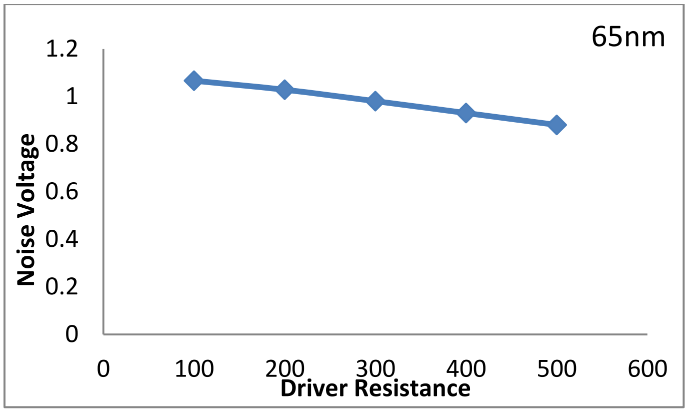


Figure 10. Noise voltage with change in driver resistance for 65nm

Figure 6. shows the experimental setup used for simulation in AWR software.

Figure 7.to figure 11. Shows the variation in crosstalk noise voltage with the change in driver resistance for different technology nodes.

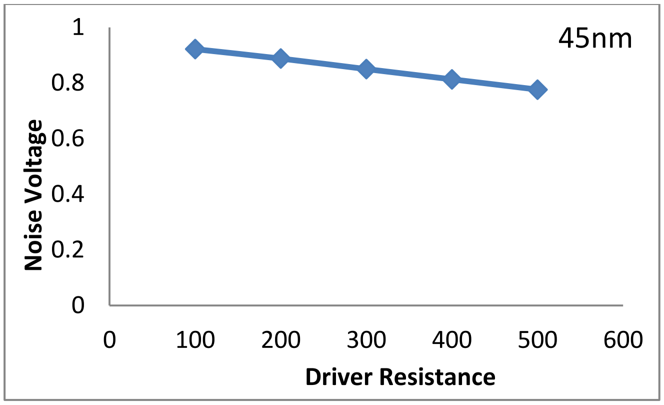


Figure 11. Noise voltage with change in driver resistance for 45nm

III. Conclusion

In this paper, we presented analysis for crosstalk noise model which incorporates all victim and aggressor driver/interconnect physical parameters including coupling locations on victim and aggressor nets, distributed RC characteristics of interconnects. Crosstalk noise minimization technique using driver sizing also developed and validated for deep submicron technologies. Output voltage is observed for increased driver size and shown that crosstalk can be minimized by driver optimization.

References

[1]. S. I. Association. The international technology roadmap for semiconductors, 1999.

[2]. P. D. Gross, R. Arunachalam, K. Rajagopal, and L. T. Pileggi. Determination of worst-case aggressor alignment for delay calculation. In Proceedings of the IEEE International Conference on Computer-Aided Design, ICCAD-98, 1998.

[3]. S. Sirichotiyakul, D. Blaauw, C. Oh, R. Levy, V. Zolotov, and J. Zuo. Driver modeling alignment for worst-case delay noise. In Proceedings of Design Automation Conference DAC, pages 720–725, June 2001.

[4]. S. Alwar, D. Blaauw, A. Dasgupta, A. Grinshpon, R. Levy, C. Oh, B. Orshav, S. Sirichotiyakul, and V. Zolotov. Clarinet: A noise analysis tool for deep submicron design. In Proceedings of Design Automation Conference DAC, pages 233–238, June 2000.

[5]. K. L. Shepard and V. Narayanan. Noise in deep submicron digital design. In Proceedings of ICCAD-96 Intl. Conference on Computer Aided Design, pages 524–531, November 1996.

[6]. T. Sakurai. Closed-form expression for interconnect delay, coupling, and crosstalk in VLSIs. IEEE Transactions on Electron Devices, 40:118–124, 1993

[7]. A. Vittal and M. Marek-Sadowska. Crosstalk reduction for VLSI. IEEE Transactions on Computer Aided Design, 16:290–298, March 1997.

[8]. A. Vittal, L. H. Chen, M. Marek-Sadowska, K. P. Wang, and S. Yang. Crosstalk in VLSI interconnections. IEEE Transactions on Computer Aided Design, 18:1817–1824, December 1999.

[9]. A. B. Kahng, S. Muddu, and D. Vidhani. Noise and delay uncertainty studies for coupled rc interconnects. In Proceedings of ASIC/SOC Conference, pages 3–8, 1999.

[10]. A. Devgan. Efficient coupled noise estimation for on-chip interconnects. In Proceedings of the IEEE International Conference on Computer-Aided Design, ICCAD-97, pages 147– 153, 1997.

[11]. M. Kuhlmann and S. S. Sapatnekar. Exact and efficient crosstalk estimation. IEEE Transactions on Computer Aided Design, 20(7):858–866, July 2001.

E-learning System Which Allows Students' Confidence Level Evaluation with Their Voice When They Answer to the Questions During Achievement Tests

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

Abstract— **E-learning system which allows students' confidence level evaluation with their voice when they answer to the question during achievement tests is proposed. Through experiments of comparison of students' confidence level between the conventional (without evaluation) and the proposed (with evaluation), 17-57% of improvement is confirmed for the proposed e-learning system.**

***Keywords**- learnng system; confidence level evaluation; emotion recognition with voice.*

I. INTRODUCTION

Under the ADL: Advanced Distributed Learning Initiatives [1], Sharable Content Object Reference Model: SCORM[2] which is a collection of standards and specifications for web-based e-learning is promoted [1]. It defines communications between client side content and a host system called the run-time environment, which is commonly supported by a learning management system. Reusability, accessibility, inter-operability, and maintainability are important for the SCORM standard.

One of the issues to be discussed for the conventional e-learning system is that improvement of achievement level. In other word, effectiveness of the e-learning system as well as e-learning contents is one of the major issues. Although there are many suspected causes, quality of achievement test is one of them. Namely, students can precede one step forward even if they do not have confidence. Because only think students have to do is click a supposed appropriate radio button among four or five candidate radio button as possible answers. Thus the students may get trouble when they get one step advance even if they do not have confidence.

E-learning system can be divided into two categories, synchronous and on-demand type. In particular, the synchronous type includes a quasi-real time based Q and A systems. Students may get an answer when they make a question. Therefore, effectiveness of the synchronous type is better than the on-demand type. For both types of e-learning system, achievement tests are important. The proposed e-learning system allows check confidence levels during achievement tests. Therefore, achievement test results can be evaluated much properly rather than that without confidence evaluations. Confidence level evaluation can be done with students' voice for the proposed e-learning system.

In the following section, the proposed e-learning system is described followed by some experiments with students. Then conclusion with some discussions is flowed.

II. PROPOSED E-LEARNING SYSTEM

A. Fundamentals of Confidence Level Evaluations

It is assumed that voice input and output software is installed in the proposed e-learning system in advance. In particular, voice input and output software is used for confidence level evaluation during achievement tests period. If confidence level is not high enough, then such students have to conduct another achievement test again.

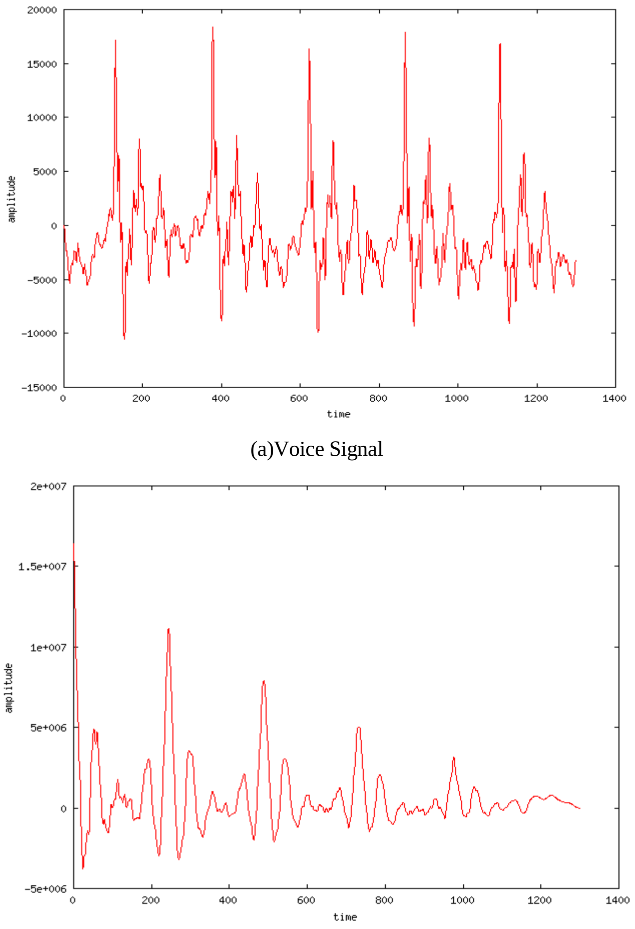
There are some methods which allow evaluation of confidence level with students' voices and moving pictures during they answer to questions in achievement tests. With moving picture, it can be recognized that students are ill at ease, or are not in a calm situation in particular during achievement test. It is much easy to check students' confidence level using their voice. Frequency components as well as loudness of voice can be used. These features are referred to pitch frequency[3] and power level, hereafter. The pitch frequency is defined as fundamental frequency which can be estimated with auto-correlation function, r_l of human voice signals (equation (1)).

$$r_l = \frac{1}{N} \sum_{t=0}^{N-l-1} x_t x_{t+l} \tag{1}$$

where N and x_t denotes the number of samples of voice signals and voice signal itself, respectively. The typical human voice signal is shown in Fig.1 (a) while typical auto-

[1] http://www.adlnet.org/
[2] http://en.wikipedia.org/wiki/Sharable_Content_Object_Reference_Model
[3] http://en.wikipedia.org/wiki/Pitch_detection_algorithm

correlation function[4] is shown in Fig.1 (b). From the auto-correlation function, pitch frequency can be determined.



(a)Voice Signal

(b)Auto-correlation function of (a)

Figure 1 Typical human voice signal and its auto-correlation function.

On the other hand, students' voice loudness, power level, P can be calculated with equation (2).

$$P=\sqrt{\frac{\sum_{N}^{i=0} x_i^2}{N}} \qquad (2)$$

B. *Evaluation of Students' Confidence Level*

Fig.2 shows an example of two dimensional scatter plots of the pitch frequency and the power level of the students' voices during achievement tests. Typical scatter plot of students' voices during achievement tests in the two dimensional distribution between pitch frequency and power level is shown in Fig.3. In general, students' voices that have a high confidence level during achievement tests are loud and include high pitch frequency components while students' voices that do not have enough confidence level during achievement tests are not loud and do not include enough high pitch frequency components.

It depends on personal voice characteristics. The student whose voice includes high pitch frequency components usually there are high pitch frequency components during achievement tests as well. The student who speaks loudly always answers to the questions loudly. Therefore, some normalization is required for pitch frequency and loudness during achievement test by using those in calm status (Normal situation).

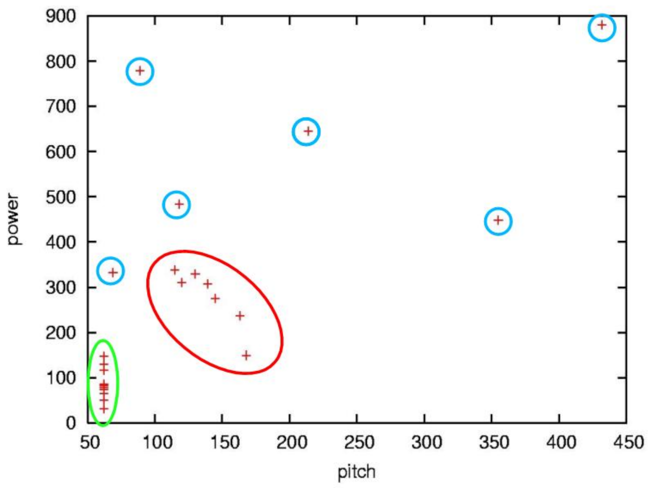


Figure 2 Example of scatter plot of students' voice between pitch frequency and power level during achievement tests.

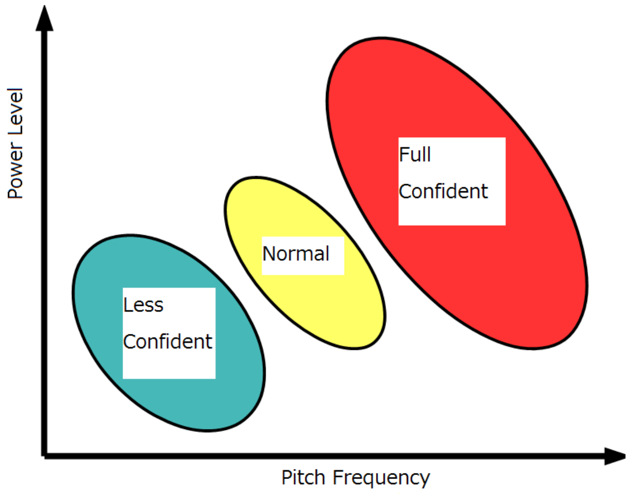


Figure 3Typical scatter plot of students' voices during achievement tests in the two dimensional distribution between pitch frequency and power level

Just before getting start achievement tests, each student has to say their student ID and their name. The proposed e-learning system, then, input their voice and plot their pitch frequency and power level on two dimensional scatter diagrams as those in calm status or normal situation. After that, student begins achievement tests. Pitch frequency and power level of students' voice is plotted on the same two dimensional feature planes. Then, gravity center[5] is calculated in a real time basis.

[4] http://en.wikipedia.org/wiki/Autocorrelation

[5] http://ejje.weblio.jp/content/center+of+gravity

Fig.4 shows gravity centers of students' voice of pitch frequency and power level. Plot #1 denotes the gravity center of the students' voice plots of which students have a high confidence level while Plot #2 denotes the gravity center of the students' voice plots of which students are in calm status or normal situation. Plot #3, on the other hand, denotes the gravity center of the students' voice plots of which students do not have a high confidence level during answering to the questions in achievement tests.

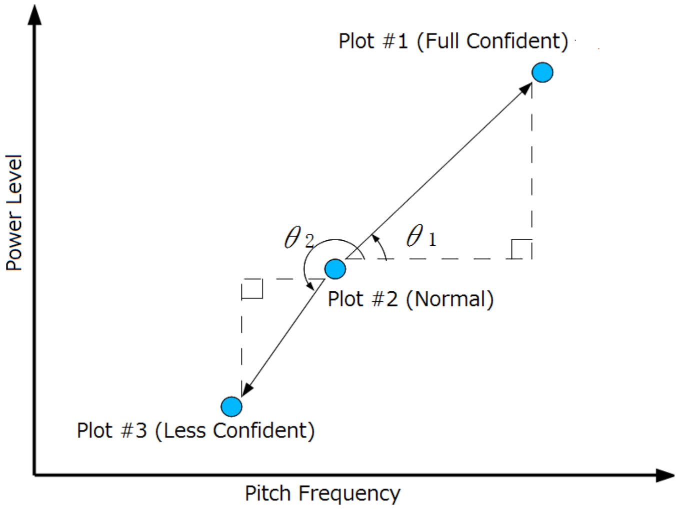


Figure 4 Gravity centers of students' voice plots on the two dimensional feature plane between pitch frequency and power level when students get start their achievement tests and when answering to the questions in achievement tests..

During scatter plots of pitch frequency and power level, cluster analysis can also be applied to the data plots. There are some clustering methods[6]. The proposed e-learning system uses hierarchical clustering method for human emotion recognitions. There are minimum distance, maximum distance, gravity, median, Ward's methods in the hierarchical clustering method. The proposed method uses Dendrogram[7] utilized Ward's method[8] [2]. It is one of hierarchical clustering method based on the following equation for representation of dis-similarity[9], d_{tr} between two clusters, t and r, which are created from cluster p and q,

$$d_{tr} = \frac{n_p + n_r}{n_t + n_r} d_{pr} + \frac{n_q + n_r}{n_t + n_r} d_{qr} - \frac{n_r}{n_t + n_r} d_{pq} \quad (3)$$

where n_i denotes the number of data in the cluster i in concern. Thus clusters are created by step by step basis as shown in Fig.6. Starting from dark blue and yellow, through light blue and orange, then purple, and finally green colored cluster is created eventually.

Process flow of these processes is shown in Fig.7.

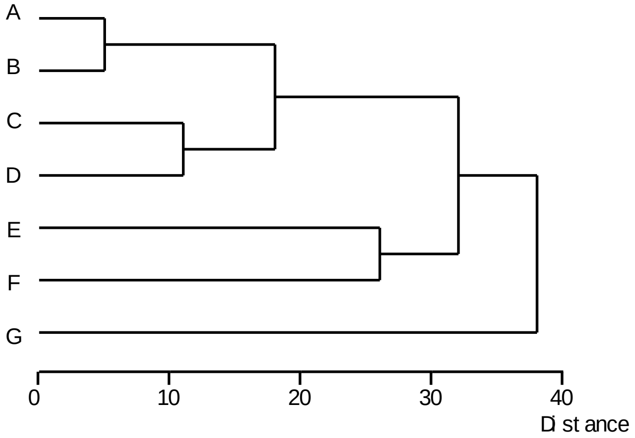


Figure 5 Example of Dendrogram

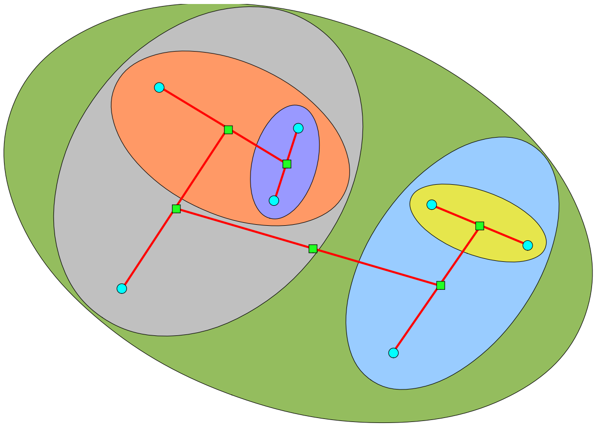


Figure 6 Clusters creation for Ward's method of clustering method

III. Implementation And Experiments

A. Implementation

The proposed e-learning system is implemented on a Windows XP OS machine. Question and answer system and e-learning contents are created with Java script with Internet Explore of web browser. On the other hand, students' emotion recognition software is created with gcc of C programming language. It contains real time voice recognition software tool. Screen shot image is shown in Fig.8.

B. Experiments

10 students are participated the experiment. Firstly, students have to input their voice, just say their names, to the proposed e-learning system in a calm status, normal situation. Then the pitch frequency and power level is plotted on feature plane. After that gravity center of the scatter plots is determined and it becomes standard axis for determination of the angle which corresponds to confidence level.

Through the experiments with 10 students, around 87.6% of confident or not confident classification performance is confirmed by comparing subjective and objective evaluation of confidence levels. 10 questions which include three programming Language related questions from Synthetic Personality Inventory: SPI test, three general questions from SPI test which are not related to programming language, and four questions of physics are provided to each student.

[6] http://en.wikipedia.org/wiki/Cluster_analysis
[7] http://en.wikipedia.org/wiki/Dendrogram
[8] http://en.wikipedia.org/wiki/Ward's_method
[9] http://en.wikipedia.org/wiki/Hierarchical_clustering

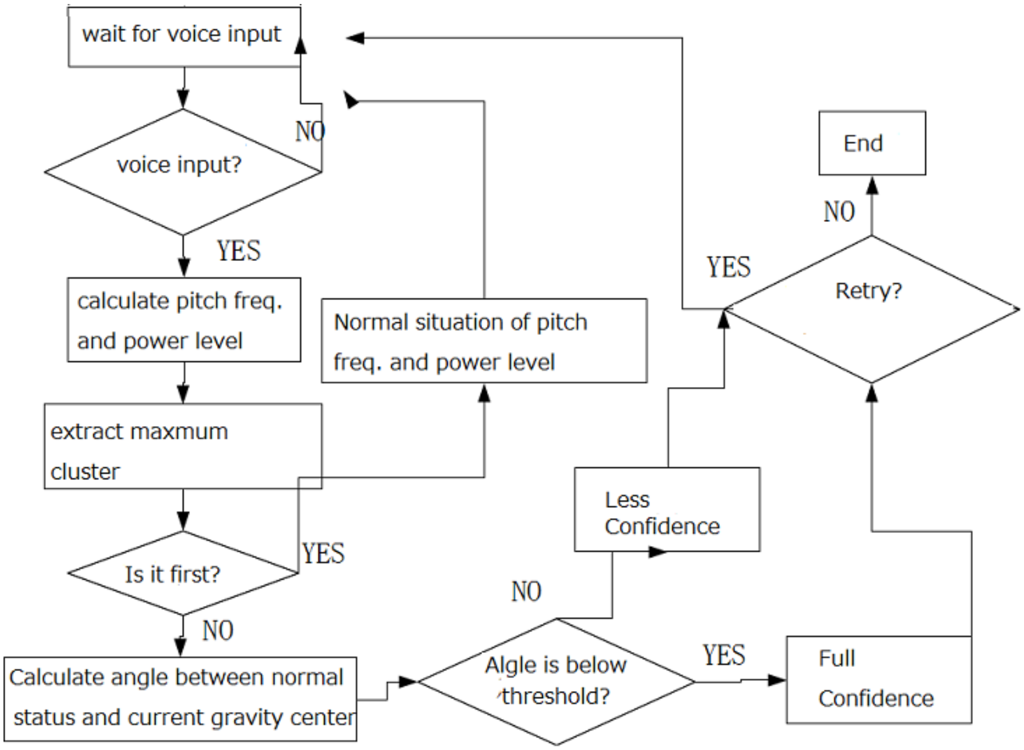
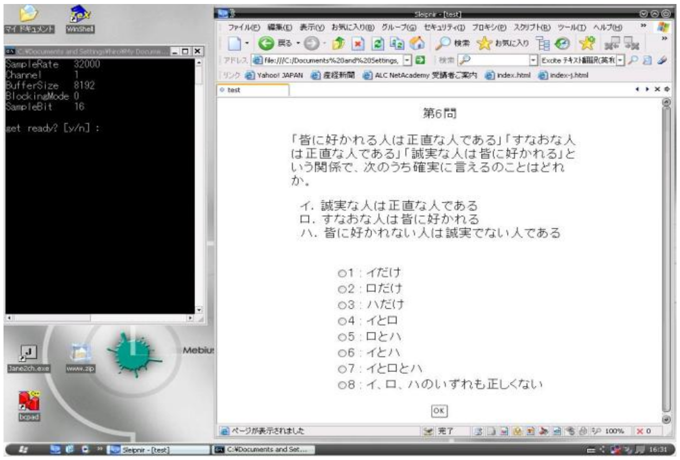

wait for voice input
voice input?
NO
YES
calculate pitch freq. and power level
Normal situation of pitch freq. and power level
extract maxmum cluster
Is it first?
YES
NO
Calculate angle between normal status and current gravity center
Algle is below threshold?
NO
YES
Less Confidence
Full Confidence
Retry?
YES
NO
End

Figure 7 Process flow of the proposed e-learning system with students’ confidence level evaluation using their voices during achievement tests



```
SampleRate    32000
Channel       1
BufferSize    8192
BlockingMode  0
SampleBit     16

get ready? [y/n] :
```

第6問

「皆に好かれる人は正直な人である」「すなおな人は正直な人である」「誠実な人は皆に好かれる」という関係で、次のうち確実に言えるのことはどれか。

イ. 誠実な人は正直な人である
ロ. すなおな人は皆に好かれる
ハ. 皆に好かれない人は誠実でない人である

○1：イだけ
○2：ロだけ
○3：ハだけ
○4：イとロ
○5：ロとハ
○6：イとハ
○7：イとロとハ
○8：イ、ロ、ハのいずれも正しくない

Figure 8 Screen shot image of e-learning content on web browser.

Then students have to take look at the explanations for each question. If the proposed e-learning system decides the student does not have enough confidence, then such students have to have another 10 questions of which the difficulty of the questions are almost same as previous questions. After that, the score of the tests before and after the retry test.

Also, pre-exercise is prepared. Pre-exercise uses the explanation of questions. The experiments are conducted with and without pre-exercise. The experimental results with and without pre-exercise is shown in Table 1. In the table, elapsed time is also evaluated. It takes much long time for the first test with pre-exercise in comparison to the elapsed time for the first test without pre-exercise. This is because students have to read the explanations for the questions first then answer to the questions. The elapsed time for the second test with pre-exercise is very fast because most of students feel confidence to their answer.

TABLE I. ACHIEVEMENT TEST RESULTS WITH AND WITHOUT PRE-EXERCISE OF THE FIRST AND SECOND TESTS

Pre Exercise		Average Score	Elapsed Time(s)
Without	1st Test	68	13'11"
	2nd Test	54	13'03"
	Improvement	-20%	26'14"
with	1st Test	48	29"33"
	2nd Test	62	8'58"
	Improvement	29%	38'31"

TABLE II. CHIEVEMENT TEST RESULTS WITH AND WITHOUT PRE-EXERCISE OF THE FIRST AND SECOND TESTS FOR EACH SUBJECT

Pre Exercise		Language	General	Physics
without	1st Test	28	18	20
	2nd Test	12	18	18
	Improvement	-117%	0%	-10%
with	1st Test	12	14	22
	2nd Test	14	22	26
	Improvement	16.70%	57.10%	18.20%

Improvement on achievement test scores is different by subjects. Improvements for programming language related questions and physics are around 16.7-18.2% while that for general questions is 57.1%. The score for the programming language related questions is essentially poor while that for physics is essentially good. That is the reason for the improvement depends on subject. On the other hand, general questions are essentially easy to answer and students feel a little bit confusion. Students have careless mistakes at the first test even if they have pre-exercises. Therefore, students answer to the questions without confidence. The confusion, however, disappears in the second test. Therefore, improvement of the score is remarkable.

After the experiments, we conduct interviews for each student. Their impressions are almost same as the previously supposed aforementioned reasons.

IV. Conclusion

E-learning system which allows students' confidence level evaluation with their voice when they answer to the question during achievement tests is proposed. Through experiments of comparison of students' confidence level between the conventional (without evaluation) and the proposed (with evaluation), 17-57% of improvement is confirmed for the proposed e-learning system.

Further improvement is required for human emotion recognition performance with several sources, not only pitch frequency and loudness of voice but also students' motions and eye movement using moving pictures.

Acknowledgment

The author would like to thank Mr. Hiroshi Yoshida for his effort to the experiments.

References

[1] Mikio Takagi, Haruhisa Shimoda Ed. Kohei Arai et al., Image Analysis Handbook, The University of Tokyo Publishing Inc., 1991.

[2] Kohei Arai, Hiroshi Yoshida, e-learning system with confidence evaluation using student voice, Technical Notes of Faculty of Science and Engineering, Saga University, 36, 1, 39-44, 2007.

Authors Profile

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science, and Technology of the University of Tokyo from 1974 to 1978 also was with National Space Development Agency of Japan (current JAXA) from 1979 to 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He was appointed professor at Department of Information Science, Saga University in 1990. He was appointed councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was also appointed councilor of Saga University from 2002 and 2003 followed by an executive councilor of the Remote Sensing Society of Japan for 2003 to 2005. He is an adjunct professor of University of Arizona, USA since 1998. He also was appointed vice chairman of the Commission "A" of ICSU/COSPAR in 2008. He wrote 30 books and published 332 journal papers.

Error Analysis of Air Temperature Profile Retrievals with Microwave Sounder Data Based on Minimization of Covariance Matrix of Estimation Error

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

Abstract— **Error analysis of air temperature profile retrievals with microwave sounder data based on minimization of covariance matrix of estimation error is conducted. Additive noise is taken into account in the observation data with microwave sounder onboard satellite. Method for air temperature profile retrievals based on minimization of difference of brightness temperature between model driven microwave sounder data and actual microwave sounder data is also proposed. The experimental results shows reasonable air temperature retrieval accuracy can be achieved by the proposed method.**

Keywords- Error analysis; leastsquare method; microwave sounder;air temperature profile.

I. INTRODUCTION

Air temperature and water vapor profiles are used to be estimated with Microwave Sounder data [1]. One of the problems on retrieving vertical profiles is its retrieving accuracy. In particular, estimation accuracy of air-temperature and water vapor at tropopause[1] altitude is not good enough because there are gradient changes of air-temperature and water vapor profile in the tropopause due to the fact that observed radiance at the specific channels are not changed by the altitude [2].

In order to estimate air-temperature and water vapor, minimization of covariance matrix of error is typically used. In the process, error covariance matrix[2] which is composed with the covariance of air temperature and water vapor based on prior information and the covariance of observed brightness temperature[3] based on a prior information as well as difference between model driven and the actual brightness temperature. Error analysis [4] is important for design sensitivity and allowable observation noise of microwave sounder. For this reason, error analysis is conducted for the conventional air temperature profile retrieval method. Other than this, this paper propose another air temperature profile retrieval method based on minimization of brightness temperature difference between model driven and actual brightness temperature acquired with real microwave sounder [5]. Experiment is conducted for the proposed method. Reasonable retrieval accuracy is confirmed.

The following section describes the conventional air temperature and water vapor profile retrieval method followed by excremental results. Then another retrieval method is proposed with some experimental results. Finally, conclusion is followed together with some discussions.

II. ERROR ANALYSIS

A. Microwave Sounder

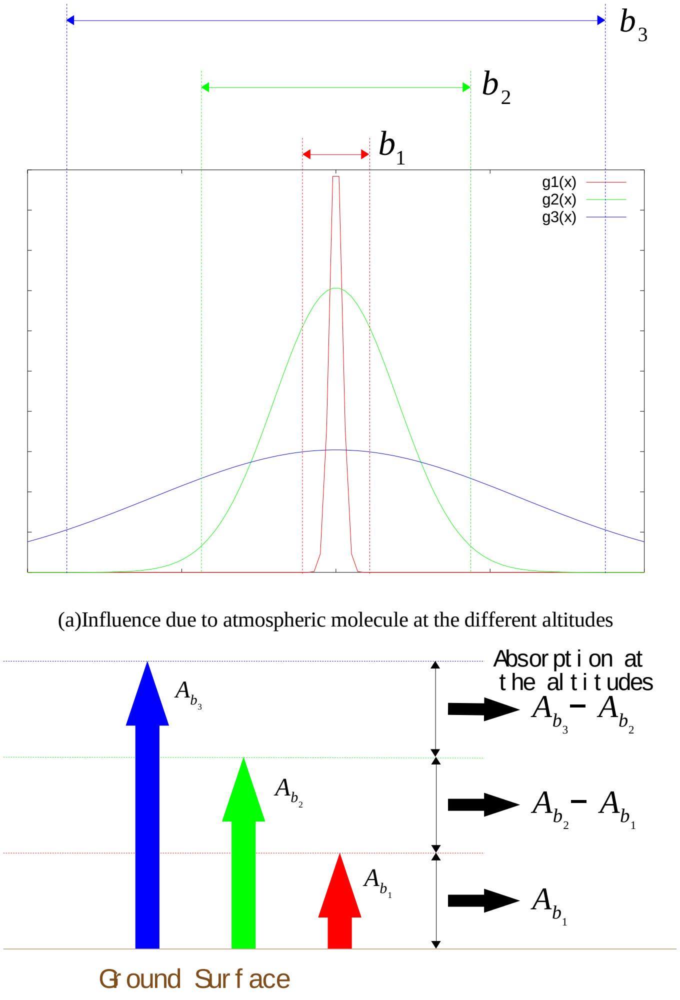
Air temperature profile can be retrieved with the microwave sounder data at absorption wavelength due to oxygen while water vapor profile can be estimated with the microwave sounder data at the absorption wavelength due to water. The microwave sounder which is onboard AQUA satellite [6] as well as NOAA-15, 16, 17 is called Advanced Microwave Sounding Unit: AMSU[7]. Description of AMSU is available in Analytical Theoretical Basis Document: ATBD document[8]. Observation frequency ranges from 23.8 GHz to 89 GHz. 22.235 GHz is the absorption frequency due to water while absorption frequency due to oxygen is situated in 60 GHz frequency bands. At the absorption frequency, observed brightness temperature is influenced by the molecule, oxygen, water. The influence due to molecule depends on the observation altitude as shown in Fig.1 (a). Also absorption due to atmospheric molecules depends on the observation altitudes as shown in Fig.1 (b). Therefore, it is possible to estimate molecule density of oxygen and water at the different altitude results in air temperature and water vapor profiles retrievals.

[1] http://en.wikipedia.org/wiki/Tropopause
[2] http://en.wikipedia.org/wiki/Covariance_matrix
[3] http://en.wikipedia.org/wiki/Brightness_temperature
[4] http://en.wikipedia.org/wiki/Error_analysis
[5] http://en.wikipedia.org/wiki/Advanced_Microwave_Sounding_Unit
[6] http://en.wikipedia.org/wiki/Aqua_(satellite)
[7] http://disc.sci.gsfc.nasa.gov/AIRS/documentation/amsu_instrument_guide.shtml
[8] http://eospso.gsfc.nasa.gov/eos_homepage/for_scientists/atbd/docs/AIRS/atbd-airs-L1B_microwave.pdf

Weighting function [9] is defined as the gradient of atmospheric transparency against altitude. The weighting function depends on observation frequency. Observed brightness temperature at the frequency, therefore, is influenced depending on the weighting function. Therefore, the altitude of which peak of weighting function is situated is the most influencing to the observed brightness temperature at the observation frequency. The following observation frequencies are selected for estimation of oxygen absorption (air temperature at the following altitudes,

15, 18, 20, 23, 14, 19, 7 km

58.7, 59.3, 60.2, 60.5, 61.8, 62.3, 63.7 GHz



(a)Influence due to atmospheric molecule at the different altitudes

(b)Absorption due to atmospheric molecule at the different altitudes

Figure 1 Absorption and influence due to atmospheric molecules at the different altitudes.

The weighting functions for these observation frequencies are shown in Fig.2. Using Millimeter wave Atmospheric Emission Simulator: MAES[10] of radiative transfer calculation software code [11] provided by National Institute for Communication Technology, Japan, NICT [12], atmospheric transparency can be calculated at the observation frequency. In this case, Mid. Latitude Summer of atmospheric model[13] is selected. Then gradient of atmospheric transparency against altitude is calculated results in weighting function calculations.

B. *Conventional Air Temperature and Water Vapor Profile Retrieval Method*

In order to estimate air-temperature and water vapor, minimization of covariance matrix of error is typically used. In the process, covariance matrix which is composed with the covariance of air temperature and water vapor based on prior information [14] and the covariance of observed brightness temperature based on a prior information as well as difference between model driven and the actual brightness temperature. Covariance matrix of estimation error is defined as follows,

$$X - X_0 = (S_x^{-1} + A^T * S_\epsilon^{-1} * A)^{-1} * A^T * S_\epsilon^{-1} * (G - G_0) \qquad (1)$$

where X_0, S_x, A, S_E, G, G_0 denote air temperature at each altitude, covariance matrix of air temperature for a prior information, Jacobian matrix[15] for brightness temperature of each frequency band, covariance matrix of observation error for a prior information, model driven brightness temperature, and estimated brightness temperature, respectively.

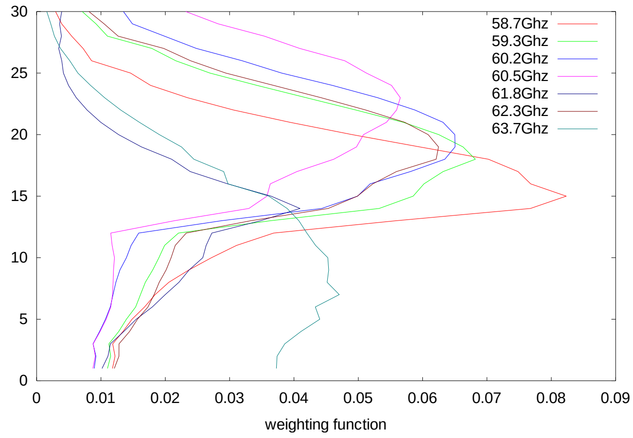


Figure 2 Weighting functions for observation frequencies, 58.7, 59.3, 60.2, 60.5, 61.8, 62.3, 63.7 GHz

A can be determined from equation (2).

$$\begin{matrix} B(T_{\lambda_1},\lambda_1)\times K_{\lambda_1} & \cdots & B(T_{\lambda_7},\lambda_1)\times K_{\lambda_1} \\ \vdots & \ddots & \vdots \\ B(T_{\lambda_1},\lambda_7)\times K_{\lambda_7} & \cdots & B(T_{\lambda_7},\lambda_7)\times K_{\lambda_7} \end{matrix} \qquad (2)$$

where B, T_λ, λ , K_λ denotes Plank function, air temperature at the peak of weighting function, frequency, and weighting

[9] http://www.lmd.jussieu.fr/~falmd/TP/results_interpret_AMSU/AMSU.pdf

[10] http://www.sat.ltu.se/workshops/radiative_transfer/minutes.php

[11] http://en.wikipedia.org/wiki/Atmospheric_radiative_transfer_codes

[12] http://www.nict.go.jp/

[13] http://www.arm.gov/publications/proceedings/conf05/extended_abs/mlawer_ej.pdf

[14] http://andrewgelman.com/2011/03/prior_informati/

[15] http://andrewgelman.com/2011/03/prior_informati/

function at the peak altitude, respectively. On the other hand, G_0 can be calculated with equation (3).

$$\sum_{h=1}^{H} B(T_h,\lambda_1)\times K(\lambda_1,h)$$
$$\vdots$$
$$\sum_{h=1}^{H} B(T_h,\lambda_7)\times K(\lambda_7,h) \qquad (3)$$

where h, H, T_h denotes altitude, peak altitude at which weighting function is maximum, and air temperature at altitude.

C. Inverse Problem Solbing Based Mtheod with Microwave Sounder Data

As aforementioned, A can be calculated in advance for air temperature profile retrievals. A is square matrix. Therefore, it is easy to calculate inverse matrix of A. Using inverse matrix A, air temperature profile can be retrieved as follows,

$$T=T_0+A^{-1}(G-G_0) \qquad (4)$$

where T_0, G, G_0 denotes air temperature at the designated altitude, brightness temperature derived from the acquired AMSU data, and model derived brightness temperature, respectively. This method is referred to Inverse Matrix Method: IMM hereafter.Fig.3 shows the weighting functions for assumed observation frequencies, 52.8, 55.5, and 57.29 GHz, respectively.

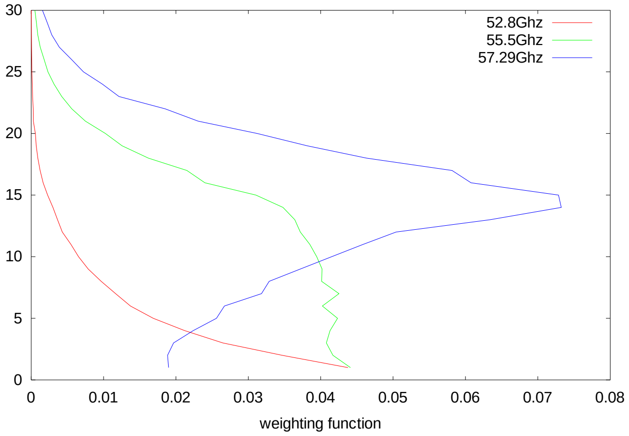


Figure 3 Weighting functions for the designated observation frequencies of 52.8, 55.5, and 57.29 GHz

III. Experiments

A. Error Analysis on Air Temperature Profile Retrieval Accuracy for the Conventional Error Covariance Based Method

Brightness temperature at the designated observation frequency can be calculated with MAES (Mid. Latitude Summer of atmospheric model). One of the input parameters is air temperature profile. Therefore, error analysis is made through the following procedure,

(1) Designate air temperature profile

(2) Calculate observed brightness temperature at the designated observation frequencies

(3) Estimate air temperature profile based on the conventional error covariance based method

(4) Compare the designated and estimated air temperature profiles at the altitudes at which weighting function is maximum (peak weighting function altitude)

Table 1 shows estimated air temperature derived from the conventional covariance matrix based method and truth air temperature as well as estimation error. Table 1 (a) shows those for 1K of additive noise while Table 1 (b) shows those for 3K of additive noise. On the other hand, Table 1 (c) shows those for 5K of additive noise. 1, 3, 5K of noises are added to the observed brightness temperature of AMSU data.

TABLE I. Air Temperature Profile Estimation Accuracy for the Conventional Error Covariance Based Method

(a)Additive Noise = 1K

Altitude(km)	Estimated	Truth	Error
7	256.356	254.7	1.658
14	217.713	215.7	2.031
15	217.876	215.7	2.176
18	219.529	216.8	2.729
19	219.691	217.9	1.791
20	220.712	219.2	1.512
23	224.517	222.8	1.717

(b)Additive Noise=3K

Altitude(km)	Estimated	Truth	Error
7	258.391	254.7	3.691
14	219.93	215.7	4.23
15	219.483	215.7	3.783
18	220.787	216.8	3.987
19	221.762	217.9	3.862
20	223.24	219.2	4.04
23	226.808	222.8	4.008

(c)Additive Noise=5K

Altitude(km)	Estimated	Truth	Error
7	260.309	254.7	6.609
14	220.009	215.7	4.309
15	221.181	215.7	5.481
18	223.253	216.8	6.453
19	223.553	217.9	5.653
20	227.823	219.2	8.612
23	227.258	222.8	4.458

Trend of the estimation error against additive noise shows exponential function as shown in Fig.4. The estimation error at additive noise is zero (without any observation noise is added to brightness temperature) ranges from 1.2 to 2.5 K. It is a reasonable accuracy of air temperature profile.

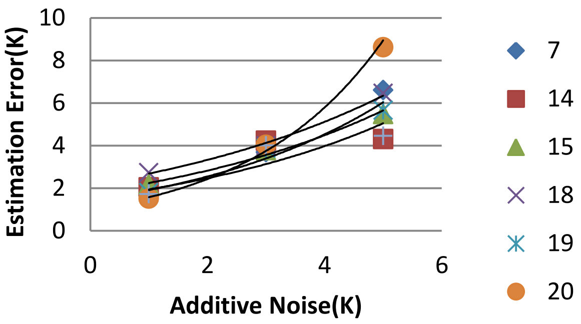


Figure 4 Estimation error trend of air temperature profile as a function of additive noise.

B. AMSR Data Used

The proposed method which minimizing the difference between model derived and the actual microwave sounder data derived air temperature is validated with AMSU data of suburban of London (Longitude: 0 degree West, Latitude: 51.3 North) which is acquired on July 8 2004.

Fig.5 (a), (b), (c) shows brightness temperature of the AMSU Channel 4, 8, and 9, respectively.

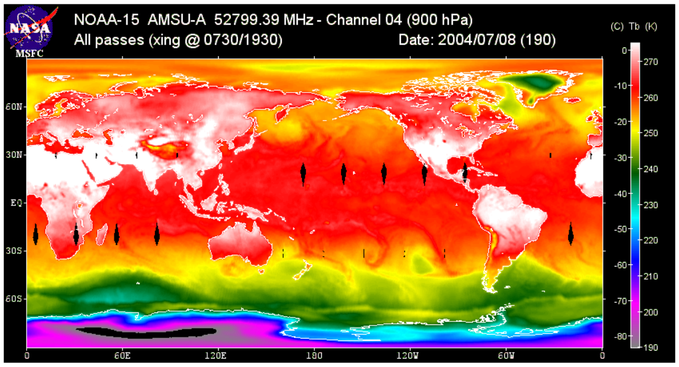
The brightness temperature at the test location for the designated three frequency bands are as follows,

52.8GHz (247.2 K),

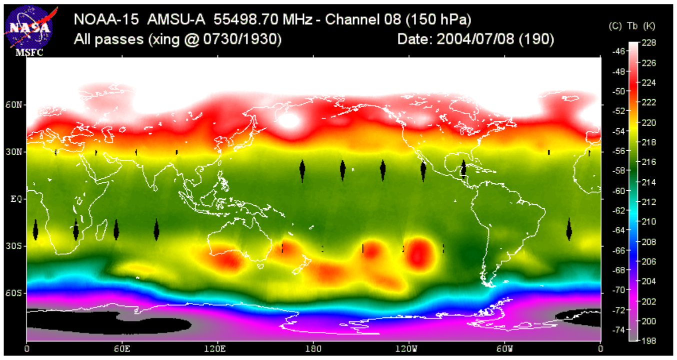
55.5GHz (213.3K), and

57.29GHz (210.6K)

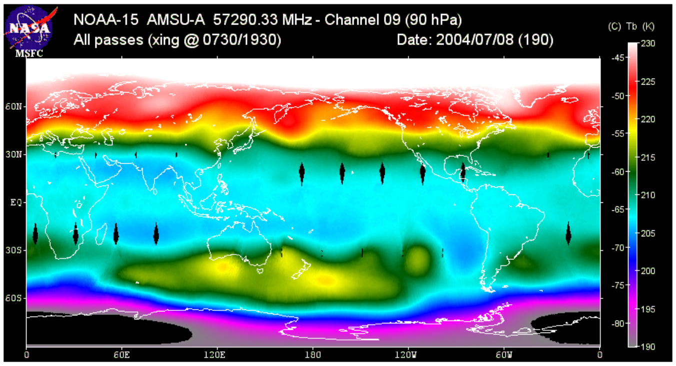
Assuming Mid. Latitude Summer of atmospheric model, brightness temperature of these three observation frequency bands is estimated.



(a)Channel 4 which corresponds to 900hPa



(b)Channel 8 which corresponds to 150 hPa



(c)Channel 9 which corresponds to 90 hPa

Figure 5 AMSU data used

C. Air TemperatureEstimation Accuracy

Using these brightness temperature, air temperature profile is estimated with the proposed method. Fig.6 and Table 2 shows the estimated and model derived air temperature profiles.

The estimation error at the altitudes of 7 and 14 km are common to the conventional method and the proposed method. Therefore, the averaged estimation error at altitude of 7 and 14 km are compared. The result is shown in Table 3.

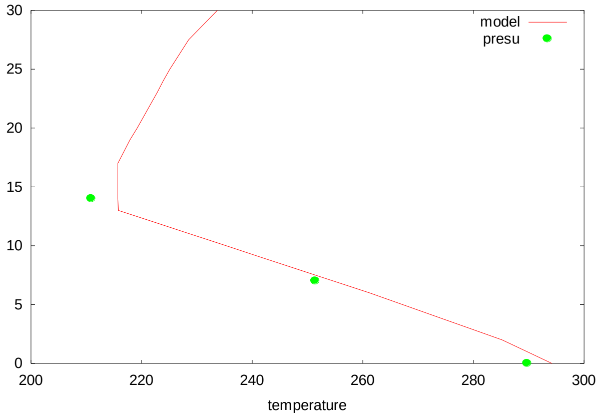


Figure 4 Model derived and the estimated air temperature profiles

TABLE II. AIR TEMPERATURE PROFILE ESTIMATION ACCURACY FOR THE PROPOSED INVERSE MATRIX BAED METHOD

Altitude(km)	Estimated	Truth	Error
0	289.745	294.2	4.456
7	251.429	254.7	3.271
14	210.913	215.7	4.787

TABLE III. AVERAGE AIR TEMPERATURE ESTIMATION ERROR BETWEEN ERROR AT THE ALTITUDE OF 7 AND 14KM FOR BOTH OF THE CONVENTIONAL AND THE PROPOSED MTHEODS

Additive Noise	1K	3K	5K
Conventional Method	1.845	3.961	5.459
Proposed Method	4.029		

Even though, the estimation error of the proposed method do not take into account any additive noise, the estimation error is corresponding to the error of the conventional method with 3K of additional noise. Although the proposed method is not so accurate retrieval method for air temperature profile, it is quit fast and does not required huge computer resources because only thing we have to do is to calculate inverse matrix of *A*. It is 10 times faster than the conventional method.

IV. CONCLUSION

Error analysis of air temperature profile retrievals with microwave sounder data based on minimization of covariance matrix of estimation error is conducted. Additive noise is taken into account in the observation data with microwave sounder onboard satellite. Method for air temperature profile retrievals based on minimization of difference of brightness temperature between model driven microwave sounder data and actual microwave sounder data is also proposed.

The experimental results shows reasonable air temperature retrieval accuracy can be achieved by the proposed method. The air temperature estimation error of the proposed Inverse Matrix Based Method is around 4K and is corresponding to that of the conventional method with 3K of observation noise. Also it is found that air temperature estimation error of the conventional error covariance based method ranges from 1.2 to 2.5K and is getting large exponentially in accordance with increasing of observation noise.

ACKNOWLEDGMENT *(HEADING 5)*

The author would like to thank Mr. Taizo Nakamura for his effort to experimental study.

REFERENCES

[1] Kohei Arai, Lecture Notes on Remote Sensing, Morikita Publishing Inc., 2004

[2] Kohei Arai and XingMing Liang, sensitivity analysis for air temperature profile estimation method around the tropopause using simulated AQUA/AIRS data, Advances in Space Research, 43, 3, 845-851, 2009.

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science, and Technology of the University of Tokyo from 1974 to 1978 also was with National Space Development Agency of Japan (current JAXA) from 1979 to 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He was appointed professor at Department of Information Science, Saga University in 1990. He was appointed councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was also appointed councilor of Saga University from 2002 and 2003 followed by an executive councilor of the Remote Sensing Society of Japan for 2003 to 2005. He is an adjunct professor of University of Arizona, USA since 1998. He also was appointed vice chairman of the Commission "A" of ICSU/COSPAR in 2008. He wrote 30 books and published 332 journal papers

Fast DC Mode Prediction Scheme For Intra 4x4 Block In H.264/AVC Video Coding Standard

Tajdid Ul Alam[1]
Department of Electronics and Telecommunication Engineering
Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh

Jafor Ikbal[2]
Department of Electronics and Telecommunication
Engineering Rajshahi University of Engineering and
Technology, Rajshahi, Bangladesh

Touhid Ul Alam[3]
Department of Computer Science and Engineering,
Ahsanullah University of Science and
Technology, Dhaka, Bangladesh

***Abstract*— In this paper, the researchers proposed a new scheme for DC mode prediction in intra frame for 4X4 block. In this scheme, the upper and left neighboring pixels would be used to predict each of the pixels in the 4X4 block with different weight. The prediction equations for each position of the 4X4 block in DC mode would be static with fixed weighting coefficients. As a result, the computational time for intra frame would be decreased considerably with an increase in PSNR.**

Keywords- Intra; macroblock (MB); DC mode; luma; prediction.

I. Introduction

H.264 or MPEG-4 Part 10 Advanced Video Coding (AVC) is a joint venture of video coding experts from both the Joint Video Team (JVT) of ITU-T and Motion Picture Experts Group (MPEG) of ISO [1].

H.264/AVC is an important format, which is most commonly used in recording, compression and distribution of high definition video. It was first released in May, 2003.

After its release, it has been found that it has very broad application that covers all forms of digital compressed video from low bit-rate internet streaming applications to HDTV broadcast and digital cinema applications where nearly lossless coding is very important [2].

A number of new features have been added to the H.264/AVC. Some of these features are multiple reference frames, variable block size motion estimation, multiple motion vectors per macroblock, weighted prediction, spatial prediction from the edges of neighboring block for intra coding rather than DC only etc. These features allow H.264/AVC to compress video more efficiently but complexity of implementation also increases.

In H.264/AVC standard, spatial redundancy is removed in intra frame by using 9 modes [3]. DC mode is one of the modes that are used for encoding intra 4X4 block.

In DC mode, the luminance values of entire pixels of current 4X4 block are predicted by computing average of the luminance values of pixels from upper and left neighbors. For 4X4 block, there are 8 such pixels in total.

We observed that most of the variations of luminance values in intra 4X4 block are removed by the standard DC mode as the luminance value of 16 pixels are predicted by a single value, which is calculated by averaging the 8 pixels. Therefore, we are enthusiastic to propose a new pattern for DC mode which will preserve the variations of luminance values. We call this pattern as “ZIG-ZAG DC mode prediction”.

A lot of efforts have been made to improve the intra prediction scheme of H.264/AVC. Many researches have been done to reduce computational complexity. But it has been witnessed that in most of the cases the researchers have less contribution in improving coding efficiency.

This ZIG-ZAG DC mode prediction can improve coding efficiency by preserving more details and can reduces computational time with significant increase in PSNR.

II. Intra Prediction Scheme In H.264/Avc Standard

Prediction for intra macroblocks (MB) is called intra prediction. Intra prediction is used to remove spatial redundancy of intra frames. Intra frames are encoded without any reference of other frames. Intra prediction is performed in pixel domain. It uses only transform coding and the neighboring blocks of the same frame to predict block values. Intra prediction is performed on 16X16 luma and 4X4 luma blocks.

No intra prediction is performed on 8X8 luma blocks [4]. In this standard, intra prediction forms predictions of pixel values as linear interpolations of pixels from the adjacent edges of neighboring MBs that are decoded before the current MB that is to be encoded. The interpolations are directional in nature, with multiple modes, each implying a spatial direction of prediction [3].

A 4X4 luma block contains samples as shown in figure 1 (a). There are nine intra prediction modes as shown in figure 1 (b). The modes are given in table 1.

TABLE 1: INTRA PREDICTION MODES FOR 4X4 LUMA BLOCK.

Nu mber	Intra 4X4 prediction mode
0	Vertical
1	Horizontal
2	DC
3	Diagonal down left
4	Diagonal down right
5	Vertical right
6	Horizontal down
7	Vertical left
8	Horizontal up

Modes 0, 1, 3, 4, 5, 6, 7, 8 are directional prediction modes as indicated in figure 1 (b) and mode 2 is the DC prediction mode with no direction. The prediction block is calculated from the samples A-M as shown in figure 1 (a). In figure 1 (b), the arrows indicate the direction of prediction in each mode.

While predicting a sample in the current 4X4 block, the neighboring samples of upper side and left side have previously been encoded and reconstructed. The neighboring samples are available to the encoder and decoder to form prediction references.

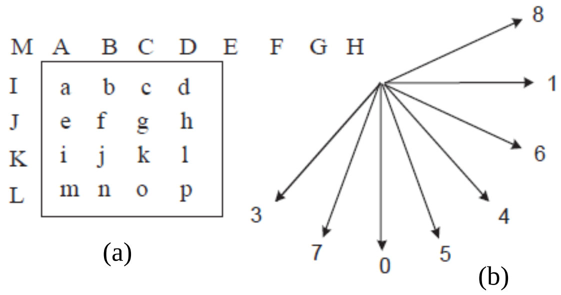


Figure 1: (a) Identification of samples used for intra spatıaı prediction, (b) intra prediction directions.

I. Standard Algorithm For Dc Mode Prediction

The standard form of DC mode prediction for 4X4 block is to replace all pixels in the current 4X4 block by the mean value of the neighboring pixels. The pixels in the current 4X4 block that are to be predicted are shown in figure 1(a) by using a box. The reference pixels are A, B, C, D, I, J, K and L. These are also shown in the same figure. The DC prediction is accomplished by using the following rules [4]:

1. If all samples A, B, C, D, I, J, K, L are available, all samples are predicted by (A+B+C+D+I+J+K+L+4)>>3
2. If A, B, C, D are not available and I, J, K, L are available, all samples shall be predicted by (I+J+K+L+2)>>2
3. If I, J, K, L are not available and A, B, C, D are available, all samples shall be predicted by (A+B+C+D+2)>>2
4. If all eight samples are unavailable, the prediction for all luma samples in the 4X4 block shall be 128.

A block therefore can always be predicted in DC mode. A typical case for luma prediction using DC mode in 4X4 block is shown in figure 2 [5].

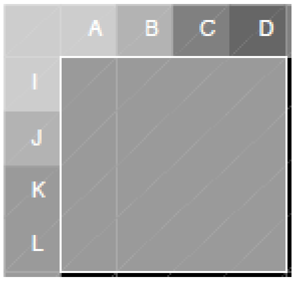


Figure 2: A typical case of DC Mode prediction.

Figure 2 shows that all pixels in the current 4X4 block are predicted by an average of their neighboring pixels. So, most of the variations of luminance value are removed by using DC prediction mode.

III. Proposed Algorithm For Dc Mode Prediction

The DC mode prediction scheme of pixel values in current 4X4 block can be improved by using ZIG-ZAG DC mode prediction. In ZIG-ZAG DC mode prediction, the 16 pixels in current 4X4 block are predicted one by one using a ZIG-ZAG order. The prediction sequence of the pixels in current 4X4 block is indicated in figure 3 using arrows. The reference pixels are A, B, C, D which belong to the upper neighboring block and I, J, K, L which belong to the left neighboring block. Besides these reference pixels, we also used previously predicted pixels of the current 4X4 block to predict another pixel belonging to the same block. But the priority of pixels from upper and left neighboring block was high and priority of pixels belonging to the same block was low. This was done by assigning different weighting coefficient to different pixels.

For example, we use predicted luminance values of pixels in the positions of i, e, f and g to predict the luminance value of pixel in the position of j. As shown in the figure, i, e, f and g are previously predicted pixels positions. Similarly, luminance value of a pixel in the position of c is predicted from the predicted luminance value of pixels in the positions of f, b, B, C and D.

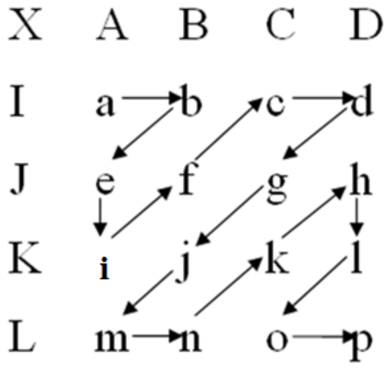


Figure 3: ZIG-ZAG DC mode prediction

Each pixel in current 4X4 block was predicted using different prediction equations. Each equation contains the luminance value of the pixels that are used to predict the luminance value of the current pixel in 4X4 block. The weight assigned to each of the reference pixels was different and it also depended on the pixel's position in current 4X4 block. We assigned higher weight to the upper and left neighboring pixels than the pixels belonging to the current 4X4 block.

The equations that are used for the prediction of luminance value of each pixel in the 4X4 block in ZIG-ZAG DC mode are given as follows:

$$a = 0.25A + 0.125B + 0.25X + 0.25I + 0.125J$$

$$b = 0.25A + 0.25B + 0.0833X + 0.0833I + 0.0833J + 0.25C$$

$$e = 0.05A + 0.05B + 0.3K + 0.3I + 0.3J$$

$$i = 0.06A + 0.06B + 0.06X + 0.2I + 0.2J + 0.2K + 0.2L$$

$$f = \frac{a + b + e + i}{4}$$

$$c = \frac{(B + C + D) + b + f}{3}$$

$$d = \frac{c + (C + D)}{2}$$

$$g = \frac{b + c + d + f}{4}$$

$$j = \frac{e + f + g + i}{4}$$

$$m = \frac{i + j + (K + L)}{3}$$

$$n = \frac{i + j + k}{3}$$

$$k = \frac{f + g + j + n}{4}$$

$$h = \frac{k + j + c + d}{4}$$

$$l = \frac{g + h + k}{3}$$

$$o = \frac{j + k + l + n}{4}$$

$$p = \frac{o + k + l}{3}$$

We can see from the equations that the pixels in current 4X4 block are predicted using both the previously encoded pixels of upper and left neighbor and previously predicted pixels of the same block with different weighting coefficients.

From these equations we can also tell that, the luminance values of each position of the current 4X4 block are not same. So, the proposed scheme preserves more details than the previous scheme which leads to an increase in PSNR.

IV. Simulation Result

The proposed algorithm was coded and simulated in JM software version 18.2 [http://iphome.hhi.de/suehring/tml/]. Using the reference software, first we encoded the 10 sequences with the standard DC mode prediction algorithm and collected the PSNR(Y), computational time and bit rate. Then we coded and simulated the ZIG-ZAG DC mode prediction algorithm in JM software version 18.2 and then encoded the same sequences with the proposed ZIG-ZAG DC mode prediction algorithm and collected the PSNR(Y), computational time and bit rate of those sequences.

We encoded three frames for each sequences and all sequences were encoded in QCIF resolution. We compared the PSNR(Y), computational time and bit rate of the encoded sequences for both the standard DC mode prediction algorithm and ZIG-ZAG DC mode prediction algorithm. We observed the improvement of PSNR(Y) and computational time for most of the sequences where the bit rates for the sequences were almost the same.

The difference in PSNR(Y), computation time and bit rate between standard algorithm and proposed ZIG-ZAG DC mode prediction algorithm was calculated. The machine was Intel® Core™ i5-2430M @ 2.40GHz with 4.00GB RAM. Operating system was 64 bit OS. The condition for testing of standard algorithm and proposed algorithm were same. Observed change in results was tabulated as follows:

TABLE 2: SIMULATION RESULTS OF VARIOUS SEQUENCES.

Sequence	ΔPSNR (dB)	Δ bit rate (kbps)	Δ time (ms)
Miss America	0.048	1.44	0.361
Bus	-0.009	-20.16	-14.725
Foreman	1.588	-0.12	-12.372
Mobile	-0.095	9.96	-12.53
Mother	2.679	2.76	0.057
Highway	0.066	2.28	-8.594
Coastguard	0.119	8.28	-0.777
Grandma	0.059	1.68	-0.008
Salesman	0.417	-3.60	-68.291
Paris	0	0	-9.627
Average	0.4872	0.252	-12.651

In table 2, we calculated the differences of the PSNR(Y), computational time and bit rate of the standard algorithm from the ZIG-ZAG DC mode prediction algorithm.

As we can see from table 2, the PSNR(Y) increases in most of the cases. For the sequence "Paris", the PSNR(Y) was same. For the sequences "Bus" and "Mobile", the decrease in PSNR(Y) for ZIG-ZAG DC mode prediction algorithm is almost negligible.

Also we observe that, the decrease in computational time for the ZIG-ZAG DC mode prediction algorithm is noticeable for most of the sequences. For the sequences “Miss America” and “Mother-daughter” the difference in computational time is almost the same.

We calculated the average of the change in PSNR(Y), computational time and bit rate. From the average value, we observed that, increase in PSNR(Y) and computational time were satisfactory. The change in bit rate was almost negligible.

Our target was to increase the PSNR(Y) value and decrease the computational time. As we observed that the standard algorithm for DC mode prediction looks for several conditions for predicting each 4X4 block, the computation time for standard DC mode prediction algorithm is higher. The conditions are whether the upper samples are available or not, whether the left samples are available or not, whether the upper and left samples are available together, whether there is no sample available.

These four steps consume most of the time for predicting the pixel values in DC mode. As in our proposed ZIG-ZAG DC mode prediction algorithm, there is no such condition to check for, the computational time decreases than the standard DC mode prediction algorithm.

Again in standard DC mode prediction algorithm, all pixels in current 4X4 block are predicted by a single value as discussed earlier. But in our proposed ZIG-ZAG DC mode prediction algorithm, each pixel is predicted separately and variation in luminance values are preserved more in current 4X4 block. As a result for most of the cases, the PSNR(Y) increases.

Figure 4 to 6 shows the graph of PSNR(Y), computation time and bit rate values respectively. Each graph contains values for 10 sequences. The result for standard algorithm and proposed algorithm are shown side by side on same graph.

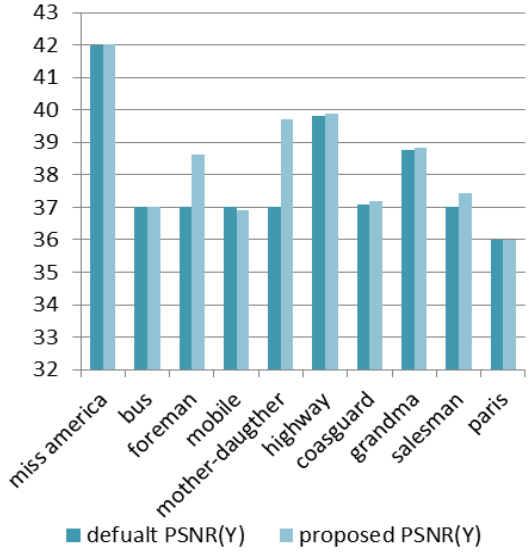
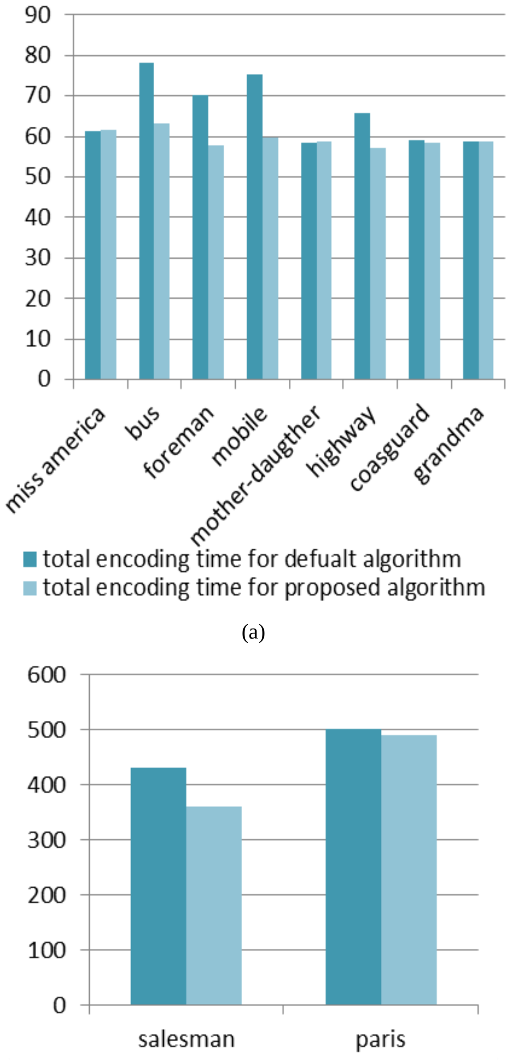


Figure 4: Comparison of PSNR(Y) of standard and ZIG-ZAG DC mode prediction algorithm.



(a)

(b)

Figure 5 (a) & (b): Comparison of total encoding time of standard and ZIG-ZAG DC mode prediction algorithm.

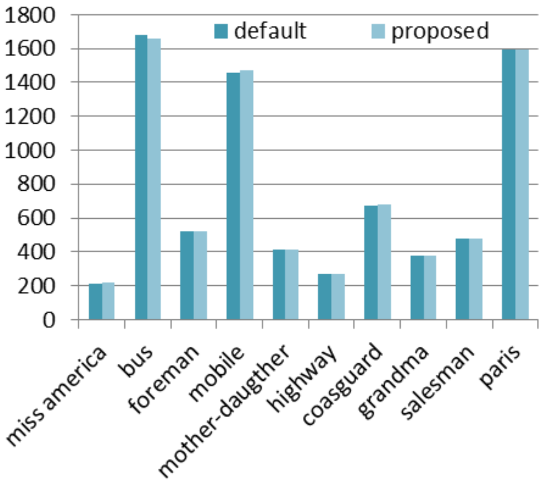


Figure 6: Comparison of bit rate (kbps) of standard and ZIG-ZAG DC mode prediction algorithm.

In standard DC mode, due to the replacement of all the pixels with a single value, most of the variations in luminance were omitted. But with our proposed ZIG-ZAG DC mode prediction algorithm, this factor is improved and as a result the PSNR increases. The decrease in processing time with the ZIG-ZAG DC mode prediction algorithm is due to the absence of conditions as involved with the standard DC mode prediction algorithm.

Actually in ZIG-ZAG DC mode prediction algorithm, the concepts of all the eight modes of prediction are combined. As a result, the prediction accuracy is improved.

V. CONCLUSION

From the observation and comparison of simulation results, we see that the computation time decrease which makes the ZIG-ZAG DC mode prediction algorithm more efficient for live video broadcasting application as well as teleconferencing application where lowering computation time adds advantage. Also increase in PSNR improves signal quality for the video signal.

VI. FUTURE WORK

In future, we will try to further decrease the computation time for intra frame in H.264/AVC standard. We can do this by observing the relative usage of each of the nine modes in intra frame and discarding the less used modes for prediction of intra frame. If the changes in PSNR(Y) and bit rate were not worse, then the reduction in prediction mode in intra frame would reduce the computation time.

REFERENCES

[1] Atul Puri, Xuemin Chen, Ajay Luthra, “Video coding using the H.264/MPEG-4 AVC compression standard”, Elsevier: Signal Processing: Image Communication 19 (2004) 793–849

[2] http://en.wikipedia.org/wiki/H.264/MPEG-4_AVC

[3] Soon-kak Kwon, A. Tamhankar, K.R. Rao, “Overview of H.264/MPEG-4 part 10”, science direct, J. Vis. Commun. Image R. 17 (2006) 186–216

[4] A. Tamhankar and K. R. Rao, “AN OVERVIEW OF H.264 / MPEG4 PART 10”, EC-VIP-MC 2003,4th EURASIP Conference focused on Video / Image Processing and Multimedia Communications, 2-5 July 2003, Zagreb. Croatia

[5] Gary J. Sullivan, Pankaj Topiwala, and Ajay Luthra, “The H.264/AVC Advanced Video Coding Standard:Overview and Introduction to the Fidelity Range Extensions”, Gary J. Sullivan, Pankaj Topiwala, and Ajay Luthra, Presented at the SPIE Conference on Applications of Digital Image Processing XXVII, Special Session on Advances in the New Emerging Standard: H.264/AVC, August, 2004

[6] JM Software: http://iphome.hhi.de/suehring/tml/

[7] Iain Richardson, “White Paper: H.264 / AVC Intra Prediction”, Vcodex, © 2002-2011

[8] Thomas Wiegand, Gary J. Sullivan, Gisle Bjøntegaard, and Ajay Luthra, “Overview of the H.264/AVC Video Coding Standard”, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 13, NO. 7, JULY 2003

[9] Yung-Lyul Lee, Ki-Hun Han, Dong-Gyu Sim, and Jeongil Seo, "Adaptive Scanning for H.264/AVC Intra Coding",aETRI Journal, Volume 28, Number 5, October 2006

[10] ITU-T Recommendation H.264 and ISO/IEC 14496-10, “Advanced video coding for generic audiovisual services,” May 2003 (and subsequent amendment and corrigenda).

[11] Wes Simpson, "Video Over IP IPTV, Internet Video, H.264, P2P, Web TV, and Streaming: A Complete Guide to Understanding the Technology",Second Edition, Published by Elsevier Inc.

[12] L. Hanzo, P. J. Cherriman and J. Streit, "Video Compression and Communications From Basics to H.261, H.263, H.264, MPEG4 for DVB and HSDPA-Style Adaptive Turbo-Transceivers, Second Edition, John Wiley & Sons, Ltd

[13] Mohammed Ebrahim Al-Mualla, C. Nishan Canagarajah and David R. Bull, "Video Coding for Mobile Communications Efficiency Complexity and Resilience", ACADEMIC PRESS

[14] John Watkinson, "The MPEG Handbook—MPEG-1, MPEG-2, MPEG-4", Focal Press

[15] Atul Puri, Tsuhan Chen, "Multimedia Systems, Standards, and Networks", MARCEL DEKKER, INC.

Free Open Source Software: FOSS Based GIS for Spatial Retrievals of Appropriate Locations for Ocean Energy Utilizing Electric Power Generation Plants

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

Abstract— **Free Open Source Software: FOSS based Geographic Information System: GIS for spatial retrievals of appropriate locations for ocean wind and tidal motion utilizing electric power generation plants is proposed. Using scatterometer onboard earth observation satellites, strong wind coastal areas are retrieved with FOSS/GIS of PostgreSQL/GIS. PostGIS has to be modified together with altimeter and scatterometer database. These modification and database creation would be a good reference to the users who would like to create GIS system together with database with FOSS.**

Keywords- free open source software; postgres SQL; GIS; spatial retrieval.

I. Introduction

Geographic Information System: GIS is used for exploration and spatial retrieval of appropriate locations and areas. GIS software is widely available now a day [1]. GIS can be created with Free Open Source Software: FOSS [2]. Functionalities of GIS is as follows,

・Display superimposing the thematic maps and imagery data

・Spatial and temporal retrieval of the maps and data

・Quantitative analysis (length, area, etc.)

・Simulation (assessment, 3D scenery analysis, etc.)

There are the following issues should be discussed,

・Difficulty on customization of the GIS to specific applications

・Expensive system and database updating cost

・Reverse retrievals for the spatial and temporal search for confirmation of the original data

・Apply image processing to the retrieved data

In order to overcome the aforementioned problems of FOSS [1] based GIS, example of customization of PostgreSQL [2]/GIS (PostGIS [3]) for the specific purpose of spatial retrieval of ocean energy[4], or marine energy[5], ocean wind and tidal motion utilizing power generation plant locations are attempted.

In order for that, retrievals of the appropriate ocean areas in the Japanese vicinity for exploration of ocean related energy sources, geoid, ocean winds, wave heights and tidal effects are required [3]. The data for the aforementioned energy sources are available from satellite based radar altimeter [6] and scatterometer[7]. Then exploration of possible areas for ocean related power generations is followed by [4], [5]. Also, create the database containing geoid[8], tides, ocean winds, wave height and so on from the NASA/JPL PODAAC [9] (Topex/Poseidon[10] and Jason satellites[11] data) by extracting the geo-referenced and time stamped data from the PODAAC has to be done. After that, access to the database through php[12] and Mapscript [13] then display the retrieval results of the appropriate ocean areas for the ocean energy exploration on the php web browser. These procedures are demonstrated in this paper. Also the GIS is used as Neural Network [6], [7].

The following section describes the proposed PostGIS followed by the data descriptions required for spatial retrievals of appropriate locations for electric power generation plants utilizing ocean energy. Then demonstration is followed by. Finally, conclusion and some discussions are followed.

II. Proposed FOSS/GIS

A. Availability of FOSS/GIS

There are not so small numbers of FOSS/GIS systems which are available and downloadable from their web sites. Table 1 shows just a small portion of available FOSS/GIS. As for the well-known Grass of GIS, it is easy to install it on your computer through the following procedure,

[1] http://e-words.jp/w/FOSS.html
[2] http://www.postgresql.org/
[3] http://postgis.refractions.net/

[4] http://www.renewableenergyworld.com/rea/tech/ocean-energy
[5] http://en.wikipedia.org/wiki/Marine_energy
[6] http://en.wikipedia.org/wiki/Radar_altimeter
[7] http://en.wikipedia.org/wiki/Scatterometer
[8] http://en.wikipedia.org/wiki/Geoid
[9] http://podaac.jpl.nasa.gov/
[10] http://sealevel.jpl.nasa.gov/
[11] http://ilrs.gsfc.nasa.gov/satellite_missions/list_of_satellites/jas2_general.html
[12] http://ja.wikipedia.org/wiki/PHP:_Hypertext_Preprocessor
[13] http://mapserver.org/mapscript/index.html

Link to Grass of GIS software

Installation of GRASSLink[14]s

GRASSLinks is web interface for GRASS GIS of PDS installation

Before using GRASSLinks, GRASS GIS has to be installed

Information on Installation of GRASS is available from the http://www.media.osaka-cu.ac.jp/~raghavan/grassinfo/.

TABLE I. EXAMPLES OF AVAILABLE FOSS/GIS

Tool	Category	Functionality	Remarks
MapServer[15]	Web Mapping engine	Thematic and the other maps generation and services	Useful tool for map services
PostGIS	RDBMS middleware extension	Space retrievals extending data types to the PostgreSQL	Useful tool for geological retrieval services
Grass[16]Ver.6	Client based GIS software	Geological contents management	Useful tool for register and edition of the contents

As for the well-known Mapserver, it is easy to install it on your computer through the following procedure,

MapServer international version (i18n) (i18n Version of Mapserver: Package)

MapServer 4.0.1 source code and patch for the international use

As for the well known PostgreSQL/GIS, PostGIS, it is easy to install it on your computer through the following procedure,

PostGIS allows store the objects in concern to the GIS (Geographic Information Systems) database

PostgtreSQL extension of PostGIS supports fundamental functions for analysis of GIS objects and spatial R-Tree index of the GiST base

PostgreSQL can be downloaded from the http://www.postgresql.org/

PostGIS is source code tree of the PostgreSQL and can be installed by using the definition of installation process of the PostgreSQL

PostGIS can be compiled with GNU C[17], gcc and/or ANSI C[18] complier

GNU Make, gmake and/or make can be used for making the PostGIS. GNU make is the default version of make. Version can be confirmed with “make –v”. Make file of PostGIS will not be processed properly when the different version of make is used

Proj4 is the library of the map projection conversion tools as one of the options of the PostGIS. Proj4 is available from the http://www.remotesensing.org/proj

In order to utilize Mapserver, the following procedure is required,

Minnesota Mapserver is the internet Web mapping server and is compatible to the mapping server specification

Mapserver is available from the http://mapserver.gis.umn.edu/

Web Map specification of OpenGIS is available from the http://www.opengis.org/techno/specs/01-047r2.pdf

B. FOSS of GIS

The required systems are as follows,

PostgreSQL

FOSS of relational database system

SQL：Structured Query Language

PostGIS

GIS extension of PostgreSQL

Good interface to the GIS database

MapServer

Web mapping engine

Database access with php and MapScript

MapServer(php/MapScript)

php

Interface to database with php

Retrievals are then available through php Web page

Submit queries then the retrieved results are displayed from the database table

MapScript

Map engine allows displaying the retrieved results superimposing the other existing thematic maps

Multiple layers

Raster and vector data of maps, meshed data and images through the php web browser

C. Example of Database Creation

Conversion of binary data to GIS database is required together with analysis program. After that, Modification of the analysis program is required followed by extraction of the required data. Conversion of the extracted data to GIS database (database table can be created with PostgreSQL) is

[14] http://ippc2.orst.edu/glinks/

[15] http://mapserver.org/

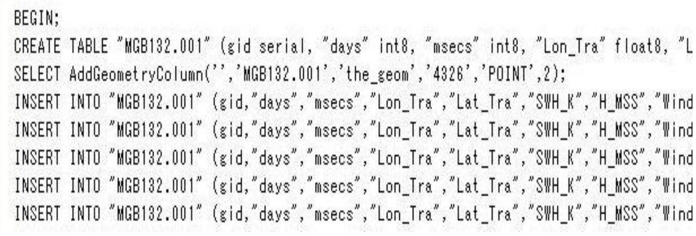
[16] http://grass.fbk.eu/

[17] http://gcc.gnu.org/

[18] http://ja.wikipedia.org/wiki/C%E8%A8%80%E8%AA%9E

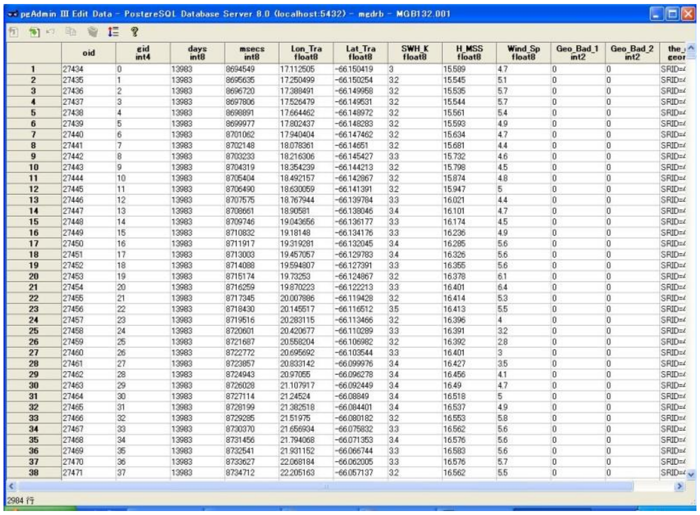
also required. Table 2 shows example of the database in PostGIS. Database is described with table style.

TABLE II. EXAMPLES OF DATABASE IN POSTGIS



```
BEGIN;
CREATE TABLE "MGB132.001" (gid serial, "days" int8, "msecs" int8, "Lon_Tra" float8, "L
SELECT AddGeometryColumn('','MGB132.001','the_geom','4326','POINT',2);
INSERT INTO "MGB132.001" (gid,"days","msecs","Lon_Tra","Lat_Tra","SWH_K","H_MSS","Wind
INSERT INTO "MGB132.001" (gid,"days","msecs","Lon_Tra","Lat_Tra","SWH_K","H_MSS","Wind
INSERT INTO "MGB132.001" (gid,"days","msecs","Lon_Tra","Lat_Tra","SWH_K","H_MSS","Wind
INSERT INTO "MGB132.001" (gid,"days","msecs","Lon_Tra","Lat_Tra","SWH_K","H_MSS","Wind
INSERT INTO "MGB132.001" (gid,"days","msecs","Lon_Tra","Lat_Tra","SWH_K","H_MSS","Wind
```

Editing of the database can be done in accordance with PostgresSQL data manipulation as shown in Fig.1. Fig.2 shows example of PostgresSQL database server.

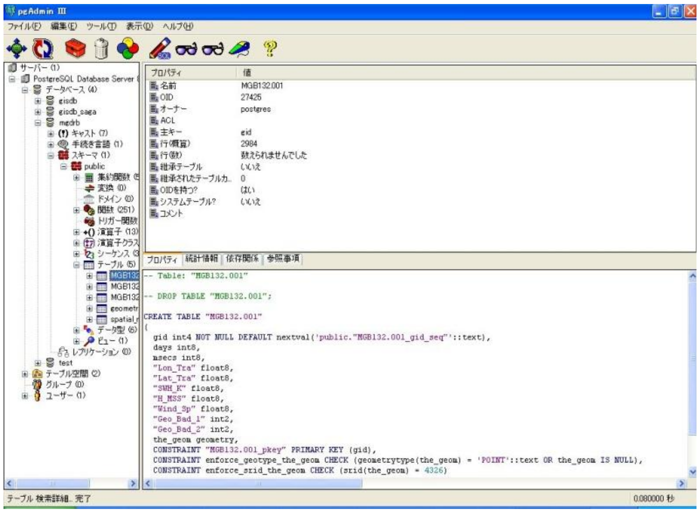


pgAdmin III Edit Data - PostgreSQL Database Server 8.0 (localhost:5432) - mgdrb - MGB132.001

	oid	gid int4	days int8	msecs int8	Lon_Tra float8	Lat_Tra float8	SWH_K float8	H_MSS float8	Wind_Sp float8	Geo_Bad_1 int2	Geo_Bad_2 int2	the_ geor
1	27434	0	13983	8694549	17.112505	-66.150419	3	15.589	4.7	0	0	SRID=4
2	27435	1	13983	8695635	17.250499	-66.150254	3.2	15.545	5.1	0	0	SRID=4
3	27436	2	13983	8696720	17.388491	-66.149958	3.2	15.535	5.7	0	0	SRID=4
4	27437	3	13983	8697806	17.526479	-66.149531	3.2	15.544	5.7	0	0	SRID=4
5	27438	4	13983	8698891	17.664462	-66.148972	3.2	15.561	5.4	0	0	SRID=4
6	27439	5	13983	8699977	17.802437	-66.148283	3.2	15.593	4.9	0	0	SRID=4
7	27440	6	13983	8701062	17.940404	-66.147462	3.2	15.634	4.7	0	0	SRID=4
8	27441	7	13983	8702148	18.078361	-66.14651	3.2	15.681	4.4	0	0	SRID=4
9	27442	8	13983	8703233	18.216306	-66.145427	3.3	15.732	4.6	0	0	SRID=4
10	27443	9	13983	8704319	18.354239	-66.144213	3.2	15.798	4.5	0	0	SRID=4
11	27444	10	13983	8705404	18.492157	-66.142867	3.2	15.874	4.8	0	0	SRID=4
12	27445	11	13983	8706490	18.630059	-66.141391	3.2	15.947	5	0	0	SRID=4
13	27446	12	13983	8707575	18.767944	-66.139784	3.3	16.021	4.4	0	0	SRID=4
14	27447	13	13983	8708661	18.90581	-66.138046	3.4	16.101	4.7	0	0	SRID=4
15	27448	14	13983	8709746	19.043656	-66.136177	3.3	16.174	4.5	0	0	SRID=4
16	27449	15	13983	8710832	19.18148	-66.134176	3.3	16.236	4.9	0	0	SRID=4
17	27450	16	13983	8711917	19.319281	-66.132045	3.4	16.285	5.6	0	0	SRID=4
18	27451	17	13983	8713003	19.457057	-66.129783	3.4	16.326	5.6	0	0	SRID=4
19	27452	18	13983	8714088	19.594807	-66.127391	3.3	16.355	5.6	0	0	SRID=4
20	27453	19	13983	8715174	19.73253	-66.124867	3.2	16.378	6.1	0	0	SRID=4
21	27454	20	13983	8716259	19.870223	-66.122213	3.3	16.401	6.4	0	0	SRID=4
22	27455	21	13983	8717345	20.007886	-66.119428	3.2	16.414	5.3	0	0	SRID=4
23	27456	22	13983	8718430	20.145517	-66.116512	3.5	16.413	5.5	0	0	SRID=4
24	27457	23	13983	8719516	20.283115	-66.113466	3.2	16.396	4	0	0	SRID=4
25	27458	24	13983	8720601	20.420677	-66.110289	3.3	16.391	3.2	0	0	SRID=4
26	27459	25	13983	8721687	20.558204	-66.106982	3.2	16.392	2.8	0	0	SRID=4
27	27460	26	13983	8722772	20.695692	-66.103544	3.3	16.401	3	0	0	SRID=4
28	27461	27	13983	8723857	20.833142	-66.099976	3.4	16.427	3.5	0	0	SRID=4
29	27462	28	13983	8724943	20.97055	-66.096278	3.4	16.456	4.1	0	0	SRID=4
30	27463	29	13983	8726028	21.107917	-66.092449	3.4	16.49	4.7	0	0	SRID=4
31	27464	30	13983	8727114	21.24524	-66.08849	3.4	16.518	5	0	0	SRID=4
32	27465	31	13983	8728199	21.382518	-66.084401	3.4	16.537	4.9	0	0	SRID=4
33	27466	32	13983	8729285	21.51975	-66.080182	3.2	16.553	5.8	0	0	SRID=4
34	27467	33	13983	8730370	21.656934	-66.075832	3.3	16.562	5.6	0	0	SRID=4
35	27468	34	13983	8731456	21.794068	-66.071353	3.4	16.576	5.6	0	0	SRID=4
36	27469	35	13983	8732541	21.931152	-66.066744	3.3	16.583	5.6	0	0	SRID=4
37	27470	36	13983	8733627	22.068184	-66.062005	3.3	16.576	5.7	0	0	SRID=4
38	27471	37	13983	8734712	22.205163	-66.057137	3.2	16.562	5.5	0	0	SRID=4

2984 行

Figure 1 Editing of the database can be done in accordance with PostgresSQL data manipulation



プロパティ	値
名前	MGB132.001
OID	27425
オーナー	postgres
ACL	
主キー	gid
行(概算)	2984
行(数)	数えられませんでした
継承テーブル	いいえ
継承されたテーブルカ..	0
OIDを持つ?	はい
システムテーブル?	いいえ
コメント	

```
-- Table: "MGB132.001"

-- DROP TABLE "MGB132.001";

CREATE TABLE "MGB132.001"
(
  gid int4 NOT NULL DEFAULT nextval('public."MGB132.001_gid_seq"'::text),
  days int8,
  msecs int8,
  "Lon_Tra" float8,
  "Lat_Tra" float8,
  "SWH_K" float8,
  "H_MSS" float8,
  "Wind_Sp" float8,
  "Geo_Bad_1" int2,
  "Geo_Bad_2" int2,
  the_geom geometry,
  CONSTRAINT "MGB132.001_pkey" PRIMARY KEY (gid),
  CONSTRAINT enforce_geotype_the_geom CHECK (geometrytype(the_geom) = 'POINT'::text OR the_geom IS NULL),
  CONSTRAINT enforce_srid_the_geom CHECK (srid(the_geom) = 4326)
```

Figure 2 Example of PostgresSQL database server

D. *The Data Required for Spatial Retrievals of Appropriate Locations for Ocean Energy Utilizing Electric Power Generation Plants*

There are some of required data for finding appropriate locations of ocean energy utilizing electric power generation plants. Namely,

(1) Topex/Poseidon

Topex/Poseidon was launched on Aug. 10 1992. This is the joint mission between U.S.A. and France. Specific features are the followings,

Microwave altimeter

Non sun-synchronous

Inclination: 66°

Global coverage within 10 days

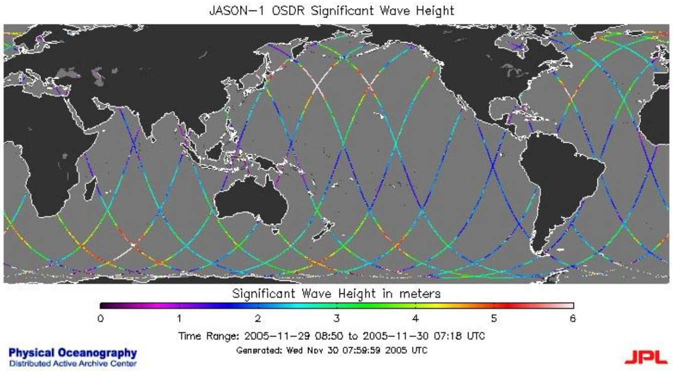

JASON-1 OSDR Significant Wave Height
Significant Wave Height in meters
0
1
2
3
4
5
6
Time Range: 2005-11-29 08:50 to 2005-11-30 07:18 UTC
Generated: Wed Nov 30 07:59:59 2005 UTC
Physical Oceanography
Distributed Active Archive Center
JPL

Figure 3 Topex/Poseidon observes ocean surface along with its orbit

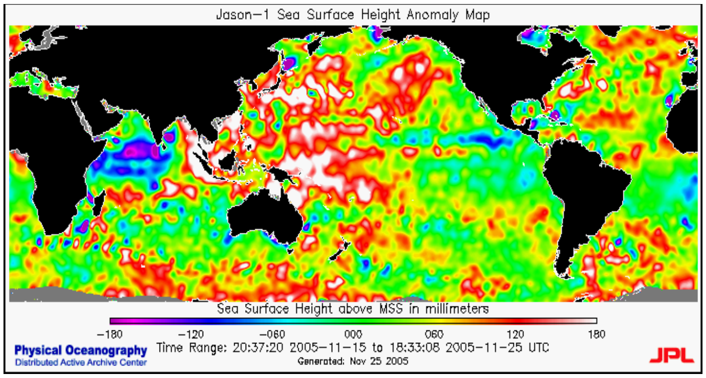

Jason-1 Sea Surface Height Anomaly Map
Sea Surface Height above MSS in millimeters
-180
-120
-060
000
060
120
180
Physical Oceanography
Distributed Active Archive Center
Time Range: 20:37:20 2005-11-15 to 18:33:08 2005-11-25 UTC
Generated: Nov 25 2005
JPL

Figure 4 Geoid potential and wave height is estimated with the altimeter onboard Topex/Poseidon satellite

(2) Scatterometer

Ocean wind direction and speed can be estimated with scatterometer data. One of the scatterometers onboard satellites is SeaWinds [19] on Advanced Earth Observing Satellite: ADEOS-II [20]. Major specification of SeaWinds is shown in Table 3.

[19] http://winds.jpl.nasa.gov/missions/seawinds/

[20] http://en.wikipedia.org/wiki/ADEOS_II

TABLE III. MAJOR SPECIFICATION OF SEAWINDS

Radar:	13.4 gigahertz; 110-watt pulse at 189-hertz PRF
Antenna:	1-meter-diameter rotating dish producing 2 spot beams sweeping in a circular pattern
Mass:	200 kilograms
Power:	220 watts
Average Data Rate:	40 kilobits per second

Along with satellite orbit, scatterometer observes ocean surface as shown in Fig.5. Global coverage can be done. Then ocean wind direction and speed is estimated as shown in Fig.6 (a), (b).

5 days average of wind speed and vector wind is shown in Fig.6 (a) while 5 days average of dynamic height[21] and winds are shown in Fig.6 (b), respectively.

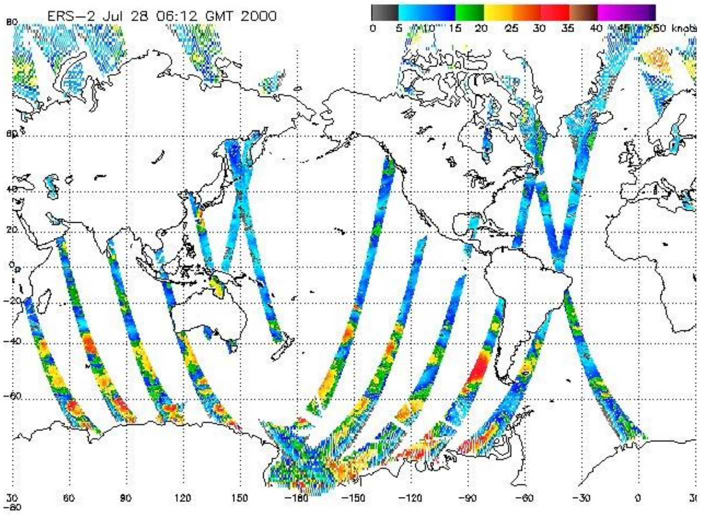
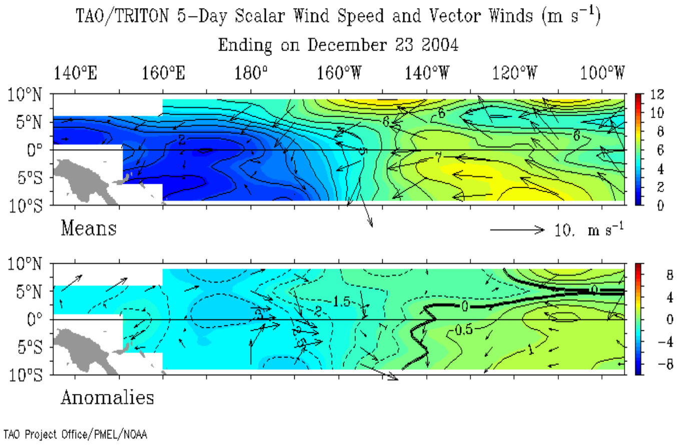
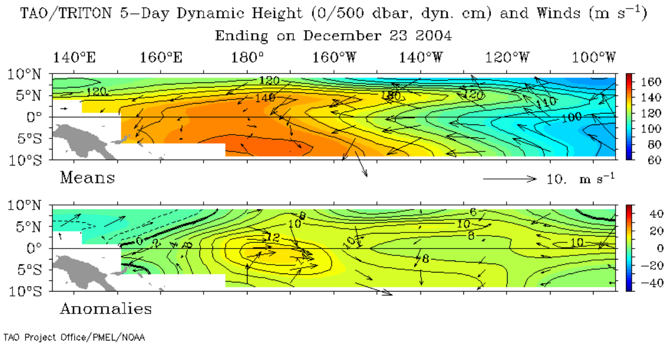


Figure 5 Example of scatterometer observed ocean wind along with satellite orbit.



(a)Wind speed and vector winds

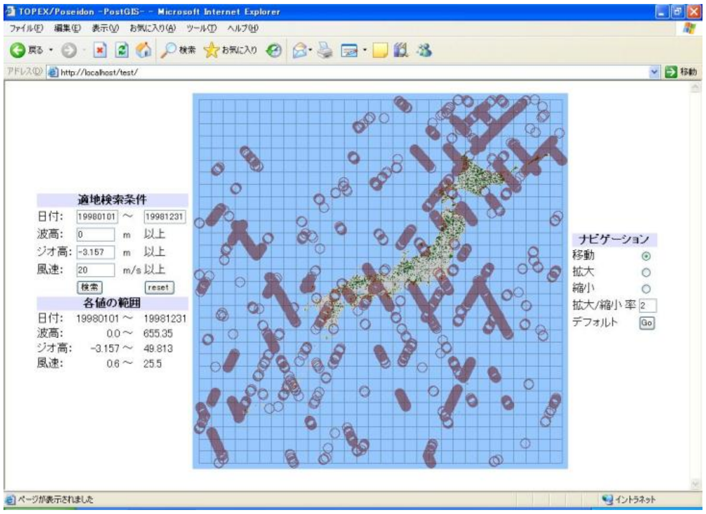


(b)Dynamic height and winds

Figure 6 Example of estimated ocean wind direction and speed

III. EXPERIMENTS WITH THE PROPOSED FOSS/GIS

From the MapServer site, Japan and its vicinity of map is downloaded. Web site is designed with php. Under the web site, there is the PostGIS with the databases of Topex/Poseidon altimeter data derived geoid and significant wave height as well as ADEOS-II scatterometer (SeaWinds) data derived wind direction and wind speed. Through web site, search conditions of ocean wind speed, significant wave height, and geoid can be input using radio button. Then search results are obtained after the input. If the search condition of wind speed (it is greater than 20 m/s) is input, then spatial retrieval result is displayed as shown in Fig.7.


TOPEX/Poseidon -PostGIS- - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る
検索
お気に入り
アドレス(D)
移動
ページが表示されました
イントラネット

http://localhost/test/

適地検索条件

日付: 19980101 ～ 19981231
波高: 0 m 以上
ジオ高: -3.157 m 以上
風速: 20 m/s 以上
検索 reset

各値の範囲

日付: 19980101 ～ 19981231
波高: 0.0 ～ 655.35
ジオ高: -3.157 ～ 49.813
風速: 0.6 ～ 25.5

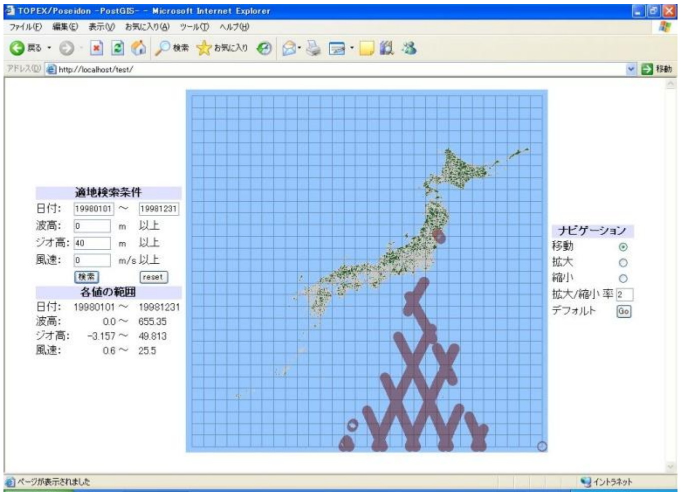
ナビゲーション

移動 ●
拡大 ○
縮小 ○
拡大/縮小 率 2
デフォルト Go

Figure 7 Spatial retrieval result with the condition of which wind speed is greater than 20 m/s

Spatial retrieval result with the search condition of which geoid potential is greater than 40 m is shown in Fig.8. Also, spatial retrieval result with the search condition of which geoid potential is greater than 40 m and wind aped is greater than 20 m/s is shown in Fig.9. Meanwhile, Fig.10 shows spatial retrieval result with the search conditions of which wind speed, significant wave height, geoid potential, and the distance from the nearest coastal lines. Such these spatial retrievals are specific feature of GIS.

[21] http://www.euro-argo.eu/content/download/21538/311050/file/04_guinehut_euro_argo_01.pdf


TOPEX/Poseidon -PostGIS- - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る
検索
お気に入り
アドレス(D)
移動
検索
reset
ナビゲーション
Go
ページが表示されました
イントラネット

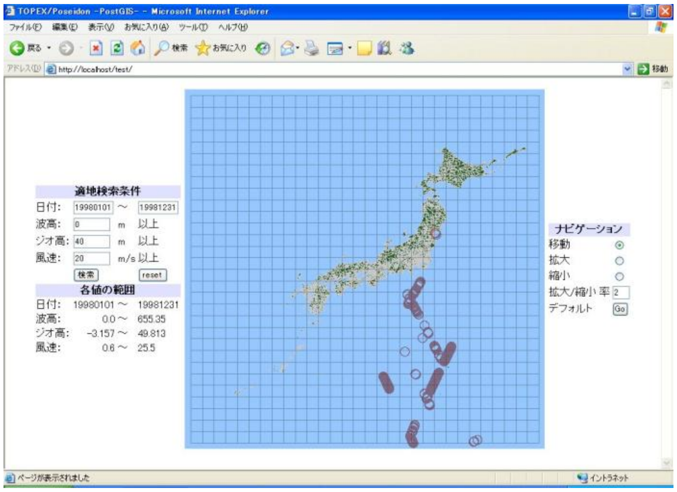
http://localhost/test/

適地検索条件
日付: 19980101 〜 19981231
波高: 0 m 以上
ジオ高: 40 m 以上
風速: 0 m/s 以上

各値の範囲
日付: 19980101 〜 19981231
波高: 0.0 〜 655.35
ジオ高: -3.157 〜 49.813
風速: 0.6 〜 25.5

移動 ◉
拡大 ○
縮小 ○
拡大/縮小 率 2
デフォルト

Figure 8 Spatial retrieval result with the search condition of which geoid potential is greater than 40m


TOPEX/Poseidon -PostGIS- - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る
検索
お気に入り
アドレス(D)
移動
検索
reset
ナビゲーション
Go
ページが表示されました
イントラネット

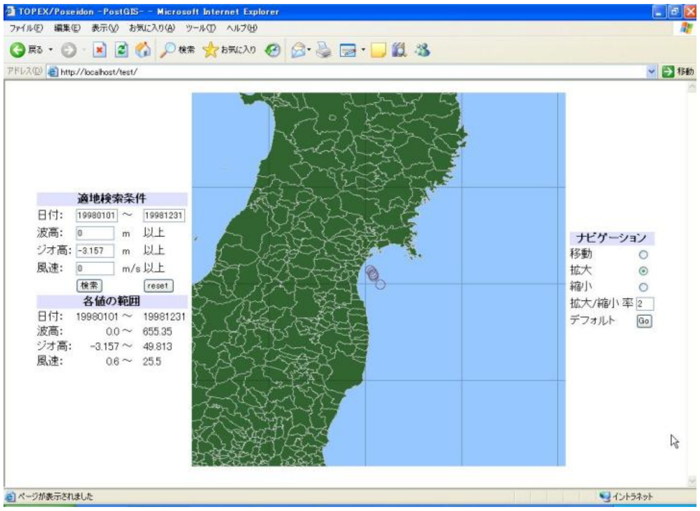
http://localhost/test/

適地検索条件
日付: 19980101 〜 19981231
波高: 0 m 以上
ジオ高: 40 m 以上
風速: 20 m/s 以上

各値の範囲
日付: 19980101 〜 19981231
波高: 0.0 〜 655.35
ジオ高: -3.157 〜 49.813
風速: 0.6 〜 25.5

移動 ◉
拡大 ○
縮小 ○
拡大/縮小 率 2
デフォルト

Figure 8 Spatial retrieval result with the combined conditions between wind speed (greater than 20 m/s) and geoid potential (greater than 40m).


TOPEX/Poseidon -PostGIS- - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る
検索
お気に入り
アドレス(D)
移動
検索
reset
ナビゲーション
Go
ページが表示されました
イントラネット

http://localhost/test/

適地検索条件
日付: 19980101 〜 19981231
波高: 0 m 以上
ジオ高: -3.157 m 以上
風速: 0 m/s 以上

各値の範囲
日付: 19980101 〜 19981231
波高: 0.0 〜 655.35
ジオ高: -3.157 〜 49.813
風速: 0.6 〜 25.5

移動 ○
拡大 ◉
縮小 ○
拡大/縮小 率 2
デフォルト

Figure 9 Spatial retrieval result with the search conditions among wind speed, significant wave height, geoid potential, and the distance from the nearest coastal lines.

IV. CONCLUSION

Customization can be done smoothly. It is easy to customize the PostGIS (extension of PostgreSQL) with Mapserver through the php. Also, it is confirmed that the most of functionalities of PostGIS (Submittion of queries, retrievals of the appropriate data from the database, display the retrieved results on the php web browser). Furthermore, image processing and analysis are also available and can be applied to the retrieved data.

Because the proposed PostGIS is modified version of free open source software, everybody may download and install and modify easily. The proposed system is open to the public upon request with the condition of credibility of the name of Arai/Saga University.

ACKNOWLEDGMENT

The author would like to thank to Mr. Shuichi Uchiyama for his effort to the experiments.

REFERENCES

[1] Kohei Arai, Earth observation satellite imagery data processing methods by means of Java language, Morikita Publishing Inc., 2001.

[2] Kohei Arai, Free Open Source Software GIS, Proceedings of the Computer Assisted Teaching Conference (CATCON 2006), ISPRS Technical Commission VI Symposium"E-LEARNING AND THE NEXT STEPS FOR EDUCATION" 2006

[3] Kohei Arai, Open GIS with spatial and temporal retrievals as well as assimilation functionality, Proceedings of the Asia Pacific Advanced Network Natural Resource Workshop, Utilization of Earthy Observation Satellite-Digital Asia Special Session 1,p. 8, 2003.

[4] Kohei Arai, Measuring and analysing methods for ocean environment by means of satellite remote sensing, Proceedings of the Advanced Science and Technology for Utilization of Ocean Energy, p11, 7, 2007.

[5] Kohei Arai, Ocean environment measuring and ocean energy exploration methods by means of satellite remote sensing, Proceddings of the Advanced Science and Technology for Utilization of Ocean Energy, p46-53, 7, 2007.

[6] Kohei Arai, Geographic information system: GIS based on neural network for appropriate parameter estimation of geophysical retrieval equations with satellite remote sensing data, Proceedings of the IEEE Geoscience and Remote Sensing, PID 220128, 2006.

[7] K.Arai, Sea Surface Temperature estimation method for ocean areas and seasons using Geographic Information System as a neural network, Sediment and Ecohydraulics, Proc.in Marine Science, 9, Elsevier ISBN978-0-444 53184-1, 2007.

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science, and Technology of the University of Tokyo from 1974 to 1978 also was with National Space Development Agency of Japan (current JAXA) from 1979 to 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He was appointed professor at Department of Information Science, Saga University in 1990. He was appointed councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was also appointed councilor of Saga University from 2002 and 2003 followed by an executive councilor of the Remote Sensing Society of Japan for 2003 to 2005. He is an adjunct professor of University of Arizona, USA since 1998. He also was appointed vice chairman of the Commission "A" of ICSU/COSPAR in 2008. He wrote 30 books and published 332 journal papers.

Knowledge Sharing Protocol for Smart Spaces

Jussi Kiljander
VTT Technical Research Centre of Finland
Oulu, Finland

Francesco Morandi
ARCES, University of Bologna
Bologna, Italy

Juha-Pekka Soininen
VTT Technical Research Centre of Finland
Oulu, Finland

Abstract— **In this paper we present a novel knowledge sharing protocol (KSP) for semantic technology empowered ubiquitous computing systems. In particular the protocol is designed for M3 which is a blackboard based semantic interoperability solution for smart spaces. The main difference between the KSP and existing work is that KSP provides SPARQL-like knowledge sharing mechanisms in compact binary format that is designed to be suitable also for resource restricted devices and networks. In order to evaluate the KSP in practice we implemented a case study in a prototype smart space, called Smart Greenhouse. In the case study the KSP messages were on average 70.09% and 87.08% shorter than the messages in existing M3 communication protocols. Because the KSP provides a mechanism for automating the interaction in smart spaces it was also possible to implement the case study with fewer messages than with other M3 communication protocols. This makes the KSP a better alternative for resource restricted devices in semantic technology empowered smart spaces.**

Keywords- ***Semantic Web; SPARQL; Ambient Intelligence; Ubiquitous Computing; embedded system; M3.***

I. Introduction

Smart spaces are realizations of ubiquitous computing (ubicomp) [1] and ambient intelligence (AmI) [2] visions. A typical smart space consists of a large amount of devices which in co-operation provide services for users. In order to provide relevant services in the right situations the devices need to share knowledge about the smart space with each other. Fortunately, a lot of knowledge representation (KR) technologies have been developed for the emerging Future Internet paradigm, called the Semantic Web [3] that could be also exploited in ubicomp/AmI domain. This has also been proposed by Lassila [4], and Chen [5], for example. The M3 concept [6] is a recent example of ubicomp interoperability framework which utilizes semantic technologies for knowledge representation.

Many of the devices in smart spaces are resource restricted in terms of memory, processing capacity, and energy. Additionally, typical communication technologies in smart spaces such as the 6LowPAN [7], and Bluetooth low energy (BLE) [8], for example, possess limited capabilities when compared to the technologies used in the Web. On the other hand, the current technologies enabling KR in the Semantic Web such as Resource Description Framework (RDF) [9], RDF Schema (RDFS) [10], Web Ontology Language (OWL) [11], and SPARQL [12] use either Extensible Markup Language (XML) based or human readable[1] syntax. These formats both require a large amount of memory and are slow to process in low capacity computing platforms. As a result they are not as such feasible for real-life smart spaces. Binary XML formats such as Efficient XML Interchange (EXI) [13] and X.694 [14] provide feasible solutions for compressing XML, but cannot be used with non-XML based semantic technologies such as the SPARQL, for example. Another interesting approach for Semantic Web based KR in resource restricted devices is the Entity Notation (EN) [15]. As a lightweight KR notation the EN is a good alternative for typical RDF serialization formats such as the RDF/XML, Turtle, and N-Triple, but it does not provide SPARQL-like mechanisms to query and update knowledge in smart spaces.

In M3 applications the problem with resource restricted computing platforms has been typically solved by utilizing gateways which transform the proprietary format data from low capacity devices to semantic format. However, this approach complicates the system unnecessarily as for each new device a new gateway is needed (or new interface to existing gateway needs to be added). If all the communication between smart space agents would be based on common knowledge sharing protocol instead, the smart spaces would be much easier to develop and maintain. Additionally, since the common knowledge sharing protocol would enable also resource restricted devices to access the information published by other devices it would be easier to develop context-aware embedded systems capable of providing relevant services for users.

In this paper we present a novel Knowledge Sharing Protocol (KSP) for M3-like semantic interoperability frameworks. The KSP provides all kind of agents from low capacity embedded systems to high end personal computers with SPARQL-like mechanisms for accessing and manipulating the knowledge in smart spaces. Unlike normal SPARQL, however, the KSP is designed to be suitable for smart spaces. Instead of the sparse human readable syntax used in SPARQL the KSP uses a compact binary format which allows significantly shorter messages to be created. Features such as the persistent update and max request size option make it also easier to exploit semantic technologies in resource restricted devices and networks. Additionally, to support the

[1] By human readable syntax we refer to notations designed to be used by developers as such. These kinds of formats are used, for example, in SPARQL, Notation3, and Turtle.

heterogeneous nature of smart spaces the KSP defines various bindings for typical communication and networking technologies used in smart spaces. In the paper bindings for the most typical transports User Datagram Protocol (UDP) and Transmission Control Protocol (TCP) are presented.

The rest of the paper is structured as follows. In the section 2 we present short overview of the SPARQL and M3 concepts as necessary background information for the paper. The section 3 describes the KSP in high detail. In the section 4 we illustrate the KSP with practical example and compare it with existing M3 communication protocols (M3CP). In the section 5 conclusions and future work directions are presented.

II. Background

A. SPARQL

SPARQL provides the standard way to access and manipulate RDF data in the Semantic Web. The importance of SPARQL to the Semantic Web is formulated by the W3C Director Tim Berners-Lee as follows: “Trying to use the Semantic Web without SPARQL is like trying to use a relational database without SQL.” [16]

SPARQL 1.1 defines four types of query forms: SELECT, CONSTRUCT, ASK and DESCRIBE. All these query forms use solutions obtained via pattern matching to form result sets or RDF graphs. The SELECT query returns the bound variables as such. The CONSTRUCT query returns an RDF graph constructed by substituting variables in a set of RDF triple patterns with bound variables obtained from the pattern matching. ASK query returns a Boolean value indicating whether the query has a solution or not. The DESCRIBE query returns a single RDF graph that contains data about the requested resource.

In addition to the query language the SPARQL 1.1 specification defines an update language for manipulating RDF data. The update operations are classified into two groups: graph management and graph update. Management type operations provide mechanisms for creating, destroying, moving and copying named graphs, or adding the contents of one graph to another. In contrast to the management operations used for operating on the graph level the update operations are used for modifying triples inside the graphs. The update operations are: INSERT DATA, DELETE DATA, DELETE/INSERT, LOAD and CLEAR. The DATA format operations operate on concrete RDF triples meaning that the use of variables is prohibited. Blank nodes are also not permitted in the DELETE DATA operation. The DELETE/INSERT operation is used to modify triples in the graph store based on bindings obtained via query pattern matching. It is possible to omit either DELETE or INSERT part of the operation in which case only the remaining part of the operation is executed. LOAD operation is used to insert triples from a RDF document specified by an URI into the graph store. CLEAR operation is used to remove all triples from the specified graphs.

SPARQL provides also many features that can be used to modify the outcome of query and update operations. These features include, for example, OPTIONAL graph pattern, FILTER, BIND, subquery, GROUP BY, and various solution sequence modifiers. The OPTIONAL graph pattern is used to declare a graph pattern which does not need to match in order for the query pattern match to succeed. FILTER keyword provides a way to restrict the solution to those in which the FILTER expression evaluates as true. The BIND feature allows a variable to be bound with a new value. The subquery feature makes it possible to combine results of multiple queries by allowing queries to be put inside of queries. The GROUP BY keyword allows sets of data to be grouped together so that aggregate functions such as average, minimum, maximum, sum, and count can be performed on the individual groups.

In SPARQL the results of the query operations can be modified using modifiers such as ORDER BY, LIMIT, OFFSET, and DISTINCT. The ORDER BY keyword can be used to sort the results of the query to either ascending or descending order. The LIMIT keyword allows the number of results to be restricted. The OFFSET keyword provides a way to skip a number of results. The DISTINC feature can be used to remove duplicates from the results set.

B. M3 Concept

The M3 aims to combine Semantic Web technologies with publish/subscribe-based blackboard [17] architecture to provide multi-device, multi-domain and multi-vendor solution to semantic level interoperability in smart spaces. Like in typical blackboard systems the idea in M3 is that knowledge is shared between various knowledge sources via common knowledge base.

The main advantage of the blackboard architecture is flexibility. In blackboard system the knowledge sources are not tied together in any way. The knowledge sources do not in fact have to know anything about each other – they just see the information published to the knowledge base. This makes it possible to add new, or remove existing knowledge sources without the need to reconfigure other knowledge sources. The blackboard system is also independent of any particular application domain. This means that the blackboard architecture can be used for sharing any kind of information and the shared information can be used in all kind of applications. Because of these features the blackboard architecture is ideal for smart spaces where it is not possible to *a priori* determine all the components or features needed in the future.

The difference between M3 and typical blackboard systems is that the knowledge in M3 is presented with semantic technologies such as RDF, RDFS, and OWL. There are many advantages in using the semantic technologies to present the knowledge in ubiquitous computing systems. First, because semantic technologies provide a natural way to describe any kind of knowledge as common machine-interpretable ontologies they make it easier to develop context-aware applications to smart spaces. Second, because of the flexible data model of RDF it is possible to add new information and create links between the information available in the knowledge base without breaking the existing applications. Third, the ontologies described with OWL and RDFS make it possible for knowledge sources to learn new concepts much in the same way as humans use encyclopedias to describe unfamiliar words.

In M3 the blackboard is called Semantic Information Broker (SIB). The Knowledge Processor (KP) is the name for the knowledge sources. M3 applications are created by KPs who provide services for end-users by interacting with each other via the SIB. The Smart Space Access Protocol (SSAP) defines the rules for KP-SIB interaction.

There are currently two serialization formats for the SSAP available: SSAP/XML and SSAP/WAX. Both formats define eight operations: *join*, *leave*, *insert*, *remove*, *update*, *query*, *subscribe*, and *unsubscribe*. The SSAP/XML version provides three types of query operations: Triple, Wilbur, and SPARQL 1.0. The M3 implementation using the SSAP/XML is called Smart-M3 [18]. A Word Aligned XML (WAX) version of the SSAP was introduced to better support the exploitation of M3 in low capacity devices [19]. In SSAP/WAX encoding scheme each XML tag is 32 bit long. This allows more compact messages to be constructed. The SSAP/WAX is used in RIBS version of the M3 [20]. Wilbur queries are not supported in the RIBS.

III. Knowledge Sharing Protocol

A. Overview

The objective of the KSP is to define the methods and syntax for knowledge sharing in AmI/ubicomb domain. The idea was to develop similar protocol to Semantic Web enhanced ubicomb systems, as the Constrained Application Protocol (CoAP) [21] is to the embedded web (i.e. the idea was to do the same for SPARQL/HTTP, as the CoAP has done for the HTTP). Initially we even planned to implement the KSP solely on top of CoAP. However, we soon understood that the heterogeneous nature of smart spaces requires the KSP to be usable also directly with various lower level transport technologies. In this paper bindings for TCP and UDP transports are presented. The principal model of KSP stack is illustrated in the fig. 1.

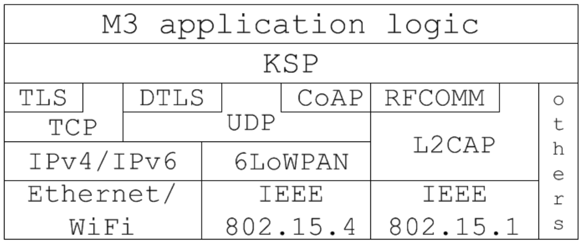


Figure 1. Knowledge Sharing Protocol stack

The different requirements of various transport technologies is managed in the KSP header presented in the section C. The KSP messages can contain also transport independent data and options fields. The structure for these fields is presented in sections D and E respectfully. All KSP messages are encoded in little endian format.

There are five major differences with the KSP and existing M3CPs (namely the SSAP/XML and SSAP/WAX). First, to provide a feasible solution for low capacity devices a binary format for the messages is used in KSP. The binary format is more compact and faster to parse, but not as versatile as the XML format used in the SSAP. Second, all the transactions in KSP are based on the SPARQL 1.1 whereas in SSAP the SPARQL is only one of the three query formats. Third, KSP does not require join and leave operations because the access control parameters can be added to any KSP message when needed. This makes the basic KP-SIB interaction more scalable because the SIBs do not have to keep track of the state of KPs. Fourth, the KSP allows KPs to define the maximum size for SIB responses. This is very useful for low capacity devices because the memory requirements can be estimated at the compile time. The fifth main difference is that the KSP defines persistent format also for update operations (i.e. DELETE, INSERT, and UPDATE). The persistent type update operations work so that the data manipulation part of the operation is executed if a solution is found from the query pattern matching. Similarly to the query operation the persistent update operations are re-evaluated every time the content of the SIB change. By allowing KPs to create simple rules to the SIB the persistent update operations provide a way to reduce the traffic in resource restricted networks and lighten the load on low capacity devices. The persistent operation are terminated with TERMINATE operation.

B. Messaging Model

The KP always initiates the transaction by sending a request (REQ) message to the SIB. To support the needs of the wide range of communication technologies the KSP defines two types of requests: Non-confirmable (NON), and Confirmable (CON). With NON request the SIB sends a response (RES) message only if results are found or an internal error has occurred. The NON requests are useful for TCP-like transports that provide a reliable delivery of messages and do not therefore require extra control from the KSP. Additionally, the NON request are useful for reducing the network traffic when it is not required that every message is delivered to the SIB. For example, a low capacity accelerometer updating acceleration value in high intervals does not have to care if a message is lost because it would take more time to update the old value than to insert a new one. The NON requests are also useful when KSP is used in multicast/broadcast manner to discover the available SIBs in the network. With CON type requests the SIB always sends a response to the KP. Therefore, the CON type request is typically used with transport technologies such as the UDP that do not provide reliable delivery of messages. The RES message is always send to the same socket where the REQ message was received

In addition to normal RES messages, the KSP defines indication (IND) messages which are used to notify the KP about changes in the persistent operations. The IND messages are typically used when the results of a persistent query operation (i.e. subscribe) change. Indications are also used to inform KP when a persistent operation needs to be terminated for some reason. Because transports such as UDP do not provide reliable delivery of messages the KSP provides acknowledgement (ACK) messages to be used with unreliable communication protocols to notify the SIB when the IND message has been received. The decisions on how long to wait before retransmitting and how many times to retransmit the

IND message are left for various SIB implementations. It is even possible to implement a SIB that can dynamically adjust to various networks by making these values modifiable via a specific graph in the SIB. The fig. 2 presents a sequence chart which illustrates the message exchange between KP and a SIB in a scenario where a KP1 subscribes to a triple and KP2 modifies the triple using UPDATE operation. Confirmable and Non-confirmable type requests are used by KP1 and KP2 respectfully.

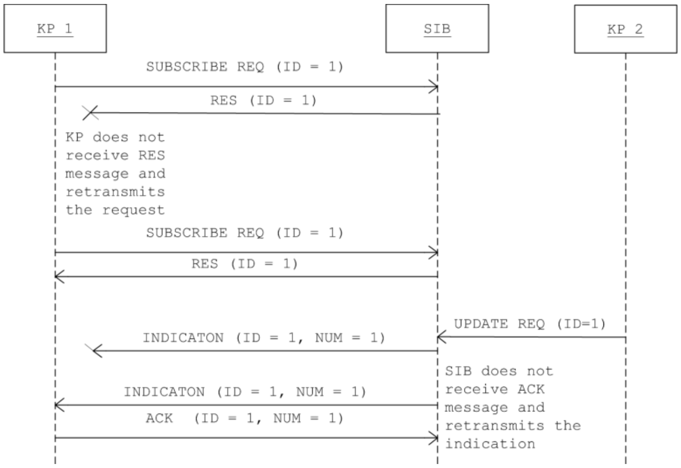

KP 1
SIB
KP 2
SUBSCRIBE REQ (ID = 1)
RES (ID = 1)
KP does not receive RES message and retransmits the request
SUBSCRIBE REQ (ID = 1)
RES (ID = 1)
UPDATE REQ (ID=1)
INDICATON (ID = 1, NUM = 1)
SIB does not receive ACK message and retransmits the indication
INDICATON (ID = 1, NUM = 1)
ACK (ID = 1, NUM = 1)

Figure 2. Example of message exchange with CON and NON requests

C. *Message Format and Semantics: Header field*

The fixed size header field contains parameters such as the version, transaction type, request type, and transaction identifier that are common for all transactions. The structure of the header field depends on the message type (REQ, RES, IND, or ACK) and the transport technology. The header size is eight bytes for TCP, four bytes for UDP REQ/RES/ACK messages, and six bytes for UDP Indications. Fig. 3 and fig. 4 illustrate the header formats for different message types with TCP and UDP transports.

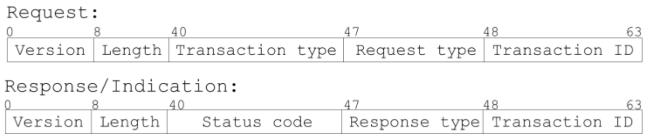


Figure 3. KPS header formats for TCP

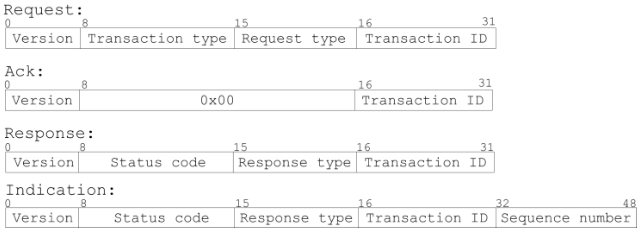


Figure 4. KSP header formats for UDP

The 8-bit *Version* field specifies the KSP version number. Value 0x01 is used for the version presented in the paper. With TCP the length of the KSP messages is defined in the 32-bit *Length* field. The length of the message is needed with TCP because the KSP message can be divided into multiple TCP segments. With UDP the *Length* field is not present and the whole KSP message must fit to a single UDP datagram. The maximum size for UDP datagram depends on the underlying protocols. For example, with 6LowPAN the maximum message size is 1024 bytes. In KP initiated messages the 7-bit *Transaction type* field specifies the type of the operation and the 1-bit *Request type* the type of the request (either NON or CON). Table 1 presents the code values for different transaction types. With connectionless transports the ACK message is identified by assigning value 0x00 for the *transaction type* and *request type* fields.

TABLE I. TRANSACTION TYPES

Transaction type	Code
DELETE DATA	0x01
INSERT DATA	0x02
UPDATE DATA	0x03
DELETE	0x04
INSERT	0x05
UPDATE	0x06
SELECT	0x07
ASK	0x08
CONSTRUCT	0x09
DELETE_PERSISTENT	0x0a
INSERT_PERSISTENT	0x0b
UPDATE_PERSISTENT	0x0c
SELECT_PERSISTENT	0x0d
ASK_ PERSISTENT	0x0e
CONSTRUCT_ PERSISTENT	0x0f
TERMINATE	0x10
RESET	0x11
CREATE	0x12
DROP	0x13
COPY	0x14
MOVE	0x15
ADD	0x16

In RES and IND messages the *Transaction type* and *Message type* fields are replaced with 7-bit *Status code* and 1-bit *Response type* fields (either RES or IND). The Status code specifies whether operation was successfully executed. Available status codes are illustrated in the table 2.

TABLE II. STATUS CODES

Status	Code
OK	0x00
ERROR: KSP version not supported	0x01
ERROR: Invalid transaction type	0x02
ERROR: Invalid message type	0x03
ERROR: Invalid data field format	0x04
ERROR: Invalid option type	0x05
ERROR: Invalid option format	0x06
ERROR: SIB internal error	0x07

In order to be able to pair RES and IND messages with the requests each KSP transaction is identified with 16-bit Transaction ID. UDP-like transports (no ordered delivery of messages) need also an additional identifier for the IND messages. The 16-bit *Sequence number* field is used for this purpose. The first indication for each persistent transaction must use a value 0x01 and the value is incremented by one for each following indication. A KP may discard an indication as outdated under the following condition:

$$\frac{-2^{16}}{2} < I_2 - I_1 < 0 \quad (1)$$

Where I_1 is the sequence number of the previous (not discarded) indication and the I_2 is the sequence number of the most recent indication. The right side of the equation checks if the sequence number for I_2 is smaller than for I_1. The left side of the equation checks whether the sequence number for the I_2 has wrapped around.

D. Message Format and Semantics: Data field

The KSP transactions can be divided into three categories: query, update, and terminate. In query and update operations the REQ message *Header* field is followed by the *Data* field which contains the transaction specific information. With TERMINATE operation the *Data* field is not needed because transaction identifier in the request header can be used to define the active persistent transaction to be terminated. In RES messages only query operations contain the *Data* field.

1) Encoding Format for RDF graph

Since the KSP is a query and update protocol for RDF the RDF graphs play a central role in almost all KSP messages. The common structure for *Graph* fields is illustrated in the fig. 5.

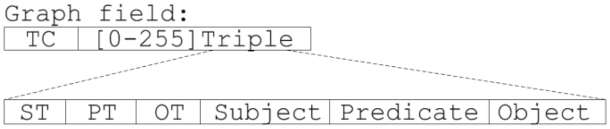


Figure 5. RDF graph field structure

The *Graph* field consists of 8-bit triple count (*TC)* field and a zero or more (maximum 255) *Triple* fields. Each *Triple* field starts with 3-bit *ST*, 2-bit *PT*, and 3-bit *OT* fields, which specify the content of the following *Subject*, *Predicate*, and *Object* fields respectfully. Possible types for these triple members (subject, predicate, and object) are presented in the table 3. The *Literal* type can be only used in objects.

TABLE III. TRIPLE MEMBER TYPES

Type	Code
Empty	0x00
URI	0x01
Reserved Word	0x02
Variable	0x03
Literal	0x04

The field structure for the various triple members is illustrated in the fig. 6. In *URI* field the 8-bit *Prefix index* field specifies the URI from the prefix list that is concatenated with the local URI to form the full URI (see prefix option). For example, with prefix index value 0x03 the third URI is selected form the prefix list. Prefix index value 0x00 is used for full URIs. The 8-16 bit *URI length* field specifies the length of the actual URI string field. The first bit of the *URI length* field denotes whether the length of the URI is presented with one or two bytes. This kind of length encoding allows not only compact messages, but also makes it possible to use longer URIs when necessary. The drawback is that the parser needs to perform extra work when decoding/encoding the messages. Similar length encoding is also used in other strings in the KSP.

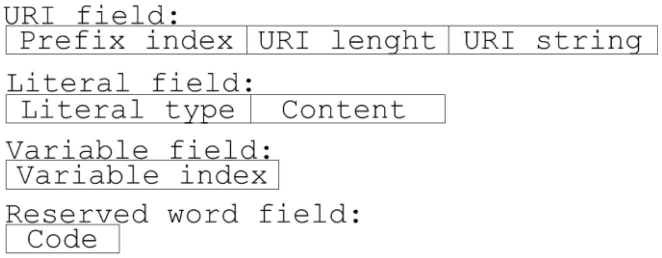


Figure 6. Field structure for triple member types

Literals in RDF can be either plain or typed. Only typed Literals are used in KSP however. The first eight bits in the Literal field is reserved for the type. The table 4 presents the supported Literal types in the KSP version 1.0.

TABLE IV. LITERAL TYPES

Type	Value
xsd:string	0x00
xsd:interger	0x01
xsd:float	0x02
xsd:dateTime	0x03
xsd:Boolean	0x04

With *xsd:string* type literal the first 8-32 bits is reserved for the length of the string. The two most significant bits specify the number of bytes used for the length field and following 6-30 specify the length of the *xsd:string*. The *xsd:integer* and *xsd:float* fields are 32-bit long. The IEEE 32-bit floating-point format is used for the *xsd:float* type. The *xsd:dateTime* field is 19 byte ASCII string. The *xsd:Boolean* field contains a 8-bit unsigned integer. Value 0x00 is reserved for “false” and 0x01 for “true”.

Variable type triple member field contains 8-bit variable index which is used as an identifier for the variable in the KSP message. With *n* variables the largest variable index is *n-1*. In SELECT operation it is also required that the indexes for projected variables (i.e. variables whose bindings are returned) start from zero. For example, in SELECT query with six variables of which two are projected variables the variable indexes 0x00 and 0x01 are used for projected variables and indexes from 0x02 to 0x05 are reserved to other variables.

The target in KSP is to provide as compact messages as possible and therefore two triple member types are designed just for this purpose. With *Empty* type the field (subject, predicate, or object) is not present and the corresponding value from the previous triple is used. The purpose of the *Empty* type is to enable more compact messages to be constructed by grouping triples with common subjects, predicates, and objects. The basic idea is similar to the Predicate-Object and Object lists in SPARQL, but the *Empty* type is more versatile because it allows both predicates without common subject, and objects without common subject-predicate to be grouped. The *Reserved word* type is another mechanism for shortening KSP messages. The idea in reserved words is to present common vocabulary with eight bit unsigned integers. The code values for various reserved RDF, XML Schema (XMLS), RDFS, and OWL words are presented in the table 5.

TABLE V. RESERVED WORDS

Word	Value
rdf:type	0x00
rdfs:Class	0x01
rdfs:subClassOf	0x02
rdfs:property	0x03
rdfs:subPropertyOf	0x04
rdfs:range	0x05
rdfs:domain	0x06
owl:TransitiveProperty	0x07
owl:SameAs	0x08
xsd:string	0x09
xsd:interger	0x0a
xsd:float	0x0b
xsd:dateTime	0x0c
xsd:Boolean	0x0d

2) *Data field format for query operations*

KSP defines six types of query operations. These operations include the transient and persistent formats for SELECT, ASK and CONSTRUCT. The persistent query operations in KSP are very similar to the SSAP equivalents. The only difference is that in KSP the SIB does not inform the KP about new and obsolete results, but just sends the current status of the query when the results have changed. The *Data* field format for transient and persistent query REQ and RES messages is presented in the fig. 7.

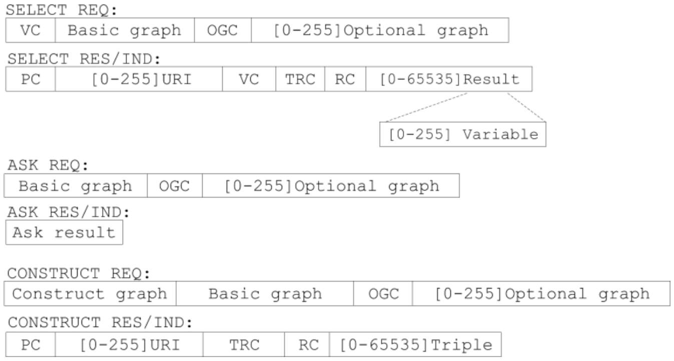


Figure 7. Data field format for query requests and responses

In SELECT requests the 8-bit VC field specifies the number of variables whose bindings are returned in the RES message. The following three fields are reserved for the query pattern which is matched against the RDF-database of the SIB. The query pattern consists of a Basic graph, and a number of Optional graph fields. The Basic graph field defines a triple pattern that must match in order for there to be a solution. The Optional graph fields contain graphs that do not reject the solution if no match for the graph pattern is found. The optional graphs are useful when a KP wants to receive some results even though not all the requested triples exist in the SIB. The 8-bit OGC field defines the number of optional graphs in the query pattern. The ASK request is otherwise identical to the SELECT request except there is no VC field. The CONSTRUCT request is also similar to the SELECT request expect the VC field is replaced by a Construct graph field. The Construct graph field defines triples to be constructed by replacing the variables with RDF terms obtained from the query pattern matching.

As already mentioned the query RES and IND messages are identical in KSP. This simplifies parser and SIB implementations when compared to the SSAP where the indications contain both new and obsolete results. The SELECT RES/IND messages start with 8-bit *PC* field which specifies the number of following URIs associated to the prefix indexes. The 16-bit *TRC* and *RC* fields specify the number of total results and results respectfully. The difference between these fields is that the total result count specifies the total number of results of the query operation whereas the result count defines the number of results in the RES message. These values can be different if all the results do not fit to a single message due the max response size. Each *Result* field in SELECT RES/IND message contains a number of bound variables. The exact number of variables (maximum 255) in a single *Result* field is defined by the 8-bit *VC* field. The first eight bits in the *Variable* field specify the type of the RDF term to which the variable is bound. Possible types for bound variables include *Empty*, *URI*, *Reserved Word*, and *Literal*. Same formats for these RDF terms fields are used as in the *Triple* field (see fig. 6). The CONSTRUCT RES/IND messages are otherwise similar to the SELECT messages expect there is no *VC* field and the *Results* field contains triples instead of RDF terms. The *Data* field in ASK RES/IND messages contains 8-bit unsigned integer defining whether solutions were found.

3) *Data field format for update operations*

In addition to the query operations the KPS provides various operations for manipulating the data in the SIB. These operations can be roughly divided into two groups: update and management. The difference between these operation types is that update operations are used to modify triples inside the graphs whereas the management operations provide mechanisms for creating, destroying, *etc.* complete graphs. Fig. 8 illustrates the data field formats for the data manipulation requests.

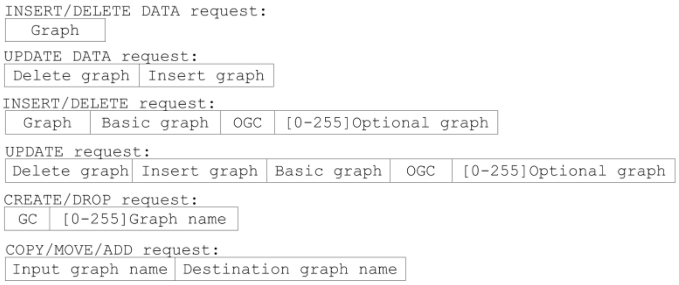


Figure 8. Data field structure for update request

The DATA format KSP update requests consist of various graphs which depending on the operation define either the triples to be deleted or/and inserted. The difference between plain and DATA type operations is that variables are not allowed in DATA operations. The advantage of DATA type operations is that large updates can be done without the need to first query bindings to the variables. The plain update operations provide also many advantages over the DATA type operations, however, and it depends on the situation which type should be used. The plain update operations are especially useful in situations where a KP would need to first query information and then use the information for modifying triples in the SIB. It is also possible to create simple rules to the SIB by defining the update to be persistent. In persistent update operation a new query pattern match is executed always when information in the SIB is modified. If solutions are obtained from the query pattern matching the data manipulation part of the operation is executed.

With SSAP it is only possible to access a single graph in the SIB. Sometimes it is feasible to be able to create separate graphs for different purposes however (e.g. privacy management). A simple way to manage privacy in M3 is to create separate graphs for various user groups and make the graphs accessible only to specific users. In KSP the CREATE and DROP operations are used for creating and destroying graphs respectfully. The 8-bit *GC* field defines the number of following *Graph name* fields which specify the URIs for the graphs to be created or destroyed. The encoding of the *Graph name* field is identical to the *URI* field. In COPY, MOVE and ADD operations the *Source graph* and *Destination graph* fields specify the source and destination graphs of the operation respectfully. Same encoding is used as for the *Graph name* field.

E. Message Format and Semantics: Options field

One of the main advantages of XML based protocols is extendibility. In KSP options are a way to achieve a certain level of extendibility in a non-XML protocol. Another advantage of options is that because options are, as the name implies, optional they make it possible to create more compact messages by leaving out the parts that are not needed in the particular message. The *Options* field follows the *Data* field in the KSP REQ messages. The 8-bit *OC* field defines the number of options in the request. Eight bits are reserved for the type in each option field. The table 6 presents the code values for the various option types.

TABLE VI. OPTION TYPES

Option	Code
PREFIX	0x00
DELETE GRAPH	0x01
INSERT GRAPH	0x02
QUERY GRAPH	0x03
OPTIONAL PATTERN	0x04
FILTER	0x05
SOLUTION MODIFIER	0x06
BIND	0x07
MAX RESPONSE SIZE	0x08
CREDENTIALS	0x09

The fig. 9 illustrates the field formats for the various KSP options. One of the most important design guidelines for the KSP was to support low capacity computing platforms. There are two options specified for this purpose: *Prefix* and *Max response size*. The *Prefix* option works as the PREFIX keyword in SPARQL and it is useful for shortening URIs appearing multiple times in a message. In *Prefix* field the first eight bits specify the number of following *URI* fields. In KSP messages the order number (prefix index) of the URI is used in the same way as the prefix label in SPARQL. By allowing the maximum size for RES and IND messages to be specified in the REQ message, the 32-bit *Max response size* option field makes it easier for a KP to estimate the memory requirements at compile time.

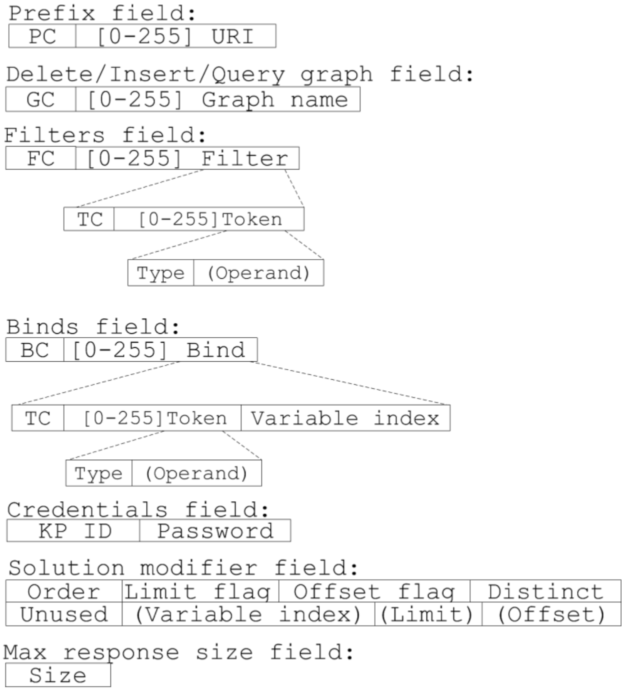


Figure 9. Field structures for options

Normally the query and update operations affect to all the graphs in the SIB (assuming KP has access rights). It is sometimes useful to make a KSP operation to affect only certain graphs in the SIB however. The *Delete graph*, *Insert graph* and *Query graph* options are used for this purpose. The first eight bits specify the number of named graphs. Same encoding is used for *Graph name* fields as for normal URI fields. The various named graph options are typically used together with credentials option which provides access control for KSP communication. Credentials are needed in situations where the SIB (or some graphs inside a SIB) can be accessed only by certain KPs. The *Credentials* field consists of 128-bit *KP ID* and *Password* fields which identify each KP uniquely. The first 8-bits in the *Password* field define the length of the following password. The *Credentials* option is typically only useful with transports such as Transport Layer Security (TLS) and Datagram TLS (DTLS) that provide encryption and trust for the authentication process. It should be noted that the KSP provides only a way to present the access rights and the actual access control management is out of the scope of the paper.

The KSP provides two options for modifying the solution produced in query pattern matching: *Filter* and *Solution modifier*. The *Filter* option can be used for limiting the solution to those which filter expression evaluates true. The first eight bits in the *Filters* field are reserved for filter count (FC). Each filter expression consists of a group of operand and operator tokens presented in Reverse Polish notation (RPN). The RPN was chosen because it provides a compact notation and allows fast processing of the expressions. The 8-bit *TC* field defines the number of tokens in the expression. The type for each token is also presented with eight bits. Values from 0x00 to 0x06 are reserved for operands and values from 0x07 onwards for operators. The table 7 presents the code values for various token types. If the token is operand type the *Type* field is followed by the *Operand* field which specifies the value for the operand.

TABLE VII. TOKEN TYPES

Token type	Code
xsd:string	0x00
xsd:integer	0x01
xsd:float	0x02
xsd:dateTime	0x03
xsd:Boolean	0x04
variable	0x05
URI	0x06
=	0x07
!=	0x08
<	0x09
>	0x0a
<=	0x0b
>=	0x0c
+	0x0d
-	0x0e
*	0x0f
/	0x10
||	0x11
&&	0x12
regex	0x13

The *Solution modifier* option provides ways for altering the solution sequence. By default the solution sequence of query operations is unordered. The 2-bit *Order* field can be used to define the solution order (unordered, ascending, or descending). The 1-bit *Limit flag*, *Offset flag*, and *Distinct flag* fields define whether the *Limit*, *Offset*, and *Distinct* modifiers are used. These modifiers are identical to the SPARQL equivalents. The next 3-bits are unused in this version of the KSP. If the solution sequence is ordered the 8-bit *Variable index* field defines the variable based on which the order of the solution sequence is created. For example, alphabetic order for the solution sequence is used if the variable is bound to xsd:string type. If the *Limit* and *Offset* flags are set the 16-bit *Limit* and *Offset* fields specify the maximum number and the offset for the results respectfully. The *Limit* and *Offset* modifiers are especially useful in large queries because they allow the KP to request the results in smaller packets.

The last option type in KSP is the *Bind*. The *Bind* option allows new values for variables to be assigned. Unlike the BIND keyword of SPARQL (can be used inside the query pattern) the *Bind* option is always executed after the query pattern matching. The *Bind* option is especially useful in update operations. Certain scenarios are even quite difficult to execute without the bind. For example, let’s consider a scenario where the KP needs to increment a certain literal by one. Without the *Bind* option the KP would first need to query the value, update it by one, and then insert the new value to the SIB. However, if another KP modifies the literal between the query and the update operations, the update operation does not work correctly. The structure of the *Bind* field is similar to the *Filter* field. The only difference is the 8-bit *Variable index* field specifying the variable to be bound.

IV. VALIDATION

In order to evaluate the KSP in practice we compared it with the other M3 communication protocols in a prototype smart space called Smart Greenhouse [22]. Of all the KPs in the Smart Greenhouse we chose to implement the Autocontrol KP because it is the most complex embedded system in that smart space. First, the Autocontrol KP was implemented with SSAP/XML and SSAP/WAX using as few and as compact messages as possible. In order to reduce the amount of messages the SPARQL queries were used instead of the Triple queries. We used both the Predicate-Object and Object list, as well as, minimum number of characters in variables to make the SPARQL queries as compact as possible. Then, to make direct comparison of the M3CPs possible we implemented the Autocontrol KP with KSP using the same operations used with SSAP. Finally, to demonstrate the full capabilities of the KSP we implemented the Autocontrol KP using persistent update operations.

In Smart Greenhouse the Autocontrol KP is responsible for modifying the status of the virtual actuators (LEDs, fans, and a water pump) available in the SIB. To do this it utilizes information about plant preferences and sensor measurements for humidity, temperature, and luminosity. For example, when the temperature is too high for a given plant the Autocontrol KP publishes information stating that the fans should be turned on. The information published by the Autocontrol KP is used by Actuator KP which modifies the physical actuators accordingly.

The operations needed for modifying the status for LEDs, fans, or the water pump when the luminosity, temperature, or humidity are out of the range of plant preferences are very similar. First, the Autocontrol KP needs to query the available actuators. Then to be aware when the state of an actuator needs to be modified the KP executes two ASK subscriptions. The first subscription informs when an actuator needs to be turned on and the second when an actuator needs to be turned off. The content of the subscriptions depends on the actuator type. After performing the ASK subscriptions the Autocontrol KP waits for IND messages from the SIB. Every time the SIB notifies the Autocontrol KP that the status of an actuator needs to be modified the KP updates a new status using the SSAP update operation. Again the content of the message depends on the actuator type.

The fig. 10 illustrates the message exchange between Autocontrol KP, Sensor KP, Actuator KP and the SIB. To make the sequence diagram as clear as possible we simplified it in two ways. First, only the messages related to the temperature and fans are illustrated and it should be noted that in reality similar message exchange is executed for other measurements and actuators as well. Second, only one of the two ASK subscriptions is illustrated in the sequence chart. It is also assumed that the static plant preference values for minimum (19.5 Celsius degree) and maximum (25.7 Celsius degree) temperature have been published into the SIB. The

Autocontrol KP uses CON type request while NON type request are used by the Sensor KP and the Actuator KP.

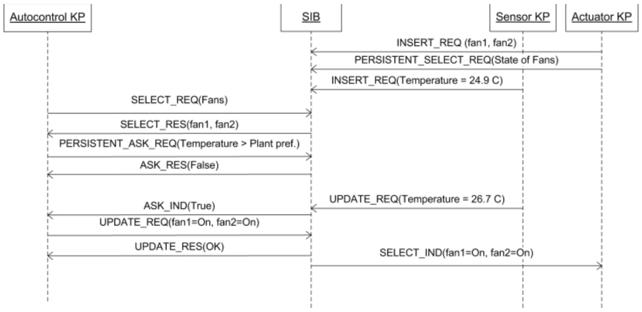

Autocontrol KP
SIB
Sensor KP
Actuator KP
INSERT_REQ (fan1, fan2)
PERSISTENT_SELECT_REQ(State of Fans)
INSERT_REQ(Temperature = 24.9 C)
SELECT_REQ(Fans)
SELECT_RES(fan1, fan2)
PERSISTENT_ASK_REQ(Temperature > Plant pref.)
ASK_RES(False)
UPDATE_REQ(Temperature = 26.7 C)
ASK_IND(True)
UPDATE_REQ(fan1=On, fan2=On)
UPDATE_RES(OK)
SELECT_IND(fan1=On, fan2=On)

Figure 10. Sequence chart illustrating the message exchange between the KPs and the SIB in Smart Greenhouse

The average size of REQ messages sent by the Autocontrol KP is illustrated in the fig. 11. The sizes for RES and IND messages are presented in the fig. 12. TCP was used as the transport technology with all M3CPs. For clarity sake we did not include the *join* and *leave* messages needed with SSAP/XML and SSAP/WAX. The first column in the fig. 11 presents the average size of messages used to query the different actuators (fans, LEDs, or a water pump) from the SIB. In the fig. 12 the first column presents the average size of the RES messages to the actuator query. The second and third columns in fig. 11 and fig. 12 represent the ASK subscription requests and indications respectfully. In fig. 11 the fourth column presents the average size for the update message that is send every time a subscribe indication is received from the SIB. The average size for the update response message is illustrated in the fourth column of the fig. 12. As can be seen from the fig. 11, the size for KSP query/subscribe request is on average only 18.63% of the corresponding SSAP/XML and 43.58% of the corresponding SSAP/WAX requests. In the case of the update request the average size of KSP message size is 15.22% of SSAP/XML and 27.50% of the SSAP/WAX requests. In the RES and IND messages the difference between KSP and SSAP is even bigger. The KSP response/indication message size is on average 6.89% of the SSAP/XML and 19.06% of the SSAP/WAX RES/IND messages.

In the previous study we illustrated how implementing the Autocontrol KP with KSP leads to a significantly shorter message sizes than with other M3CPs even when the same operations are used. The KSP provides also more advanced ways to implement the Autocontrol KP however. By using the persistent update operation both the workload on the Autocontrol KP and the network traffic can be dramatically decreased. Only two persistent update operations are needed for each actuator type and once the Autocontrol KP has performed these operations it can enter to a sleep mode. The persistent update requests are the only messages the Autocontrol KP has to send and it is therefore practical to compare them with the largest messages needed with other M3CPs. The average size of the persistent update requests sent by the Autocontrol KP is 165 bytes which is only 14.27% and 25.78% of the largest SSAP/XML and SSAP/WAX requests respectfully.

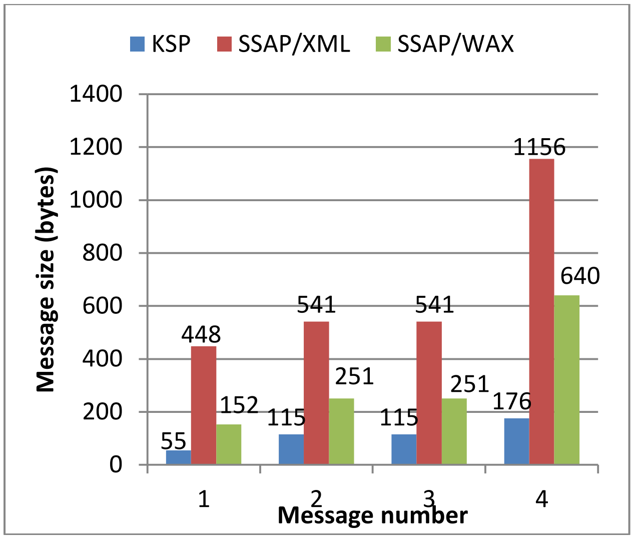


Figure 11. Average request size of Autocontrol KP with different M3 communication protocols

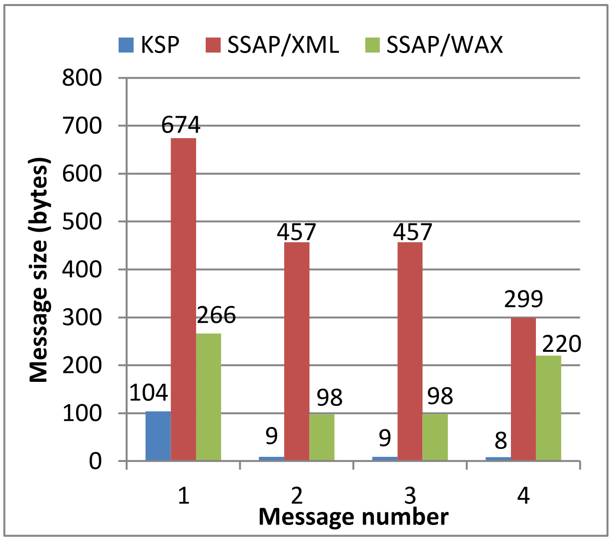


Figure 12. Average message size of the response and indication messages received by the Autocontrol KP with different M3 communication protocols

When CON type requests are used the size for each persistent update RESP message is four bytes (UDP assumed) which is only 0.59% of the largest SSAP/XML and 1.50% of the largest SSAP/WAX RES message. When used with TCP it is feasible to use NON type request in which case the persistent update response messages are not needed at all. The fig. 13 presents the message exchange between the KPs and the SIB in Smart Greenhouse when persistent update operations are used by the Autocontrol KP.

V. Conclusions and Future Work

We presented a novel knowledge sharing protocol for semantic technology empowered AmI systems. The KSP is designed for M3 applications but it can be also used with any other application that requires a compact protocol for knowledge sharing.

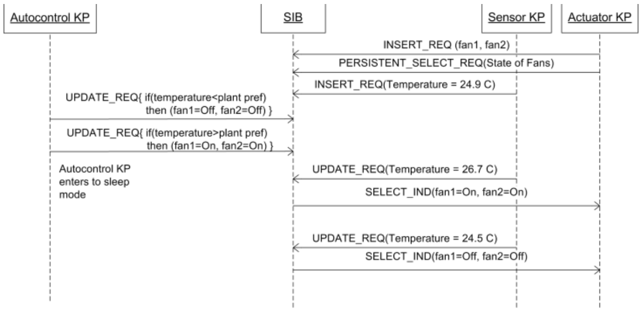

Autocontrol KP
SIB
Sensor KP
Actuator KP
INSERT_REQ (fan1, fan2)
PERSISTENT_SELECT_REQ(State of Fans)
INSERT_REQ(Temperature = 24.9 C)
UPDATE_REQ{ if(temperature<plant pref) then (fan1=Off, fan2=Off) }
UPDATE_REQ{ if(temperature>plant pref) then (fan1=On, fan2=On) }
Autocontrol KP enters to sleep mode
UPDATE_REQ(Temperature = 26.7 C)
SELECT_IND(fan1=On, fan2=On)
UPDATE_REQ(Temperature = 24.5 C)
SELECT_IND(fan1=Off, fan2=Off)

Figure 13. The message exchange between the KPs and the SIB when persistent update operations are used by the Autocontrol KP.

The main design guideline in the KSP was to implement SPARQL-like knowledge sharing mechanism in a compact binary format that is suitable for real-life smart spaces. In addition to the compact binary format, the feasibility of the KSP to resource restricted devices and networks is improved with mechanisms such as the persistent update, multi-transport support, and the max request size option, for example.

For evaluation purposes we implemented the Autocontrol KP of the Smart Greenhouse with different M3CPs. The KSP messages were on average 87.08% and 70.09% shorter than the SSAP/XML and SSAP/WAX messages respectfully. We also demonstrated how the Autocontrol KP implementation can be significantly simplified with persistent update operations. Only six (two for each actuator type) persistent update request were needed to fully implement the Autocontrol KP. It is evident that with fewer and more compact messages the KSP is more suitable for battery powered low capacity devices than the other M3CPs.

The KSP is designed to be compact and easily processable knowledge sharing protocol for ubicomb systems. This kind of format does not come without limitations however. The binary format limits both maximum amount and size of entities such as prefixes, graphs, triples, and results, for example. This may cause troubles in certain situations where huge amount of RDF triples need to be manipulated in a single operation. Another drawback in KSP is that unlike SPARQL it requires a good application programming interface (API) because the binary format is not suitable to be used by developers as such. We also choice not to implement some of the rarely used features of SPARQL 1.1 mainly because they would have made the KSP too complicated. The current version of the KSP does not support SPARQL 1.1 features such as DESCRIBE queries, Property paths, Aggregates and Subqueries, for example.

The adoption of semantic technologies in resource restricted devices is not only important for ubiquitous computing, but also for the Semantic Web. This is because, the Semantic Web is not going to get wider acceptance before there is enough data available to create meaningful Semantic Web applications. Therefore, in the future we are planning to exploit the KSP also in the field of IoT and Semantic Web. To this end we will further develop and evaluate the KSP. For example, new bindings for at least the BLE transport needs to be implemented. We are also considering whether some of the missing SPARQL 1.1 functionalities should be incorporated into the next version of the KSP.

In the future we will also need to make a more comprehensive study on the benefits and drawbacks of the KSP when compared to SPARQL/HTTP used in the Semantic Web. Because the KSP is a binary format with predefined places for parameters it is obviously much faster to parse than the SPARQL. However, the actual difference in parsing times needs to be measured with different workloads to justify the drawbacks caused by the binary format. The effect of persistent update operations on the performance of the SIB needs to also to be carefully analyzed in the future.

ACKNOWLEDGEMENTS

This work has been funded by the Merging IoT Technologies (MIoTe), the Device and Interoperable Ecosystem (DIEM), and Smart Objects for Intelligent Applications (SOFIA) projects.

REFERENCES

[1] M. Weiser, “The Computer for the 21st Century,” Scientific American, September 1991, pp. 94-100.

[2] E. Aarts, H. Harwing, and M. Schuurmans, “Ambient Intelligence,” The Invisible Future: The Seamless Integration Of Technology Into Everyday Life. Denning, P. (ed.), McGraw Hill, New York (2001).

[3] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” Scientific American, May 17, 2001, pp. 34-43.

[4] O. Lassila, “Serendipitous interoperability”, The Semantic Web Kick-off in Finland – Vision, Technologies, Research, and Applications, HIIT Publications, University of Helsinki, 2002.

[5] H. Chen, An Intelligent Broker Architecture for Pervasive and Context-Aware Systems, doctoral dissertation, University of Maryland, Baltimore County, Department of Computer Science and Electrical Engineering, 2004.

[6] A. Lappeteläinen, J. Tuupola, A. Palin, and T. Eriksson, “Networked systems, services and information – The ultimate digital convergence,” First International NoTA conference, 2008.

[7] Z. Shelby, and C. Bormann, 6LoWPAN: the Wireless Embedded Internet, John Wiley and Sons, 2010, p. 244

[8] B. SIG, “Bluetooth specification version 4.0,” Available at http://www.bluetooth.org, August 2012.

[9] G. Klyne and J. J. Carroll. 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax, W3C Recommendation, 10 February 2004, URL: http://www.w3.org/TR/rdf-concepts/.

[10] D. Brickley and R.V. Guha. 2004. RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation 10 February 2004, URL: http://www.w3.org/TR/rdf-schema/.

[11] W3C OWL Working Group. 2009. OWL 2 Web Ontology Language Document Overview. W3C Recommendation, 27 October 2009, URL: http://www.w3.org/TR/owl2-overview/.

[12] S. Harris and A. Seaborne. 2012. SPARQL 1.1 Query Language, W3C Working Draft, 5 January 2012, URL: http://www.w3.org/TR/sparql11-query/.

[13] J. Schneider and T. Kamiya, 2001. Efficient XML interchange (EXI) format 1.0. W3C Recommendation, 10 March 2001, URL: http://www.w3.org/TR/exi/

[14] International Telecommunication Union: X.694 (2004), http://www.itu.int/ITU/studygroups/com17/languages/x694.pdf

[15] X. Su, J. Riekki, and J. Haverinen, ”Entity Notation: enabling knowledge representations for resource-constrained sensors”, Personal and Ubiquitous Computing, 21 September, 2011, pp. 1-16

[16] B. DuCharme, Learning SPARQL: Querying and Updating with SPARQL 1.1, O’Reilly Media (2011)

[17] F. Buschmann, R. Meunier, H. Rohnert, P. Sommerlad, and M. Stal, Pattern-Oriented Software Architecture: A System Of Patterns, West Sussex, England: John Wiley & Sons Ltd., 1996

[18] J. Honkola, H. Laine, R. Brown, and O. Tyrkkiö, ”Smart-M3 information sharing platform,” proc. ISCC 2010, pp. 1041 – 1046

[19] A. Ylisaukko-oja, P. Hyttinen, J. Kiljander, J. Soininen, and E. Viljamaa, ”Semantic Interface for Resource Restricted Wireless Sensors,” IC3K 2nd International Workshop on Semantic Sensor Web – SSW2011, October, 2011, Paris, France.

[20] J. Suomalainen, P. Hyttinen, and P. Tarvainen, "Secure information sharing between heterogeneous embedded devices," proc. ECSA 2010, pp. 205-212

[21] Z. Shelby, K. Hartke, C Bormann, and B. Frank. B. 2012. Constrained Application Protocol (CoAP). IETF Internet-Draft 09, URL: http://datatracker.ietf.org/doc/draft-ietf-core-coap/ (2012)

[22] J. Kiljander, M. Eteläperä, J. Takalo-Mattila, and J. Soininen, “Opening information of low capacity embedded systems for Smart Spaces”, Proc. WISES 2010, pp. 23-28

AUTHORS PROFILE

Jussi Kiljander is a research scientist at VTT Technical Research Centre of Finland. He received his M.Sc. (Technology) from University of Oulu in 2010. His current research and Ph.D. studies are focused on ubiquitous computing and device interoperability with semantic technologies. He has published more than 10 scientific papers and contributed to several projects related to semantic interoperability and pervasive computing.

Francesco Morandi was born in Lugo, Italy, the 26th of July 1984. He received the degree in Electronical Engineering in 2006 and specialistic degree in Telecommunication Engineering in 2009 both from University of Bologna. In 2010 he worked for Fondazione Ugo Bordoni in Pontecchio Marconi (Bologna) as consultant for the television transition to digital (DVB-T). From 2011 he's working as researcher for the University of Bologna in ARCES (Advanced Research Center on Electronic Systems for Information and Communication Technologies E. De Castro). His current research is focused on developing and improving interoperable platforms for healthcare, maintenance and energy smart grids.

Prof. Juha-Pekka Soininen is a research professor of computing and computer architectures at VTT Technical Research Centre of Finland. He received his MSc, LicTech and Doctor of Science (Technology) degrees from University of Oulu 1987, 1997 and 2004 respectively. He has been Research scientist at VTT since 1988, senior research scientist since 1996 and research professor since 2007. He has been the leading expert in various large research projects at VTT during 1993 - 2011. These projects include contract research projects, joint research projects and European Union research projects. His current research deals with ubiquitous and distributed computing, system architectures, platform-based design methodologies, system architecture evaluation methods, and system-level design methods. He has been a reviewer in several international conferences, journals and books. His has published more than 70 scientific publications and he is a member of IEEE.

Localisation of Numerical Date Field in an Indian Handwritten Document

S Arunkumar[1], Pallab Kumar Sahu[2], Sudeep Gorai[2], Kalyan Ghosh[3]
[1]Dept of Information Technology
[2]Dept of Computer Science and Engineering
[3]Dept of Electronics and Communication Engineering
Institute of Engineering and Management
Kolkata, India

***Abstract*— This paper describes a method to localise all those areas which may constitute the date field in an Indian handwritten document. Spatial patterns of the date field are studied from various handwritten documents and an algorithm is developed through statistical analysis to identify those sets of connected components which may constitute the date. Common date patterns followed in India are considered to classify the date formats in different classes. Reported results demonstrate promising performance of the proposed approach.**

Keywords- Connected Components; Feature Extraction; Spatial Arrangement; K-NN classifier.

I. Introduction

Many institutions, business organisations etc. face the problem of processing handwritten document .No successful work regarding the decipherment of unconstrained cursive handwriting has been reported till date [1]. Nevertheless, when focused on certain restricted applications of handwritten text like revealing the location certain numerical data (phone number, pin code...), work becomes quite interesting. The deciphering of the location of the ‘date’ field in a handwritten document is one such interesting work which has been illustrated in this paper. This may find huge industrial importance as many handwritten documents are required to be sorted or categorized according to the dates mentioned on it. Our proposed algorithm is an advancement to make these industrial or organizational works automated. This will allow additional advantage to fax, photocopy and scanning machines, where sorting handwritten documents based on dates (mentioned in it) could appreciably be made automated.

Works regarding the recognition of a given date information has been reported by many [2][3][4],each establishing a unique technique of its own. These algorithms however assume that the given input is a date field (i.e. the pixel locations of the ‘date’ field is already considered to be known). The challenging task remaining, however, is the detection or identification of those pixels from handwritten documents which may constitute the date field. Our paper focuses only on this challenging issue, so that those pixels which are extracted could be fed into the above mentioned algorithms for recognition, thus making our work a pioneering one in the field of Document Image Analysis.

In India, the most commonly followed date patterns are DD-MM-YY, DD/MM/YY and DD.MM.YY. There are more date patterns like DD-MM-YYYY, DD/MM/YYYY, DD.MM.YYYY etc. but our paper focuses only on the above three patterns. It could be convincingly said that the proposed algorithm to locate the former patterns could also be used to locate the later ones with slight alterations.

In this paper we necessitate that the spatial orientation of the connected components in a numerical date field follows a specific structure and can be exploited for the localisation task. We thus target to find all classified date fields in each and every text line of the handwritten document.

II. Overview Of The Proposed Algorithm

The proposed algorithm comprises of a series of processes (depicted by a flowchart shown in Figure I) which includes Pre-processing, Scrutinization of Eight Consecutive Connected Components (ECCC) and Further Classification of DD-MM-YY and DD.MM.YY. Each of these processes is discussed in detail in the subsequent sections of the paper.

Since our study demanded us to have a well maintained database, a database was created (for both training and testing) by scanning numerous handwritten documents of various individuals. Each of these images (documents written on white paper) were scanned at 600 dpi and stored in JPEG format.

A section is also devoted to demonstrate the outcome of our experimentation. All the results obtained, having been enunciated to corroborate our study.

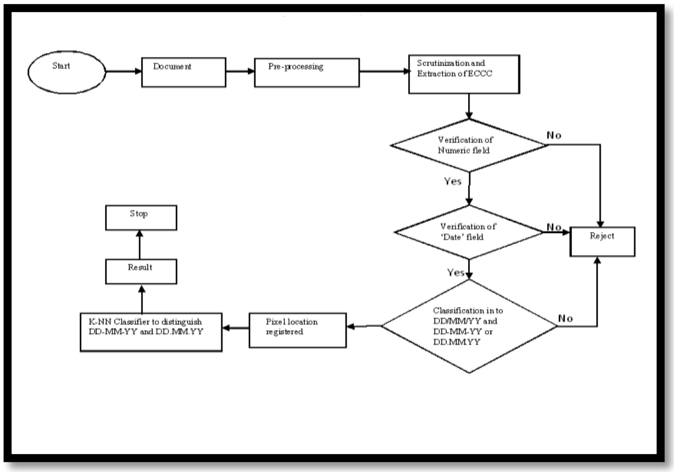
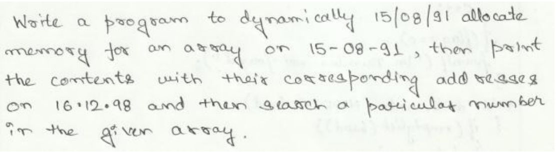

Start
Document
Pre-processing
Scrutinization and Extraction of ECCC
Verification of Numeric field
No
Yes
Verification of 'Date' field
No
Reject
Yes
Classification in to DD/MM/YY and DD-MM-YY or DD.MM.YY
No
Pixel location registered
K-NN Classifier to distinguish DD-MM-YY and DD.MM.YY
Result
Stop

Figure I: the flowchart of the proposed algorithm.

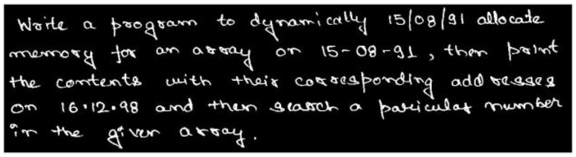
III. PREPROCESSING

Since our algorithm basically focuses on the scrutinization of the spatial arrangements of connected components and not on other aspects such as colour, texture etc, all the handwritten documents which are considered for statistical analysis or testing are converted to binary image such that the background is assigned a 'zero' pixel value and all the handwritten components are assigned a pixel value of 'one'. The overall image thus appears as shown in Figure III.



Write a program to dynamically 15/08/91 allocate memory for an array on 15-08-91, then print the contents with their corresponding addresses on 16·12·98 and then search a particular number in the given array.

Figure II: showing the original document to be processed.



Write a program to dynamically 15/08/91 allocate memory for an array on 15-08-91, then print the contents with their corresponding addresses on 16·12·98 and then search a particular number in the given array.

Figure III: showing the binary image of the converted document.

Once the document is converted into binary image (in the above mentioned way), all the text lines are extracted from it. Extraction of text lines implies grouping of connected components that belongs to the same line. For scrutinization of spatial features, the precise knowledge of these alignments is necessary. A histogram projection based text segmentation technique (inspired from [5]) is used.

IV. SCRUTINIZATION OF EIGHT CONSECUTIVE CONNECTED COMPONENTS (ECCC)

The text lines extracted are then used for further examination. Since all the above specified classes (DD-MM-YY, DD/MM/YY and DD.MM.YY) deals with eight connected components so a group of eight consecutive connected components (ECCC) is extracted one at a time (say for example $C_1, C_2, C_3.....C_8$; where all C_i belong to the same text line and C_1 is the first connected component of the ECCC). The widths of the minimum bounding rectangle enclosing these eight connected components are calculated and the maximum of these is found out and stored (say as W_{max}).A condition:- $X_{min}(C_{i+1})> X_{min}(C_i)$ is used to eliminate instance(s) like the dot of 'i', noises, disoriented connected components (shown in Figure V and Figure VI.) etc. The goal now is to decipher whether the set of ECCC may constitute a date or not?

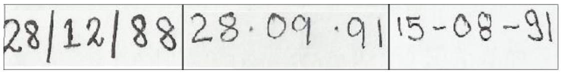


Figure IV: showing the three classes of date.

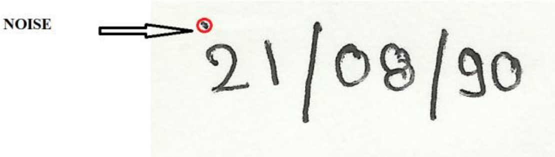


Figure V: showing the presence of noise (shown by an arrow mark).

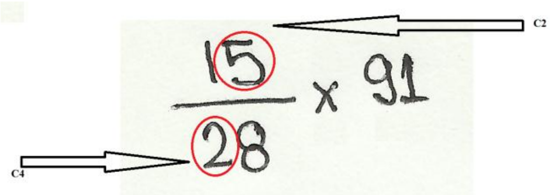


Figure VI: showing the case(s) eliminated when the condition $X_{min}(C_{i+1})> X_{min}(C_i)$ is used; the connected component C2 and C3(denoted by arrow marks) violates the above condition, hence not detected as the desired ECCC.

The outline of the process is described as follows:-

1) *The horizontal interspatial distance between the above processed eight connected components is calculated*

(say for example $S_1, S_2,.....S_7$; where S_i is the horizontal interspatial distance between C_i and C_{i+1}). It is then checked to see that the value of no S_i exceeds the value of 1.5times of $W_{max.}$. This relation has been found out experimentally to avoid cases shown in Figure VII. It is a common observation that when dates are written, all the components representing it are within a certain horizontal interspatial distance from its neighbouring.

2) *If the set of eight consecutive connect components*

(say $C_i, C_{i+1},....C_{i+7}$) obeys with the conditions of the above step(Step I), then it is sent for further examination(Step III), else the next set (i.e. C_{i+1} , C_{i+2} ,....C_{i+8}) is considered and processed(Step I). This process goes on iteratively until all the set of eight consecutive is considered for a particular text line. When a text line is checked thoroughly (i.e. all the set of eight consecutive components is scrutinized), then the next text line is processed.

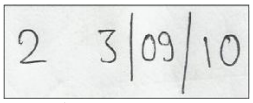


Figure VII: Incorrect formats of date:- a case that is avoided in our algorithm.

3) *Verification of numeric fields:-*

It could be easily learnt that in any classified format (as discussed above), the first, second, fourth, fifth, seventh and eighth constitute a numerical field. This process is inspired from [6] where features are defined to characterise the regularity of numerical fields. The feature vector is defined comprising of the following component f1, f2, f3, f4, f5, f6. Where for the set ECCC (say from C_i to C_{i+7}) f1= $\frac{H(Ci+1)}{H(Ci)}$, f2= $\frac{Y(Ci+1)}{Y(Ci)}$, f3=$\frac{H(Ci+4)}{H(Ci+3)}$, f4=$\frac{Y(Ci+4)}{Y(Ci+3)}$, f5=$\frac{H(Ci+7)}{H(Ci+6)}$, f6=$\frac{Y(Ci+7)}{H(Ci+6)}$; where H represents height and Y represents Y co-ordinate of the centre of gravity of the minimum bounding rectangle enclosing the connected component. A training set of 250 documents is studied to learn the range values in which these features lie. These relations of the connected components with its immediate neighbours reveal features which may characterise it as a numerical field [6].

4) Spatial Orientation of Numerical fields with respect to its Separators:-

The above classified categories of date formats accommodate three types of separators, which are slash (/), dash (-) and dot (.). Learning of the spatial orientation of the numerical field with respect to its separators is the crux of our algorithm. A pattern is studied from a database of around 250 documents which thoroughly emphasizes on the localisation of the date field and classification of it into various categories of date format. Spatial features are extracted to classify the date format into DD/MM/YY, DD-MM-YY or DD.MM.YY and NON-DATE SET. Further classification is done to distinguish among DD-MM-YY and DD.MM.YY format.

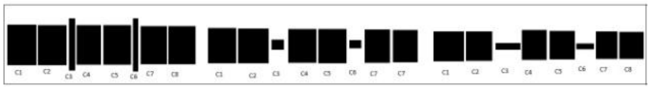


Figure VIII: showing a sample of the patterns of the minimum bounding rectangles of ECCC of all the three classes.

A feature vector is defined comprising of elements $Y_{min}(C_2)$, $Y_{min}(C_3)$, $Y_{min}(C_4)$, $Y_{min}(C_5)$, $Y_{min}(C_6)$, $Y_{min}(C_7)$, $Y_{max}(C_2)$, $Y_{max}(C_3)$, $Y_{max}(C_4)$, $Y_{max}(C_5)$, $Y_{max}(C_6)$, $Y_{max}(C_7)$; where Y_{min} and Y_{max} implies the minimum and maximum values of the Y co-ordinate of the minimum bounding rectangle.

Relationships are obtained among these features elements by training around 250 documents, these kinships are expressed (for the above defined classifications: DD/MM/YY, DD-MM-YY or DD.MM.YY and NON-DATE SET) in the form of mathematical inequalities (shown below).

For Class DD/MM/YY:

$$Y_{min}(C_3) \leq Y_{min}(C_2) \leq Y_{max}(C_3)$$
$$Y_{min}(C_3) \leq Y_{max}(C_2) \leq Y_{max}(C_3)$$
$$Y_{min}(C_3) \leq Y_{min}(C_4) \leq Y_{max}(C_3)$$
$$Y_{min}(C_3) \leq Y_{max}(C_4) \leq Y_{max}(C_3)$$
$$Y_{min}(C_6) \leq Y_{min}(C_5) \leq Y_{max}(C_6)$$
$$Y_{min}(C_6) \leq Y_{max}(C_5) \leq Y_{max}(C_6)$$
$$Y_{min}(C_6) \leq Y_{min}(C_7) \leq Y_{max}(C_6)$$
$$Y_{min}(C_6) \leq Y_{max}(C_7) \leq Y_{max}(C_6)$$

For Class DD-MM-YY or DD.MM.YY

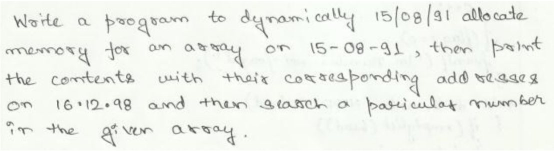
$$Y_{min}(C_2) \leq Y_{min}(C_3) \leq Y_{max}(C_2)$$
$$Y_{min}(C_2) \leq Y_{max}(C_3) \leq Y_{max}(C_2)$$
$$Y_{min}(C_4) \leq Y_{min}(C_3) \leq Y_{max}(C_4)$$
$$Y_{min}(C_4) \leq Y_{max}(C_3) \leq Y_{max}(C_4)$$
$$Y_{min}(C_5) \leq Y_{min}(C_6) \leq Y_{max}(C_5)$$
$$Y_{min}(C_5) \leq Y_{max}(C_6) \leq Y_{max}(C_5)$$
$$Y_{min}(C_7) \leq Y_{min}(C_6) \leq Y_{max}(C_7)$$
$$Y_{min}(C_7) \leq Y_{max}(C_6) \leq Y_{max}(C_7)$$

The above eight cases of inequalities (defined for each of the above two categories i.e. DD/MM/YY and DD-MM-YY or DD.MM.YY) are used to categorise a set of ECCC into the above defined date formats. A set of ECCC falls into either of the categories if and only if it satisfies all the eight conditions defining that class. Those sets of ECCC which do not fall into either of the above categories are rejected and are labelled as 'NON- DATE' sets.

5) Registering pixel locations:-

Once the set of ECCC is labelled as 'date', the pixel location range (i.e. a rectangle having the co-ordinates X_{min},$Y_{min}(C_i)$,$X_{min,}Y_{max}(C_i)$, $X_{max,}Y_{min}(C_{i+7})$, $X_{max,}Y_{max}(C_{i+7})$) is extracted. This region is now registered as 'date'. The output of a sample document (Figure IX) when processed is shown in Figure X.

The area localised is then sent for further classification if required (in case of DD-MM-YY and DD.MM.YY classes).



Write a program to dynamically 15/08/91 allocate
memory for an array on 15-08-91, then print
the contents with their corresponding addresses
on 16·12·98 and then search a particular number
in the given array.

Figure IX: showing the image of a sample document that is used as an input for the above algorithm.

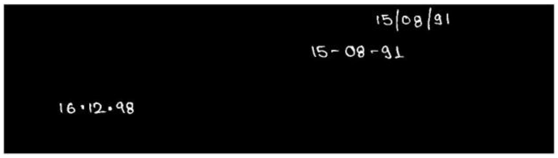


Figure X: showing the output image when the sample document (shown in Figure IX) is fetched as an input to our proposed algorithm (only the date fields are enunciated with pixel intensity value '1').

V. Further Classification Of DD-MM-YY And DD.MM.YY Formats (Or Classes)

Both these classes of dates share common spatial attributes, hence categorising them based on the above features (or conditions) is not possible. The only distinguishing factor among them is the 3rd and 6th element of the set of ECCC.

A feature vector comprising of elements W_{cc3} and W_{cc6} is defined; W_{cc3} and W_{cc6} denote the width of the 3rd and 6th connected component respectively. A database comprising of 246 handwritten dates is trained to classify these classes based on the feature vector defined. Then KNN classifier (with value K=3) is used to classify the testing data (result shown in Table I).

Table I: Enunciating the results of the K-NN classifier used to distinguish between DD-MM-YY and DD.MM.YY format.

No. of Documents	FAR (%)	FRR (%)	Efficiency (%)
75	3.86	1.43	94.71
150	3.39	1.38	95.23
246	2.66	1.06	96.28

VI. EXPERIMENT RESULTS

As mentioned earlier, the experiment was carried out (using Matlab 7.5.0.342, R2007b) on a database of 344 documents (157 of it were used for training and the remaining were used for testing). The results obtained are thus mentioned in a tabular form show in Table II.

Table II: Enunciating the results of date detection

No. of Documents	FAR (%)	FRR (%)	Efficiency (%)
50	12.00	6.00	82.00
100	10.00	4.00	86.00
187	9.09	3.20	87.71

VII. CONCLUSION AND FUTURE WORKS

The proposed algorithm shows quite an interesting result. It can be clearly seen (from table I) that FRR (False Rejection Ratio) is far less than that of FAR (False Acceptance Ratio), moreover the percentage of efficiency increases as the number of documents considered(for testing) is increased. The high percentage of FAR is due to cases as depicted by Figure XI. FRR is basically due to illegible handwriting, deviations from the normal patterns (or syntax) and occurrence of double digits (Figure XII).

Since the localisation technique does not involve any recognition process, so the overall algorithm could be rated as quite simple and fast. As mentioned earlier this prescribed algorithm could be modified to localise more classes of dates.

Future works include studying similar patterns among alpha-numeric date formats and addressing the failure in localising dates (numerical) pregnant with 'double digits'.

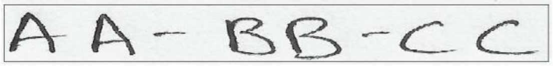


Figure XI: showing cases due to which FAR increases. The above script bears the same pattern as that of a date.

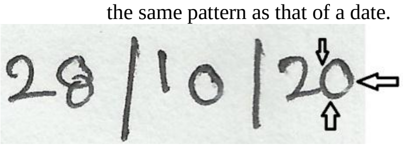


Figure XII: showing the case of double digits; the digits '2' and '0' are interconnected.

REFERENCE

[1] L. Lorette. "Handwriting recognition or reading? What is the situation at the dawn of the third millennium." IJDAR (1999), pp 2-12.

[2] Qizhi Xu et al , " Automatic Segmentation and Recognition System for Handwritten Dates on Canadian Bank Cheques", ICDAR 2003.

[3] L.Heutte et al . "Multi-bank check recognition system: consideration on the numerical amount recognition module", IJPRAI 11 (1997), pp 595-618.

[4] Marisa Mortia et al. "An HMM-based Approach for Date Recognition ", Proc of 4th International Workshop on Document Analysis System.

[5] Rodolfo . P. Dos Santos et al "Text Line Segmentation based on Morphology and Histogram Projection", 2009 10th International Conference on Document Analysis and Recognition.

[6] G.Koch, L. Heutte and T. Paquet "Numerical Sequence Extraction in Handwritten Incoming Mail Documents" ICDAR 2003.

Particle Swarm Optimization for Calibrating and Optimizing Xinanjiang Model Parameters

Kuok King Kuok
Lecturer, School of Engineering, Computing and Science, Swinburne University of Technology
Sarawak Campus, JalanSimpangTiga, 93350 Kuching, Sarawak, Malaysia

Chiu Po Chan
Lecturer, Faculty of Computer Science and Information Technology, University Malaysia
Sarawak, Kuching Samarahan Expressway, 94300
Kota Samarahan, Sarawak, Malaysia

Abstract— **The Xinanjiang model, a conceptual hydrological model is well known and widely used in China since 1970s. Therefore, most of the parameters in Xinanjiang model have been calibrated and pre-set according to different climate, dryness, wetness, humidity, topography for various catchment areas in China. However, Xinanjiang model is not applied in Malaysia yet and the optimal parameters are not known. The calibration of Xinanjiang model parameters through trial and error method required much time and effort to obtain better results. Therefore, Particle Swarm Optimization (PSO) is adopted to calibrate Xinanjiang model parameters automatically. In this paper, PSO algorithm is used to find the best set of parameters for both daily and hourly models. The selected study area is Bedup Basin, located at Samarahan Division, Sarawak, Malaysia. For daily model, input data used for model calibration was daily rainfall data Year 2001, and validated with data Year 1990, 1992, 2000, 2002 and 2003. A single storm event dated 9th to 12thOctober 2003 was used to calibrate hourly model and validated with 12 different storm events. The accuracy of the simulation results are measured using Coefficient of Correlation (*R*) and Nash-Sutcliffe Coefficient (E^2). Results show that PSO is able to optimize the 12 parameters of Xinanjiang model accurately. For daily model, the best *R* and E^2 for model calibration are found to be 0.775 and 0.715 respectively, and average *R*=0.622 and E^2=0.579 for validation set. Meanwhile, *R*=0.859 and E^2=0.892 are yielded when calibrating hourly model, and the average *R* and E^2 obtained are 0.705 and 0.647 respectively for validation set.**

Keywords - Conceptual rainfall-runoff model; Particle Swarm Optimization; Xinanjiang model calibration.

I. INTRODUCTION

Over the past half century, numerous hydrological models have been developed and applied extensively around the world. With the advent of digital computers in early 1960s, hydrologists began to develop sophisticated conceptual and physically hydrological models that are able to keep track of water movement using physical laws. One of the conceptual rainfall-runoff models developed is Xinanjiang model (Zhao *et al.*, 1980). Xinanjiang model has been successfully used in humid, semi-humid and even in dry areas mainly in China for flood forecasting since its initial development in the 1970s.

The main advantage and merit of Xinanjiang model is it can account for the spatial distribution of soil moisture storage (Liu *et al*., 2009). Generally, these spatial variations of hydrological variables are difficult to be considered (Chen *et al*., 2007). In recent decades, the distributed hydrological models have been increasingly applied to account for spatial variability of hydrological processes, to support impact assessment studies, and to develop rainfall-runoff simulations owing to their capability of explicit spatial representation of hydrological components and variables (Liu *et al.*, 2009).

In fact, no single model is perfect and best for solving all problems (Du*et al*., 2007; Das *et al*., 2008). The model performance can vary depending on model structure (distributed or lumped), physiographic characteristics of the basin, data available (resolution/accuracy/quantity), and also on how the relevant parameters are defined. Generally, Xinanjiang model consists of large number of parameters that cannot be directly obtained from measurable quantities of catchment characteristics, but only through model calibration. The aim of model calibration is to find the best set parameters values so that the model will be able to simulate the hydrological behavior of the catchment as closely as possible.

In fact, no single model is perfect and best for solving all problems (Du*et al*., 2007; Das *et al*., 2008). The model performance can vary depending on model structure (distributed or lumped), physiographic characteristics of the basin, data available (resolution/accuracy/quantity), and also on how the relevant parameters are defined.

Generally, Xinanjiang model consists of large number of parameters that cannot be directly obtained from measurable quantities of catchment characteristics, but only through model calibration. The aim of model calibration is to find the best set parameters values so that the model will be able to simulate the hydrological behavior of the catchment as closely as possible.

In early days, the model calibration was performed manually, which is tedious and time consuming due to the subjectivities involved. Besides, Xianjiang model is never applied in Malaysia, and the pioneer modeler is not confident to determine the best parameters values for using Xinanjiang model in Malaysia.

Therefore, it is necessary and useful to develop the computer based automatic calibration procedure. Some of the automatic optimization methods that have calibrated Xinanjiang model are genetic algorithm (Cheng *et al.*, 2006), shuffled complex evolution (SCE) algorithm (Duan *et al.*, 1992, 1994) and simulated annealing (Sumner *et al.*, 1997).

Among the Global Optimization Methods, Kuok (2010) found that Particle Swarm Optimization method (PSO) is more reliable and promising to provide the best fit between the observed and simulated runoff.

Xinanjiang model in Malaysia. Therefore, it is necessary and useful to develop the computer based automatic calibration procedure. Some of the automatic optimization methods that have calibrated Xinanjiang model are genetic algorithm (Cheng *et al.*, 2006), shuffled complex evolution (SCE) algorithm (Duan *et al.*, 1992, 1994) and simulated annealing (Sumner *et al.*, 1997). Among the Global Optimization Methods, Kuok (2010) found that Particle Swarm Optimization method (PSO) is more reliable and promising to provide the best fit between the observed and simulated runoff.

Even though PSO is simple in concept and easy to implement, the convergence speed is high and it is able to compute efficiently. Besides, PSO is also flexible and built with well-balanced mechanism for enhancing and adapting global and local exploration abilities (Abido, 2007). Thus, PSO is proposed to auto-calibrate Xinanjiang model in this paper.

Till to date, the application of PSO method in hydrology is still rare. Alexandre and Darrel (2006) applied multi-objective particle swarm optimization (MOPSO) algorithm for finding non-dominated (Pareto) solutions when minimizing deviations from outflow water quality targets. Bong and Bryan (2006) used PSO to optimize the preliminary selection, sizing and placement of hydraulic devices in a pipeline system in order to control its transient response. Janga and Nagesh (2007) used multi-objective particle swarm optimization (MOPSO) approach to generate Pareto-optimal solutions for reservoir operation problems. Kuok (2010) also adapted PSO to auto-calibrate the Tank model parameters.

II. Study Area

The selected study area is Bedup basin, located approximately 80km from Kuching City, Sarawak, Malaysia. The catchment area of Bedup basin is approximately 47.5km^2, which is mainly covered with shrubs, low plant and forest. The elevation are varies from 8m to 686m above mean sea level (JUPEM, 1975). The historical record shows that there is no significant land used change over the past 30 years. Bedup River is approximately 10km in length. Bedup basin is mostly covered with clayey soils. Thus, most of the precipitation fails to infiltrate, runs over the soil surface and produces surface runoff. Part of Bedup basin is covered with coarse loamy soil, thus producing moderately low runoff potential.

Bedup River is located at upper stream of Batang Sadong. It is not influence by tidal and the rating curve equation for Bedup basin is represented by Equation 1 (DID, 2007).

$$Q=9.19(H)^{1.9} \tag{1}$$

Where Q is the discharge (m^3/s) and H is the stage height (m). These observed runoff data were used to compare the model runoff.

Fig.1 presents the locality plan of Bedup basin. Sadong basin is located at southern region of Sarawak and Bedup basin is located at the upper catchment of Sadong basin. The five rainfall stations are Bukit Matuh (BM), Semuja Nonok (SN), Sungai Busit (SB), Sungai Merang (SM) and Sungai Teb (ST), and one river stage gauging station at Sungai Bedup. All these gauging stations are installed by Department of Irrigation and Drainage (DID) Sarawak.

Daily and hourly areal rainfall data obtained through Thiessen Polygon Analysis are fed into Xinanjiang model for model calibration and validation. The area weighted precipitation for BM, SN, SB, SM, ST are found to be 0.17, 0.16, 0.17, 0.18 and 0.32 respectively. Thereafter, the calibrated Xinanjiang model will carry out computation to simulate the daily and hourly discharge at Bedup outlet.

III. Xinanjiang Model Algorithms

Xinanjiang model was first developed in 1973 and published in English in 1980 (Zhao *et al*., 1980). It is a lumped hydrological model that required stream discharge and meteorological data.

The basic concept of Xinanjiang model is runoff only generated at a point when the infiltration reached the soil moisture capacity (Zhao, 1983, 1992). A parabolic curve of FC (refer Fig. 2) is used to represent the spatial distribution of the soil moisture storage capacity over the basin (Zhao *et al*., 1980):

$$\frac{f}{F} = 1 - \left(1 - \frac{WM\prime}{WMM}\right)^{b} \tag{2}$$

where $WM'$is the FC at a point that varies from zero to the maximum of the whole watershed WMM. Larger WM'means larger soil moisture storage capacity in a local area and more difficult runoff generation.

Parameter b represents the spatial heterogeneity of FC (Zhao, 1983, 1992). For uniform distribution, b always equal to zero. In contrast, large b represents significant spatial variation. The b parameter is usually determined by model calibration.

Fig.2 presents $\frac{f}{F}$ versus WM'curve. The watershed average FC (*WM*), is the integral of $\left(1-\frac{f}{F}\right)$ between WM'=0 and WM'=WMM, as represented by Equation 3.

$$WM = \frac{WMM}{(1+b)} \tag{3}$$

Meanwhile, the watershed average soil moisture storage at time *t* (W_t), is the integral of $\left(1-\frac{f}{F}\right)$, between zero and WM_t^*, which is a critical FC at time t as presented in Equation 4 and Fig.2:

$$W_t = \int_0^{WM_t^*} \left(1 - \frac{f}{F}\right) d(WM') \\ = WM\left[1 - \left(1 - \frac{WM_t^*}{WMM}\right)^{1+b}\right] \tag{4}$$

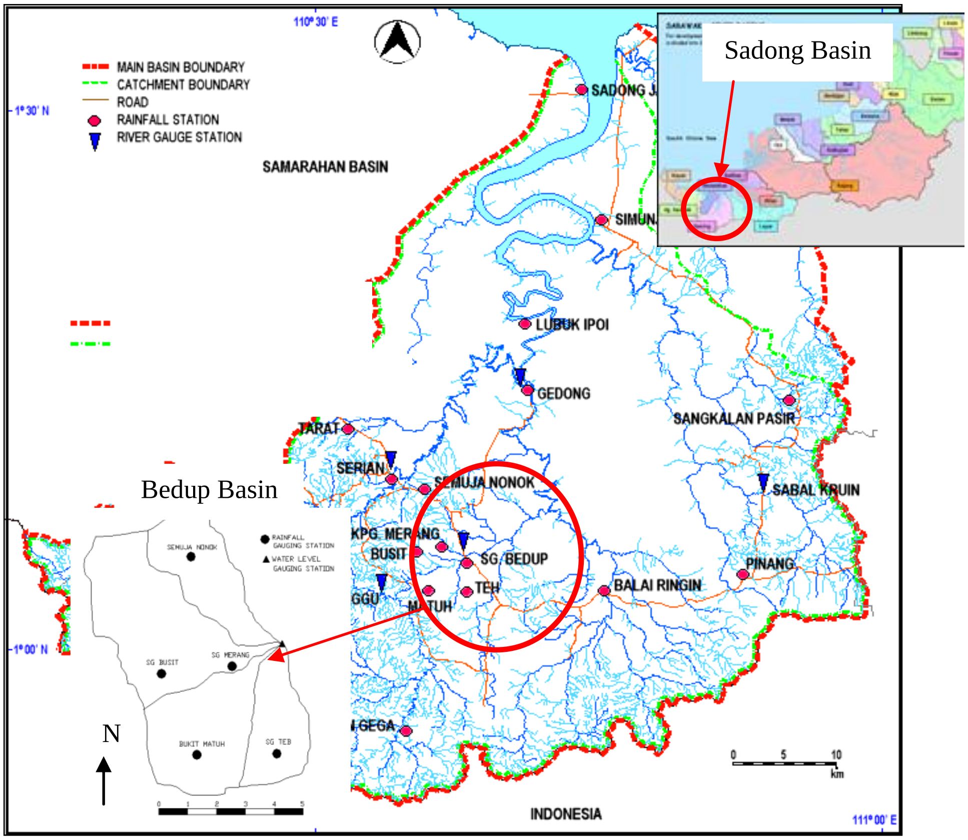


Fig 1: Locality map of Bedup basin, Sub-basin of Sadong basin, Sarawak

The critical FC (WM_t^*) corresponding to watershed average soil moisture storage (W_t) is presented in Equation 5.

$$WM_t^* = WMM\left[1-\left(1-\frac{W_t}{WM}\right)^{\frac{1}{1+b}}\right] \qquad (5)$$

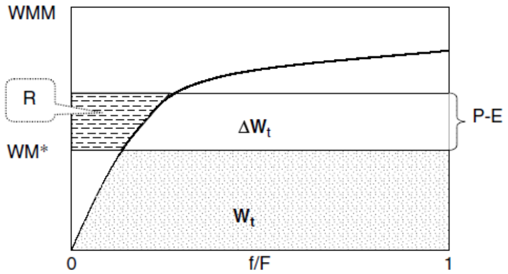


Fig. 2: FC curve of soil moisture and rainfall–runoff relationship.
Note: WMM is maximum FC in a watershed; f/F is a fraction of the watershed area in excess of FC; WM_t^* is FC at a point in the watershed; Rt is runoff yield at time t; ΔWt is soil moisture storage deficit at time t and is equal to WM-Wt ; Wt is watershed-average soil moisture storage at time t

When rainfall (P_t) exceeds evapotranspiration (E_t), P_t is infiltrated into soil reservoir. Runoff (R_t) will only be produced when the soil reservoir is saturated (soil moisture reaches FC). As shown in Fig. 2, if the net rainfall amount (rainfall minus actual evapotranspiration) in a time interval [*t* - 1, *t*] is *Pt*–*Et* and initial watershed average soil moisture (tension water) is *Wt*, the runoff yield in the time interval *Rt* can be calculated as follows:

$$\begin{aligned}
&If\ P_t - E_t - WM_t^* < WMM\\
&R_t = P_t - E_t - \Delta W_t\\
&= P_t - E_t - \int_{WM_t^*}^{P_t-E_t+WM_t^*}\left(1-\frac{f}{F}\right)d(WM')\\
&= P_t - E_t - WM + W_t\\
&\qquad + WM[1\\
&\qquad - (P_t - E_t - WM_t^*)/WMM]^{1+b}\\
&If\ P_t - E_t - WM_t^* \geq WMM\\
&R_t = P_t - E_t - WM + W_t
\end{aligned}$$

The original Xinanjiang model is divided into two components named as runoff generating component and runoff routing component. Basin is divided into series of sub-areas, and runoff is calculated from water balance component. The runoff from each sub-area is routed to the main basin outlet using Muskingum method. However, runoff generating and runoff routing components are combined together in this study as shown in Fig. 3. There are 12 parameters to be calibrated include S, Dt, K, C, B, Im, Sm, Ex, Ki, Kg, Ci and Cg. The model parameters are listed in Table 1. During the calibration, the parameter must satisfy the constraints of the Muskingum method for each channel of sub-basin.

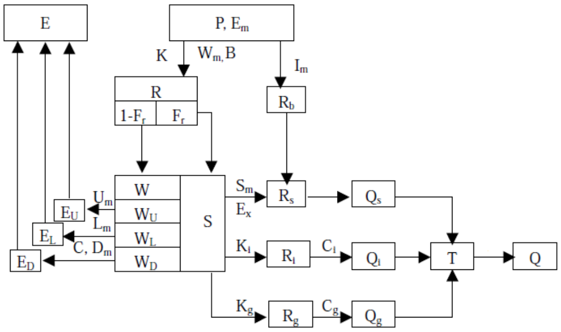


Fig.3: Flowchart of Xinanjiang Model

PSO algorithm was developed by Kennedy and Eberhart (1995). It is a simple group-based stochastic optimization technique, initialized with a group of random particles (solutions) that were assigned with random positions and velocities. The algorithm searches for optima through a series of iterations where the particles are flown through the hyperspace searching for potential solutions. These particles learn over time in response to their own experience and the experience of the other particles in their group (Ferguson, 2004). Each particle keeps track of its best fitness position in hyperspace that has achieved so far (Eberhart and Shi, 2001). For each iteration, every particle is accelerated towards its own personal best, in the direction of global best position and the fitness value for each particle's is evaluated. This is achieved by calculating a new velocity term for each particle based on the distance from its personal best, as well as its distance from the global best position.

Once the best value the particle has achieved, the particle stores the location of that value as "pbest" (particle best). The location of the best fitness value achieved by any particle during any iteration is stored as "gbest" (global best). The basic PSO procedure was shown in Fig. 4.

The particle velocity is calculated using Equation6.

IV. Particle Swarm Optimization (PSO) Algorithm

$$V_i = \omega V_{i-1} + c_1 \text{*rand()*(pbest-presLocation)}$$

$$\text{+}c_2\text{*rand()*(gbest-presLocation)} \qquad (6)$$

Table 1: Parameters for Xinanjiang Model

Notation	Definition
S	Depth of free surface water flow
Dt	Time interval
K	Ratio of potential evapotranspiration to pan evaporation
C	Coefficient of the deep layer, that depends on the proportion of the basin area covered by vegetation with deep roots
B	Exponential parameter with a single parabolic curve, which represents the non-uniformity of the spatial distribution of the soil moisture storage capacity over the catchment
Im	Percentage of impervious and saturated areas in the catchment
Sm	Areal mean free water capacity of the surface soil layer, which represents the maximum possible deficit of free water storage
Ex	Exponent of the free water capacity curve influencing the development of the saturated area
Ki	Outflow coefficients of the free water storage to interflow relationships
Kg	Outflow coefficients of the free water storage to groundwater relationships
Ci	Recession constants of the lower interflow storage
Cg	Recession constants of the groundwater storage

The particle position is updated according to Equation7.

$$\text{presLocation=prevLocation+}V_i \qquad (7)$$

where V_i is current velocity, ω is inertia weight, V_{i-1} is previous velocity, *presLocation* is present location of the particle, *prevLocation* is previous location of the particle and *rand()* is a random number between (0, 1). c_1 and c_2 are acceleration constant for gbest and pbest respectively.

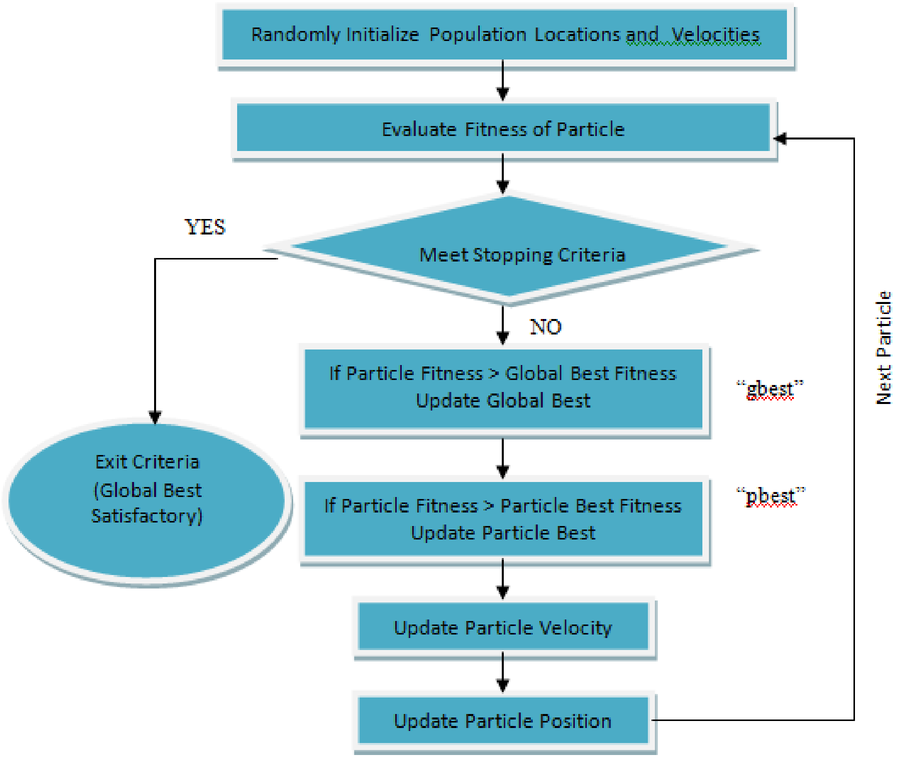


Fig. 4: Basic PSO Procedure.

V. Model Calibration And Validation

The basic calibration procedure for Xinanjiang model using PSO algorithm for both daily and hourly runoff simulation is presented in Fig. 5.

A. *Daily Model*

The Xinanjiang model for Bedup basin is calibrated with daily rainfall-runoff data Year 2001. Since the model is firstly used in Malaysia, the best parameters values are not known. Therefore, all the 12 Xinanjiang model parameters (S, Dt, K, C, B, Im, Sm, Ex, Ki, Kg, Ci and Cg) either they are related to the average climate or surface conditions of the studied region, are calibrated automatically using PSO algorithm.

At the early stage of the calibration, the parameters of PSO that will affect the calibration results are pre-set. Various sets of daily rainfall-runoff data are calibrated to find the best model configuration for simulating daily runoff. The objective function used is Root Mean Square Error (RMSE). As the calibration process is going on, the initial parameters that set previously are changed to make the simulated runoff matching the observed one. The PSO parameters investigated are:

a) Different acceleration constant for gbest (c_1) ranging from 0.5 to 2.0
b) Different acceleration constant for pbest (c_2) ranging from 0.5 to 2.0
c) Max iteration of 100, 125, 150, 175 and 200
d) 100, 125, 150, 175, 200, 225, 250, 275 and 300 number of particles

Input data series to the Xinanjiang model are daily average areal rainfall calculated using Thiesen Polygon method. Daily data from 1stJanuary 2001 to 31stDecember 2001 are used for model calibration. The model is then validated with rainfall-runoff data Year 1990, 1992, 2000, 2002 and 2003. The details of data used for model validation are presented in Table 2.

Table 2: Daily Validation Data

	Validation Daily Data Set
1	1stJanuary 1990 to 31stDecember 1990
2	1stJanuary 1992 to 31stDecember 1992
3	1stJanuary 2000 to 31stDecember 2000
4	1stJanuary 2002 to 31stDecember 2002
5	1stJanuary 2003 to 31stDecember 2003

B. *Hourly Model*

Similarly, all 12 Xinanjiang model parameters including S, Dt, K, C, B, Im, Sm, Ex, Ki, Kg, Ci and Cg are calibrated automatically using PSO algorithm for hourly runoff simulation. The objective function used is Root Mean Square Error (RMSE). PSO algorithm parameters investigated are including:

a) Different acceleration constant for gbest (c_1) ranging from 0.1 to 2.0
b) Different acceleration constant for pbest (c_2) ranging from 0.1 to 2.0
c) Max iteration of 100, 125, 150, 175 and 200
d) 100, 125, 150, 175, 200, 225, 250, 275 and 300 number of particles

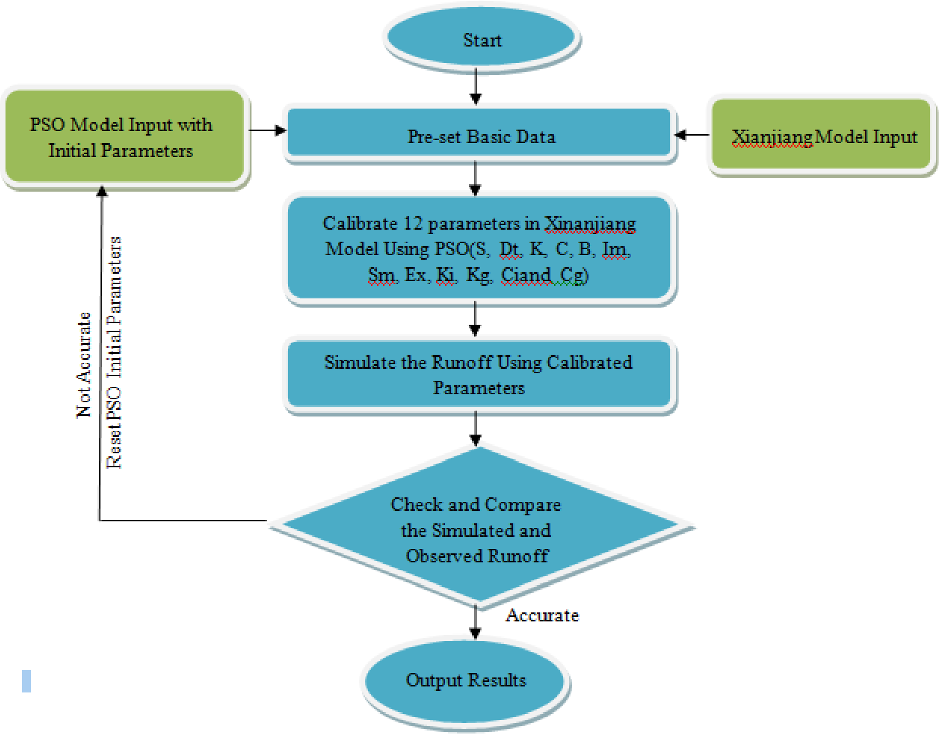

Start
PSO Model Input with Initial Parameters
Pre-set Basic Data
Xianjiang Model Input
Calibrate 12 parameters in Xinanjiang Model Using PSO(S, Dt, K, C, B, Im, Sm, Ex, Ki, Kg, Ciand Cg)
Simulate the Runoff Using Calibrated Parameters
Check and Compare the Simulated and Observed Runoff
Not Accurate
Reset PSO Initial Parameters
Accurate
Output Results

Fig.5: Calibration procedure

An average areal rainfall single storm event dated 9th to 12th October 2003 is used to calibrate and optimize Xinanjiang model parameters. Once obtained the optimal parameters, the model will be validated with 12 single storm events. The details of validation storm events are presented in Table 3.

Table 3: Hourly Validation Data

	Validation Daily Data Set
1	5th to 8th April 2000
2	26th to 31st January 1999
3	20th to 24thJanuary 1999
4	5th to 8thFebruary 1999
5	1st to 4thMarch 2002
6	11th to 15thDecember 2003
7	22nd to 25thNovember 2001
8	4th to 8thJanuary 2003
9	15th to 18thApril 2002
10	8th to 12thDecember 2004
11	17th to 21stDecember 2002
12	14th to 19thFebruary 2002

V.III Performance Measurement

The accuracy of the simulation results are measured using Coefficient of Correlation (*R*) and Nash-sutcliffe coefficient (E^2). *R* and E^2 are measuring the overall differences between observed and simulated flow values. The closer *R* and E^2 to 1, the better the predictions are. The formulas of *R* and E^2 are presented in Equations8 and 9 respectively.

$$R = \frac{\sum(obs-\overline{obs})(pred-\overline{pred})}{\sqrt{\sum(obs-\overline{obs})^2\sum(pred-\overline{pred})^2}} \qquad (8)$$

$$E^2 = 1 - \frac{\sum(obs-\overline{pred})^2}{\sum(obs-\overline{obs})^2} \qquad (9)$$

where *obs* = observed value, *pred* = predicted value, $\overline{obs}$ = mean observed values and$\overline{pred}$= mean predicted values.

VI. Results And Discussion

A. Daily ResulT

PSO algorithm achieved the optimal configuration at the RMSE of 2.3003 for daily model. The optimal configuration for PSO algorithm was found to be 200 number of particles, max iteration of 150 and c_1=1.8 and c_2=1.8. The best *R* and E^2 obtained for calibration set were found to be 0.775 and 0.715 respectively as presented in Fig. 6. The 12 parameters of Xinanjiang model optimized by PSO algorithm can be found in Table 4.

The results showed that runoff generated by Xinanjiang model optimized by PSO algorithm is controlled and dominant to 8 parameters named as S, B, Im, Sm, Ex, Ki, Kg and Ci. In contrast, Dt, K, C and Cg are less sensitive to storm hydrograph generation.

Fig. 7 shows the validation results when the optimal configuration of Xinanjiang model optimized by PSO algorithm. As *R* is referred, the results obtained for Year 2000, 2003, 2002, 1992 and 1990 are found to be 0.674, 0.649, 0.616, 0.616, 0.553 and 0.622 respectively. As E^2 is used as level mark, the E^2 obtained are ranging from 0.550 to 0.623. The average *R* and E^2 are yielding to 0.622 and 0.579 respectively.

Table 4: Optimized parameters for daily model

Parameters	Values
S	5.1424
Dt	0.00001
K	0.00001
C	0.00001
B	0.0772
Im	0.1542
Sm	30.2411
Ex	27.8412
Ki	0.0521
Kg	6.3272
Ci	7.4719
Cg	0.00001

B. Hourly Results

For hourly runoff calibration, the optimal configuration of PSO was found to be c_1= 0.6, c_2= 0.6, 200 number of particles and max iteration of 150. The best R and E^2 obtained for calibration set were found to be 0.859 and 0.892 respectively (as presented in Fig. 8). RMSE obtained by optimal configuration of PSO algorithm was 2.6303. Optimal 12 parameters of Xinanjiang model obtained for hourly runoff simulation were tabulated in Table 5.

Table 5: Optimized parameters for hourly model

Parameters	Values
S	20.0810
Dt	0.00001
K	0.2309
C	0.6296
B	0.00001
Im	13.3202
Sm	7.6331
Ex	1.5781
Ki	1.9105
Kg	4.2626
Ci	17.3510
Cg	0.00001

The results indicated that hourly runoff produced by optimized Xinanjiang model is dominant to 9 parameters. These 9 dominant parameters are S, K, C,Im, Sm, Ex, Ki, Kg and Ci. Contrary, parameters Dt, B and Cg show less sensitive to storm hydrograph generation.

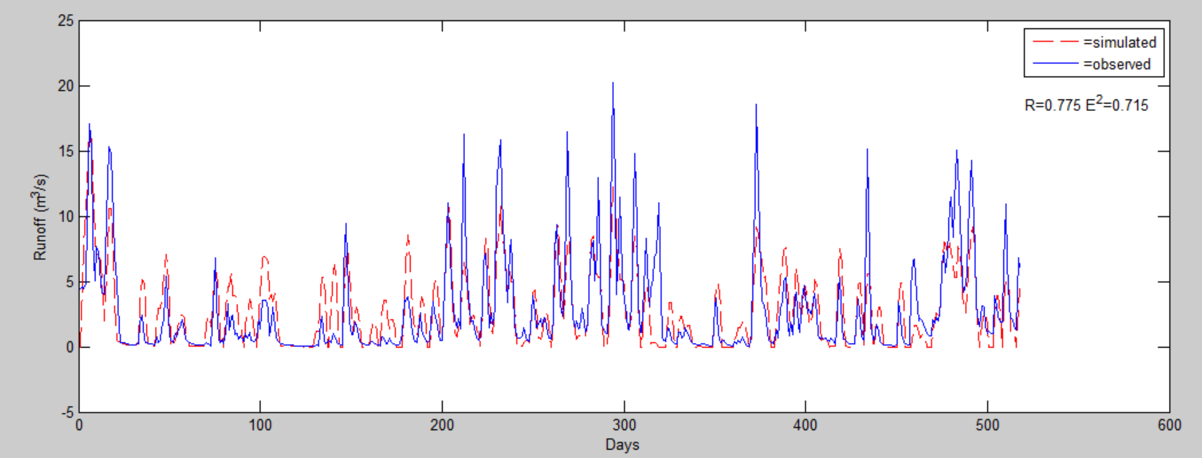


Fig. 6: Comparison between observed and simulated runoff generated by daily Xinanjiang model optimized with PSO algorithm.

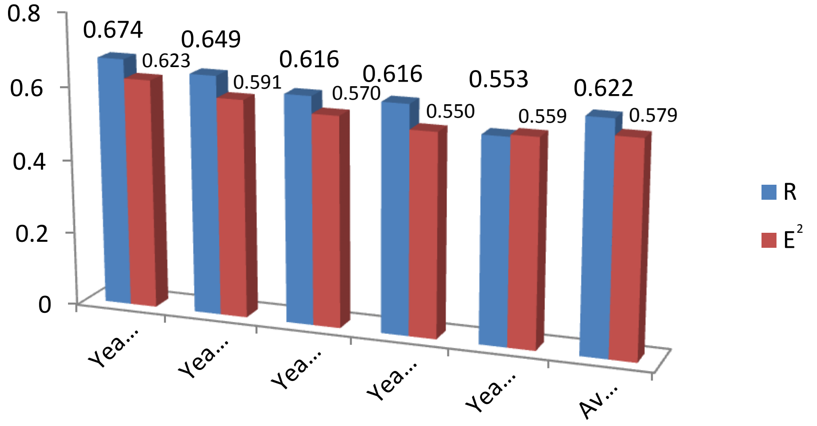


Figure 7: Daily model validation results

As optimal configuration of Xinanjiang model validated with 12 different events, the *R* values obtained are ranging from 0.552 to 0.854, whilst 0.510 to 0.763 for E^2. The average *R* and E^2 for validated storm events are 0.705 and 0.647 respectively. The validation results are presented in Fig. 9.

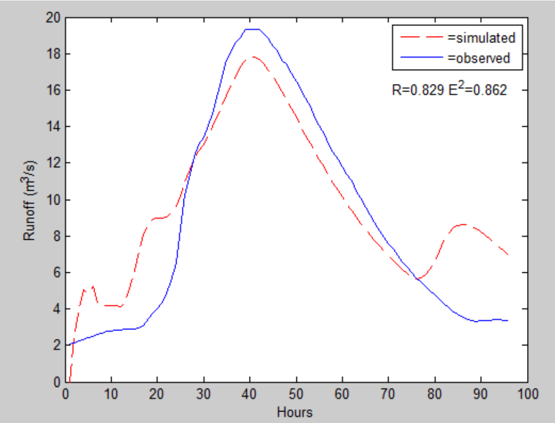


Fig. 8: Comparison between observed and simulated hourly runoff generated by Xinanjiang model optimized with PSO algorithm.

VII. Conclusion

A general framework for automatic calibration of Xinanjiang model using PSO algorithm has been successfully demonstrated for Bedup Basin, Malaysia for both daily and hourly runoff generation. The framework includes model parameterisation, choice of calibration parameters and the optimization algorithm. In this study, PSO proved its promising abilities to calibrate and optimize 12 parameters of Xinanjiang model accurately. For daily model calibration, PSO had achieved *R*=0.775 and E^2=0.715 with optimal model configuration of c_1=1.8, c_2=1.8, 200 number of particles and 150 max iteration. Besides, optimal configuration of c_1=0.6, c_2=0.6, 200 number of particles and 150 max iteration also yielded *R* and E^2 to 0.859 and 0.892 respectively for calibration of hourly model.

These results show that the newly developed PSO algorithm is able to calibrate and optimize 12 parameters of Xinanjiang model accurately. Besides, PSO had shown its robustness by validating 5 different sets of rainfall-runoff data by yielding average *R* and E^2 to 0.622 and 0.579 respectively for daily runoff simulation, and average *R*=0.705 and E^2=0.647 for hourly runoff validation.

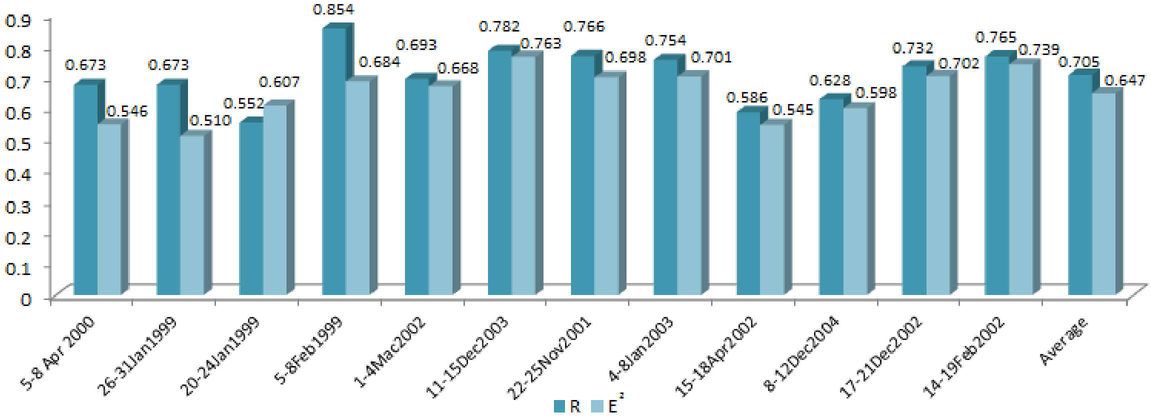


Figure 9: Hourly model validation results

These indicated that PSO optimization search method is a simple algorithm, but proved to be robust, efficient and effective in searching optimal Xinanjiang model parameters. This was totally revealed by the ability of PSO methods in searching the optimal parameters that provided the best fit between observed and simulated flows.

ACKNOWLEDGEMENTS

The authors would like to express their sincere thanks to Professor Chun-Tian Cheng from Institute of Hydropower & Hydroinformatics, Department of Hydraulic Engineering, Dalian University of Technology and Associate Professor Jintao Liu, Laboratory of Hydrology, Water Resources and Hydraulic Engineering, Hohai University, Nanjing for providing the source code of Xinanjiang model.

REFERENCES

[1] Abido, M. A. (2007) Two-Level of Nondominated Solutions Approach to Multiobjective Particle Swarm Optimization.ACM, GECCO’07, London, England, United Kingdom.

[2] Alexandre, M. B. and Darrell, G. F. (2006). A Generalized Multiobjective Particle Swarm Optimization Solver for Spreadsheet Models: Application to Water Quality. Hydrology Days 2006, 1-12

[3] Bong, S. K.; Bryan, W. K. (2006). Hydraulic optimization of transient protection devices using GA and PSO approaches. J. Water Res. Plan. Manage., 132 (1), 44-52

[4] Chen X, Chen YD, Xu CY. (2007). A distributed monthly hydrological model for integrating spatial variations of basin topography and rainfall.Hydrological Processes **21**: 242–252.

[5] Cheng CT, Zhao MY, Chau KW and Wu XY (2006).Using genetic algorithm and TOPSIS for Xinanjiang model calibration with a single procedure.Journal of Hydrology 316 (2006) 129–140.

[6] Das T, B´ardossy A, Zehe E, He E. (2008).Comparison of conceptual model performance using different representations of spatial variability.Journal of Hydrology 356: 106–118.

[7] DID (2007).Hydrological Year Book Year 2007.Department of Drainage and Irrigation Sarawak, Malaysia.

[8] Du JK, Xie SP, Xu YP, Xu CY, Singh VP. (2007). Development and testing of a simple physically-based distributed rainfall-runoff model for storm runoff simulation in humid forested basins.Journal of Hydrology 336:334–346.

[9] Duan, Q., Sorooshian, S., Gupta, V. (1994). Optimal use of the SCEUA global optimization method for calibrating watershed models.Journal of Hydrology 158, 265–284.

[10] Eberhart, R.; Shi, Y., (2001). Particle swarm optimization developments, Application and Resources. IEEE, 1, 81-86

[11] Ferguson, D., (2004). Particle swarm. University of Victoria, Canada.

[12] Kennedy, J.; Eberhart, R. C., (1995).Particle swarm optimization.Proceedings of the IEEE international joint conference on neural networks, IEEE Press.1942–1948.

[13] Kuok K. K. (2010). Parameter Optimization Methods for Calibrating Tank Model and Neural Network Model for Rainfall-runoff Modeling. Ph.D. Thesis. University Technology Malaysia, 2010.

[14] Liu JT, Chen X, Zhang JB and M. Flury. (2009) Coupling the Xinanjiang model to a kinematic flow model based on digital drainage networks for flood forecasting. Hydrological Processes.23, 1337–1348.

[15] Janga, M. R. and Nagesh, D. K. (2007).Multi-Objective Particle Swarm Optimization for Generating Optimal Trade-Offs in Reservoir Operation.Hydrological Processes. 21: 2897–2909. Published online 10 January 2007 in Wiley InterScience

[16] JUPEM (1975). Jabatan Ukur dan Pemetaan Malaysia. Scale 1:50,000.

[17] Sumner, N.R., Fleming, P.M., Bates, B.C. (1997). Calibration of a modified SFB model for twenty-five Australian catchments using simulated annealing. Journal of Hydrology 197, 166–188.

[18] Zhao RJ, Zhuang YL, Fang LR, Liu XR, Zhang QS.(1980). The Xinanjiangmodel.InHydrological Forecasting, IAHS Publication No.129. IAHS Press: Wallingford; 351–356.

[19] Zhao RJ. 1983. Watershed Hydrological Model—Xinanjiang Model and Shanbei Model. Water & Power Press: Beijing, (in Chinese).

[20] Zhao RJ. 1992. The Xinanjiang model applied in China. Journal of Hydrology **135**: 371–381.

AUTHORS PROFILE

Dr Kuok King Kuok holds PhD in Hydrology and Water Resources and Bachelor of Civil Engineering with honours, both from University Technology Malaysia, Master of Engineering major in Hydrology from University Malaysia Sarawak. He is also a Professional Engineer registered with Board of Engineers Malaysia, EMF International Professional Engineer (MY), Asean Chartered Professional Engineer. He is also a corporate member of Institution of Engineers Malaysia, ASEAN Engineer and APEC Engineer. He has authored and co-authored more than 30 national and international conference and journal papers. Currently he is lecturing at Swinburne University of Technology Sarawak Campus and also practicing as design engineer.

Chiu Po Chan graduated with Bachelor of Information Technology major in Software Engineering, and Master of Science in Computer Science, both from University Malaysia Sarawak. She has authored and co-authored more than 20 national and international conference and journal papers, mainly in application of artificial intelligence in hydrology and water resources. Currently, she is lecturing in Faculty of Computer Science and Information Technology, University Malaysia Sarawak.

Probabilistic: A Fuzzy Logic-Based Distance Broadcasting Scheme For Mobile Ad Hoc Networks

Tasneem Bano
Computer Science and Engineering,
Maulana Azad National Institute of Technology
Bhopal, Madhya Pradesh 462051, India

Jyoti Singhai
Electronics and Engineering,
Maulana Azad National Institute of technology
Bhopal, Madhya Pradesh 462051, India

***Abstract*—An on-demand route discovery method in mobile ad hoc networks (MANET) uses simple flooding method, whereas a mobile node blindly rebroadcasts received route request (RREQ) packets until a route to a particular destination is established. Thus, this leads to broadcast storm problem. This paper presents a novel algorithm for broadcasting scheme in wireless ad hoc networks using a fuzzy logic system at each node to determine its capability to broadcast route request packets, based on the node location. Our simulation analysis shows a significant improvement in performance in terms of routing overhead, MAC collisions and end-to-end delay while still achieving a good throughput compared to the traditional AODV.**

Keywords Broadcasting; Distance based broadcasting; Fuzzy; Optimization Technique; AODV.

I. INTRODUCTION

Traditionally, broadcasting means sending a message from one given node (the source station) to all the nodes in the network. In a (multi-hop) decentralized network, the broadcasted data has to be relayed by intermediate nodes in such a way that the entire network graph is spanned. In MANET, simplistic broadcast schemes result in network with redundancy, contention, and collision which is often called 'Broadcast Storm Problem'. This can prevent broadcasts from achieving the objectives of optimal delivery ratio, energy balancing, and latency.

The main objective of a broadcasting scheme is to avoid broadcast storm problems and to provide good network performance and scalability. Therefore, a route discovery technique that can guarantee an efficient utilization of the limited system resources while achieving acceptable levels of other important performance metrics such as throughput and end-to-end delay is highly desirable. Till date, research on efficient broadcasting schemes in mobile ad-hoc networks has proceeded along two main schemes:

A. Deterministic Schemes

B. Probabilistic Schemes

Deterministic Schemes use network topological information to build a virtual backbone that covers all the nodes in the network. In order to build a virtual backbone, nodes exchange information, typically about their immediate or two hop neighbors. This results in a large overhead in terms of time and message complexity for building and maintaining the backbone, especially in the presence of mobility.

Probabilistic Schemes, however, rebuild a backbone from scratch during each broadcast. Nodes make instantaneous local decisions about whether to broadcast a message or not using information derived only from overheard broadcast messages. These schemes incur a smaller overhead and demonstrate superior adaptability in changing environments when compared to deterministic schemes [1, 2, 4, 17].

The rest of the paper is organized as follows. Section II presents related work on some route discovery techniques. Section III presents analysis of node location, while section IV gives an introduction of fuzzy logic based distance route discovery method. Section V and VI gives simulation parameter, conducts a comparison and performance evaluation of the proposed route discovery methods. Finally, Section VII concludes the study and scope of future research work.

II. RELATED WORK

One of the earliest broadcast mechanisms in both wired and wireless networks is flooding, where every node in the network retransmits a message to its neighbors upon receiving it for the first time. Although flooding is simple and easy to implement, it can be costly in terms of network performance, and may lead to a serious problem, often known as the *broadcast storm problem* [4, 5, 7]. The broadcast problem is then characterized by high redundant message retransmissions, network bandwidth contention, and collision. Ni [4] have studied the flooding protocol and the results obtained have shown that rebroadcast could provide at most 61% additional coverage and only 41% additional coverage in average over that already covered by the previous broadcast. As a result, they have concluded that rebroadcasts are very costly and should be used with caution.

Probabilistic broadcasting is one of the simplest and most efficient broadcast techniques that have been suggested [6] in the literature. The advantage of probabilistic broadcasting over the other proposed broadcast methods [6, 12 and 13] is its simplicity. However, studies [6, 11] have shown that although probabilistic broadcast schemes can significantly reduce the degrading effects of the broadcast storm problem [6], they suffer from poor reachability, especially in a sparse network topology. In this approach each intermediate node rebroadcasts received packets only with a predetermined forwarding probability. Clearly, the appropriate choice of the forwarding probability determines the effectiveness of this technique.

Most probabilistic broadcast approaches that have been proposed in the literature [6, 8 and 11] have considered a fixed forwarding probability at each intermediate node. This could lead to most nodes not receiving the broadcast packet when the forwarding probability is set too low or more redundant transmissions if the probability is set too high, as discussed in [9, 10]. One of the causes for this stems from the fact that every node in the network has the same probability of rebroadcast, regardless of its local topological characteristics, such as neighboring node location. In a dense network multiple nodes may share similar transmission coverage. Therefore, if some nodes, randomly, do not forward the broadcast packet, these could save resources without degrading the delivery effectiveness. On the other hand, in a sparse network, there is much less shared coverage; thus some nodes might not receive the broadcast packet unless the rebroadcast probability is set high enough. Consequently, the rebroadcast probability should be set differently from one node to another according to their local topological characteristics.

The author in [15] introduces the concept of distance into the counter based scheme, in which nodes closer to the border of source has been given higher rebroadcast probability since they create better expected additional coverage area [6]. Distance threshold is taken to distinguish between interior and border nodes. Two distinct random assessment delays are applied to the border and interior nodes, with border nodes having shorter random assessment delay than the interior nodes.

This paper proposes a new route discovery algorithm that utilizes probabilistic broadcast methods using fuzzy logic to disseminate the RREQ packets. To evaluate the new route discovery methods we have considered using the AODV [14] routing algorithm and have compared the performance of fuzzy logic based broadcasting with the Distance-Aware Counter-Based Broadcast Scheme [15]. Our results reveal that equipping AODV [14] with the new fuzzy logic based probabilistic route discovery methods help to reduce the overall routing overhead while achieving good network throughput with improved end-to-end delay when compared to the traditional AODV [14] and [15].

I. Analysis Of Node Location

In Mobile ad hoc networks node location is one of the most important aspects in broadcasting. Each node has its transmission range within which its neighbor can receive broadcasting information [16]. Consider a simple scenario in figure below

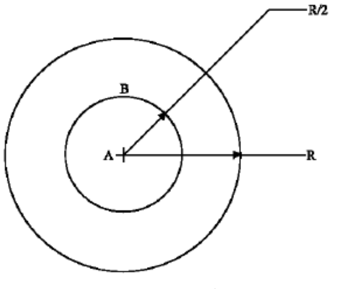


Figure 1. Coverage area of node A and B

Host a sends a broadcast message and host B decides to rebroadcast the message. Let C_A and C_B denotes the circle area covered by A's and B's transmission, respectively. The additional area provided by B's rebroadcast is denoted by C_{B-A}. Let r be the radii of C_A and C_B and d the distance between A and B. than $|C_{B-A}|$ = $|C_A|$- $|C_{A\cap B}|$ = πr^2- INTC (d), where INTC (d) is the intersection area of the two circles centered at 2 points discussed by d:-

INTC(d) = 4 $\int_{d/2}^{r} \sqrt{r^2 + r^2}$ dx

When d=r , the coverage area $|C_{B-A}|$ is the largest which equals πr^2- INTC (d) = 0.61 πr^2. This shows that 61% of additional coverage can be provided over that already covered by the previous transmission[6].

And when d< r that is B is located in As' transmission range, then the average value can be obtained by integrating the above value over the circle of radius x centered at A for x in [0,r]

$\int_0^r \frac{2\pi x\ [\pi r^2 - INTC(x)]}{\pi r^2}$ dx ~ 0.41 πr^2

So now a broadcast can cover only additional 41% area in average.

Now, by considering the expected additional coverage area of a node, different broadcasting probability can be set for nodes with d=r [border nodes] and d<r [interior nodes]. In order to distinguish these two types of node we introduce Dth [distance threshold] =200 when R[transmission radius] =250 . To calculate P, we need the relative distance between nodes D_{AB} between nodes A and B then

Pi=$\left(\frac{D_{AB}}{R}\right)^n$× 100 (1)

When n=0 the scheme is simple flooding, if n=1, the scheme broadcast with Pi if it receives the packet for first time, otherwise discard the packet. Now when n>1, Pi increases exponentially, it makes retransmit nodes concentrate towards the border of source nodes coverage area, which results in the increasing of the EAC area of next hop and less rebroadcasts. The concentration increases with n values. The larger n is selected, the more concentration of retransmission nodes to the border. The value of n is selected based on the network densities. Now varying the value of n from 0,1 and >1 and see the effect on the redundant area.

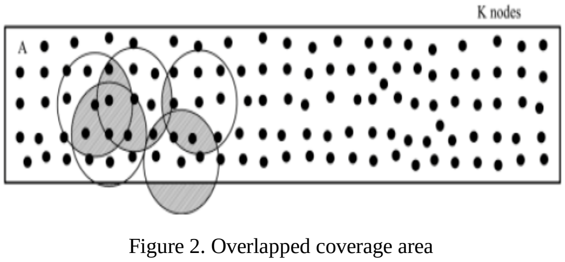


Figure 2. Overlapped coverage area

Let S is the area of our topology area A. there are m nodes in A and transmission radius is R. then every nodes' coverage area is Si= πr^2. Now when n=0

$S_Z{}^1 = \sum_{i=1}^{m} Si$

Then,

$S_Z{}^1 = \sum_{i=1}^{m} \pi R^2 = \text{m}\pi R^2$ (2)

Now in above figure there is overlapping in the coverage area as the network is dense. Then the redundant coverage area is:-

$S_r{}^2 = S_Z - S$

$= \text{m}\pi R^2$- S (3)

When n=1 the probability scheme broadcast with Pi

Si = PiπR^2 {i ≤ [1,m]}

The total area can be calculated as:

$S_Z{}^2 = \sum_{i=1}^{m} Si \rightarrow \sum_{i=1}^{n} Pi\pi R^2$

$S_Z{}^2 = \sum_{i=1}^{m} \left(\frac{Lij}{R}\right) \pi R^2 \rightarrow \sum_{i=1}^{m} Lij\pi R$

(4)

Total redundant area can be calculated as:-

$S_r{}^2 = S_Z{}^2$ – S

$= \sum_{i=1}^{m} Lij\pi R$ - S (5)

Since Lij < R, the distance between two connecting nodes will not exceed the transmission radius. So we can conclude that from equation 3 and 5

m$\pi R^2 > \sum_{i=1}^{m} Lij\pi R$

That is $S_r{}^1 > S_r{}^2$ this means the total redundant area of fixed probability is smaller than simple flooding. Similarly when n>1, total area can be calculated

$S_Z{}^3 = \sum_{i=1}^{m} \left(\frac{Lij}{R}\right)^n \pi R^2$

$S_Z{}^3 = \sum_{i=1}^{m} \pi \frac{Lij^n}{R^{n-2}}$ (6)

As derived that from equation 4

$S_Z{}^2 = \sum_{i=1}^{m} \pi LijR$

Let $S_Z{}^2 = \sum_{i=1}^{m} \pi \left(\frac{Lij}{R}\right)^{n-1} \left(\frac{Lij^n}{R^{n-2}}\right)$

Then replace the equation with $S_Z{}^3$ from equation 6 we get

$S_Z{}^2 = \sum_{i=1}^{m} \pi \left(\frac{R}{Lij}\right)^{n-1} S_Z{}^3$

As stated Lij is smaller than R and n>1 so:-

$\left(\frac{R}{Lij}\right)^{n-1} > 1$

$S_Z{}^2 > S_Z{}^3$ and $S_r{}^2 > S_r{}^3$ can be deduced

Hence the redundant area is reduced by broadcasting with a probability value based on the node location. By considering the above analysis we conclude that border nodes should have higher probability value.

III. Fuzzy Logic-Based Distance Broadcasting Scheme

A probabilistic route discovery approach can be developed which can further reduce the route discovery overhead by exploiting the problem solving control system methodology that is fuzzy logic. The Fuzzy Logic algorithm is illuminated by the powerful capability of fuzzy logic system to handle uncertainty and ambiguity. Fuzzy logic system is well known as model free. Their membership functions are not based on statistical distributions. In this paper, we apply fuzzy logic system to optimize the broadcasting scheme in AODV based on the node location. The main goal is designing the algorithm to use Fuzzy Logic Systems so as to avoid the broadcast storm problem.

For the inference process we use Mamdani's method. Mamdani's method is most commonly used in applications. There are four steps to get the crisp value from the FIS system.

1) *The first step is to evaluate the antecedent for each rule.*
2) *The second step is to obtain each rule's conclusion.*
3) *The third step is to aggregate conclusions.*
4) *The fourth and last is defuzzification*

The defuzzified output is then given as input into probability broadcasting of the AODV RREQ module. In this work fuzzy logic is embedded in Qualnet so as to have performance analysis of a dynamic nature.

B. Fuzzy Logic Controller

Fuzzy logic control is derived from fuzzy set theory introduced by Zadeh in 1965. The Fuzzy Logic Controller (FLC) shows a better performance than conventional controllers in the form of increased robustness. Fuzzy Control is based upon practical application knowledge represented by so called linguistic rule based, rather than by analytical (either empirical or theoretical) models. Fuzzy Control can be used when there is an expertise that can be expressed in its formalism**.** Other advantages of FLC are:

1) *It can work with less precise inputs.*
2) *It does not need fast processors.*
3) *It needs less data storage in the form of membership functions and rules than conventional look up table for nonlinear controllers.*

C. Fuzzification

Fuzzy logic uses linguistic variables instead of numerical variables. The process of converting a numerical variable (real number or crisp variable) into a linguistic variable (fuzzy number) is called fuzzification.

The simplest form of membership function is triangular membership function and it is used here as the reference. To determine the broadcasting probability value, one input function transforms the system inputs into fuzzy sets which is node location.

Table I below shows the membership value for the input node location and figure 3 shows the input Membership Function.

TABLE 1: INPUT LINGUISTIC VARIABLE

Input	*Membership*
Node location	Border, Internal border, Exterior, Interior

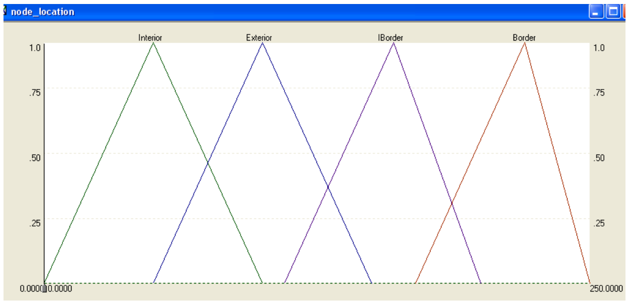


Figure 3. Input membership function

The output function is composed of four membership functions as seen in figure 4. Table II gives the output functions.

TABLE II: OUTPUT LINGUISTIC VARIABLE

Output	*Membership*
Probability	High, Medium, Low, very Low

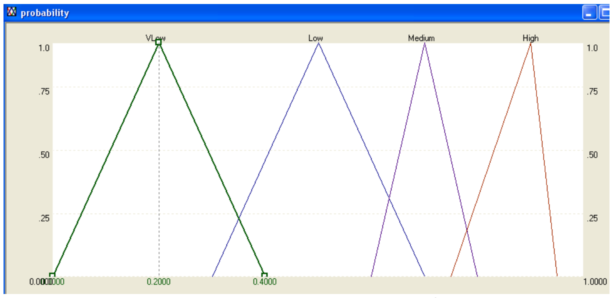


Figure 4. Output membership function

D. *Rule base table and inference engine*

The rules are in the format - „If else. Then, the if „part" of a rule is called the rule-antecedent and is a description of a process state in terms of a logical combination of atomic fuzzy propositions. The „then" part of the rule is called the rule consequent and is a description of the control output in terms of logical combinations of fuzzy propositions. In our system, we have 4 rules in the fuzzy inference. The form of the rules is: IF A, THEN C. The A and C represent node location and probability respectively.

E. *Defuzzification*

The reverse of fuzzification is called defuzzification. The use of FLC inference engine produces required output in a linguistic form. After aggregating the conclusions obtained by each rule, a defuzzification method is still needed to get the crisp value. One of the most popular defuzzification methods is the Centroid, which returns the center of the area under the fuzzy set obtained aggregating conclusions. The Centroid is shown in figure 5.

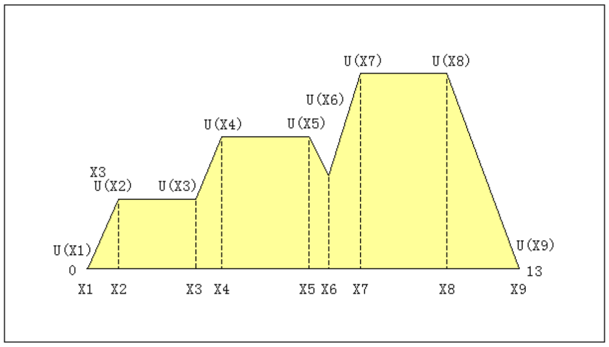


Figure 5. Centroid method

IV. SIMULATION

Qualnet is a discrete event simulator used in the simulation of mobile ad-hoc networks. To evaluate the performance of the proposed scheme for route discovery algorithms, the implementation of the AODV routing protocol in the qualnet simulator has been modified and fuzzy logic control has been embedded so as to efficiently execute the proposed algorithm.

The simulation parameters that have been used in our experiments are stated in table III.

Table III. Simulation Parameter

Parameter	*Value*
Transmitter range	250
Bandwidth	2 Mbps
Interface queue length	50 messages
Simulation time	900 seconds
Pause time	0 second
Packet size	512 bytes
Topology size	(1000*1000) m^2
Number of Nodes	25, 50, 75 and 100 nodes
Data traffic	CBR
Mobility model	Random waypoint

Extensive simulation experiments have been conducted to compare the performance of AODV, distance aware counter based broadcast [DACBB] and the fuzzy logic based Probabilistic-AODV [FPBB] and [15]

V. RESULT AND ANALYSIS

The analysis and comparison is done by considering two different settings, each designed to assess the impact of a particular network condition on the performance of the protocol.

Firstly, the impact of network density or size is assessed by deploying 25, 50, 75 and 100 mobile nodes over a fixed network size of 1000 by 1000 square meters. The second setting investigates the effects of offered traffic load on the performance of the routing protocols by varying the number of source destination pairs over the range of 5, 10, 15, 20 and 25 flows for each simulation scenario.

Figure 6, 7 and 8 shows the performance of AODV, DACBB and FPBB in terms of network throughput, saved rebroadcast and reachability versus network density. As shown, in the figures, the network throughput, saved rebroadcast and reachability generated by FPBB increases almost linearly as the network density increases.

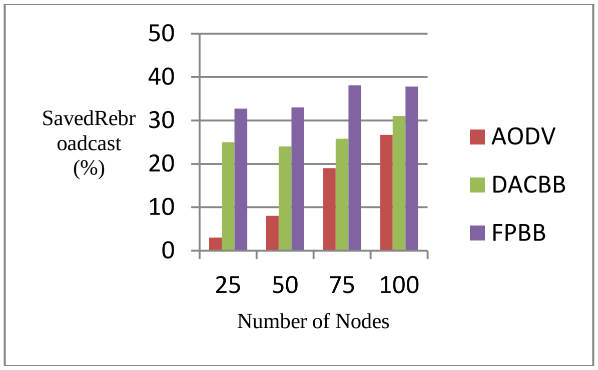


Figure 6. Saved Rebroadcast

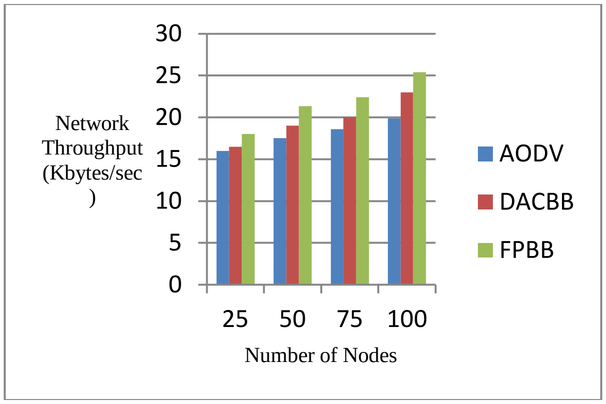


Figure 7. Network Throughput

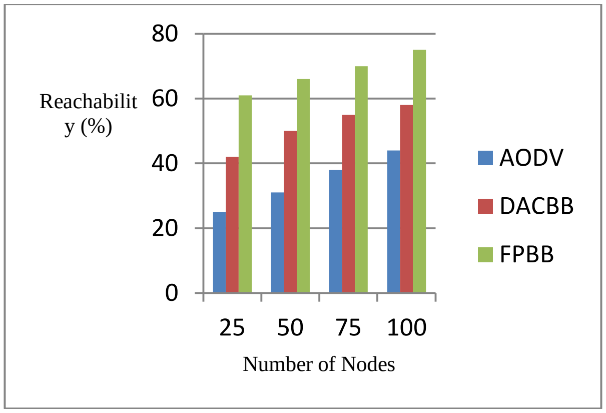


Figure 8. Reachability

Network throughput for AODV, distance aware counter based broadcast [DACBB] and the fuzzy logic based Probabilistic-AODV [FPBB] are similar and increases almost linear when the offered load is increased from 5 to 15 flows. This is because when the number of flows is increased, the number of nodes initiating route discovery operation is also increased. As a consequence, more RREQ packets are generated and transmitted which leads to a high consumption of the communication bandwidth. This phenomenon leads to a fewer number of data packets delivered at the destinations, there by degrading the network throughput.

Figure 9 shows the superiority of the FPBB over the traditional AODV and DACBB becomes more noticeable in the case of high offered load (e.g. 25 flows). At offered load of 25 flows, the network throughput is increased by around 9.24 and 5.09 percent respectively, when compared against the AODV and DACBB. The difference in the achieved network throughput is due to the reduction of the number of nodes involved in the dissemination of RREQ packets in congested networks, leading to a reduction of routing overhead and packet collisions. As a consequence more communication bandwidth is freed for data transmission.

The FBPP has the least number of rebroadcast for almost all traffic loads. We vary the traffic load by using different number of CBR source–destination connections. Figure 10 shows that as the number of connection increases the rebroadcast is saved more. Whereas when the number of CBR connection increases figure 11 shows that reachability also increases linearly with the increase in traffic load.

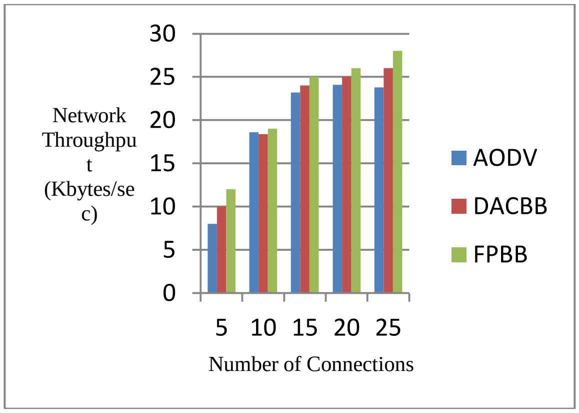


Figure 9. Network Throughput

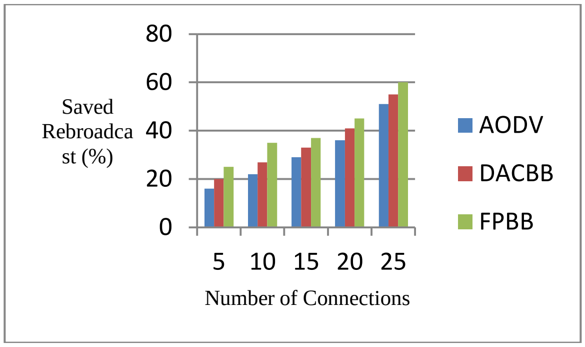


Figure 10. Saved Rebroadcast

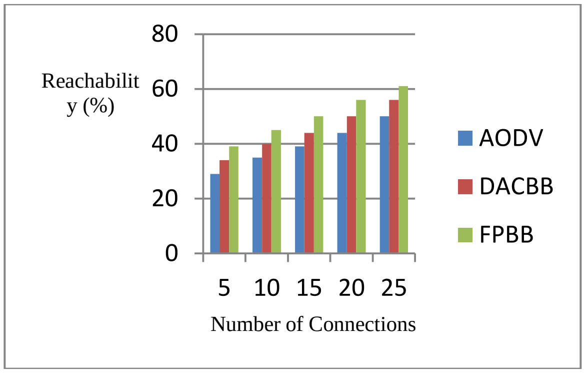


Figure 11. Reachability

Simulation results reveal that FBPP technique outperforms DACBB in most considered performance metrics such as saved rebroadcast and reachability, while maintaining comparable performance in other important performance characteristics of the network such as throughput, reachability and saved rebroadcast.

VI. Conclusion And Future Scope

This paper has evaluated the performance of fuzzy logic based distance broadcasting scheme with distance aware counter based broadcast. The present work brings out the potential advantages of applying Fuzzy Logic Control technique for generating dynamic probability value based on the node location. Fuzzy Logic Control can therefore be an effective strategy for generating varying probability value for broadcasting in MANET.

The simulation results revealed that the proposed algorithm generates much higher throughput and saved rebroadcast. The results have also shown that the degradation of the number of RREQ packet initiated and forwarded in dense network has significantly reduced. Though the analysis in this paper has been very crude, but this clearly depicts the advantage of adding the fuzzy logic controller in the conventional probabilistic broadcasting scheme. The results tend to be more broadcasting efficient. The comparisons show the superiority of Fuzzy Logic Control scheme over the smart probabilistic broadcasting schemes. As a continuation of this research work in near future, we plan to further explore the performance of the fuzzy logic based Probabilistic-AODV [FPBB] using on-demand routing protocols such as DSR. The fuzzy logic based Probabilistic-AODV [FPBB] can also be used to examine the effect on the routing table advertisements in proactive routing protocols, such as OLSR, and ZRP.

We can further explore the fuzzy logic based Probabilistic-AODV [FPBB] with varying mobility model. In the present work random waypoint mobility model is used , other model such as community based mobility model, gauss mobility model , manhattan mobility model and others can be used to study the effect on the performance of the algorithm.

References

[1] B. Williams and T. Camp, “Comparison of broadcasting techniques for mobile ad hoc networks,” in Proceedings of the 3rd ACM international symposium on Mobile ad hoc networking. ACM Press, 2002, pp. 194–205.

[2] W. Lou and J. Wu, Localized Broadcasting in Mobile Ad Hoc Networks Using Neighbor Designation. CRC Press, 2003.

[3] F. Dai and J. Wu, “Distributed dominant pruning in ad hoc networks,” in ICC, 2003.

[4] Ho, C., K. Obraczka, G. Tsudik and K. Viswanath, 1999. Flooding for reliable multicast in multi-hop ad hoc networks. Proceedings of the International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communication, Aug. 20- 20, Seattle, Washington, United States, pp: 64-71. http://portal.acm.org/citation.cfm?id=313291

[5] Jetcheva, J., Y. Hu, D. Maltz and D. Johnson, 2001. A simple protocol for multicast and broadcast in mobile ad hoc networks. Internet Draft: draft-ietf-manet-simple-mbcast-01.txt.

[6] Ni, S., Y. Tseng, Y. Chen and J. Sheu, 1999. The broadcast storm problem in a mobile ad hoc network. Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking, Aug. 15-19, Seattle, Washington, United States, pp: 151-162. http://portal.acm.org/citation.cfm?id=313451.313525.

[7] Sucec, J. and I. Marsic, 2000. An efficient distributed network-wide broadcast algorithm for mobile ad hoc networks. CAIP Technical Report 248-Rutgers University. http://citeseer.ist.psu.edu/312658.html

[8] Y.-C. Tseng, S.-Y. Ni, and E.-Y. Shih, "Adaptive approaches to relieving broadcast storms in a wireless multihop mobile ad hoc networks," Proceedings of IEEE Transactions on Computers, vol. 52, pp. 545-557, May 2003.

[9] Q. Zhang and D. P. Agrawal, "Dynamic probabilistic broadcasting in MANETs," Journal of Parallel and Distributed Computing, vol. 65, pp. 220-233, 2005.

[10] M. B. Yassein, M. O. Khaoua, L. M. Mackenzie, and S. Papanastasiou, "The Highly Adjusted Probabilistic Broadcasting in Mobile Ad hoc Networks," Proceedings of the 6th Annual PostGraduate Symposium on the Convergence of Telecommunications, Networking & Broadcasting, (PGNE T 2005), vol. ISBN 1- 902-56011-6, pp. 27-28, June 2005.

[11] Y. Sasson, D. Cavin, and A. Schiper, "Probabilistic broadcast for flooding in wireless mobile ad hoc networks," Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), March 2003.

[12] W. Peng and X. C. Lu, "On the reduction of broadcast redundancy in mobile ad hoc networks," proceedings of the ACM Symposium on Mobile and Ad Hoc Networking and Computing (MobiHoc'00), pp. 129-130, August, 2000.

[13] J. Wu and W. Lou, "Forward-node-set-based broadcast in clustered mobile ad hoc networks," Wireless Communications and Mobile Computing, vol. 3, pp. 155- 173, March 2003.

[14] C. Perkins, E. Belding-Royer, and S. Das, "Ad hoc On-Demand Distance Vector (AODV) Routing," IETF Mobile Ad Hoc Networking Working Group INTERNET DRAFT, RFC 3561,July 2003, http://www.ietf.org/rfc/rfc3561.txt. Experimental RFC, retrieved in October 2007.

[15] Chien Chen, Chin-Kai Hsu, and Hsien-Kang Wang,” A distance-aware counter-based broadcast scheme for wireless ad hoc networks”,Military communication conference-2005 IEEE, 17-20 oct, Pages 1052-1058, Vol-2.

[16] L.Zhou, G.Cui, H.Liu, Wu and D.Luo “ NPPB: A Broadcast scheme in Dense VANETs”, Information Technology Journal 9 (2): 247-256, 2010, ISSN 1812-5638, 2010 Asian Network for Scientific Information.

[17] Tasneem Bano, Jyoti Singhai “Probabilistic broadcasting Protocol in ad hoc network and its advancement: a Review”, IJCSES volume 1, number 2, Nov 2010, AIRCC publicationISSN-0976-2760 Online ISSN-0976-3252 Print.

Authors Profile

Tasneem Bano Rehman is a research scholar in Computer Science engineering department in Maulana Azad National Institute of Technology (MANIT). She holds a Master of Technology and Maulana Azad National Institute Of Technology. Her general research interests include wireless communication especially Mobile Ad-hoc Networks.

Dr. Jyoti Singhai is Associate Professor in Electronics and Communication Engineering Department in Maulana Azad National Institute of Technology (MANIT) since 1994. She did her BE in Electronics Engineering in 1991 from MANIT (formerly known as Maulana Azad College of Technology), Bhopal. She did her M.Tech. in Digital Communication in 1997 and PhD in 2005 from MANIT, Bhopal. She is recipient of “Young Scientist Award” from M.P Council of Science and Technology, Government of M.P. for the year 2002-03, “Career Award for Young Teachers” form AICTE in 2008 and “BOYSCAST Fellowship” from DST in 2010. She has published over 80 papers in various National and International conferences. She has supervised 40 UG Major projects, 41 M.Tech. projects and 6 PhD.

SAS: Implementation of scaled association rules on spatial multidimensional quantitative dataset

M. N. Doja
Professor
Faculty of Engg & Technology
Jamia Millia Islamia
New Delhi -110025,India

Sapna Jain
PhD fellow
Department of Computer Science
Jamia Hamdard
New Delhi-110062,India

M Afshar Alam
Professor
Department of Computer Science
Jamia Hamdard
New Delhi-110062,India

Abstract— **Mining spatial association rules is one of the most important branches in the field of Spatial Data Mining (SDM). Because of the complexity of spatial data, a traditional method in extracting spatial association rules is to transform spatial database into general transaction database. The Apriori algorithm is one of the most commonly used methods in mining association rules at present. But a shortcoming of the algorithm is that its performance on the large database is inefficient. The present paper proposed a new algorithm by extracting maximum frequent itemsets based on spatial multidimensional quantitative dataset. Algorithms for mining spatial association rules are similar to association rule mining except consideration of special data, the predicates generation and rule generation processes are based on Apriori. The proposed method (SAS) Scaled Aprori on Spatial multidimensional quantitative dataset in the paper reduces the number of itemsets generated and also improves the execution time of the algorithm.**

Keywords- ***association rules; spatial dataset; X tree.***

I. Introduction

Data mining and knowledge discovery have become popular fields of research. A significant subset of this research is looking at the particular semantics of space and time and the manner in which they can be sensibly accommodated into data mining algorithms [12].

Scaling is considered an important aspect in Data Mining. The problem of scalability in Data mining is not only how to process such large sets of data, but how to do it within a useful timeframe. Scalability means that as a system gets larger, its performance improves correspondingly. The main purpose of data mining techniques is to find hidden information and unknown relations within an amount of data.

Spatio-Temporal applications like climate change modeling and analysis [1], transportation systems, forest monitoring[2], diseases spreading[3], temporal geographic information systems[4] and environmental systems[5] process spatial, temporal and attribute data elements for knowledge discovery. The spatial objects are characterized by their position, shape and spatial attributes. Spatial attributes are properties of space and spatial objects located in specific positions inherit these attributes. Spatial attributes refer to the whole space and can be represented as layers, each layer represent one theme. The temporal objects are characterized by two models of time that are used to record facts and information about spatial objects.

Association rules mining are one of the major techniques of data mining and it is perhaps the most common form of local-pattern discovery in unsupervised learning systems. The technique is likely to be very practical in applications which use the similarity in customer buying behavior in order to make peer recommendations.

Association rule-based approaches focus on the creation of transactions over space so that an apriori like algorithm [28] can be used. Transactions over space can use a reference-feature centric [29] approach or a data-partition [30] approach. The reference feature centric model is based on the choice of a reference spatial feature and is relevant to application domains focusing on a specific Boolean spatial feature, e.g., incidence of cancer. Domain scientists are interested indicating theco-locations of other task relevant features (e.g., asbestos) to the reference future. Transactions are created around instances of one user specified reference spatial feature. The association rules are derived using the apriori [28] algorithm. The rules found are all related to the reference feature.

In this 21st century the developments in spatial data acquisition, mass storage and network interconnection, volume of spatial data has been increasing dramatically. Vast data satisfied potential demands of exploring the earth’s resource and environment by human being, widening exploitable information source, but the processing approaches of spatial data lag behind severely, and are unable to discover relation and rules in large amount of data efficiently and make full use of existing data to predict development trend.

This work is the extension of Aprori-UB which uses multidimensional access method, UB-tree to generate Better association rules with high support and confidence. In multidimensional databases, objects are indexed according to several or many independent attributes. However, this task cannot be effectively realized using many standalone indices and thus special indexing structures have been developed is last two decades. Indexing and querying high dimensional databases.

This paper has the following sections. Section 2 represents the previous work done in the same field .Section 3 gives the conceptual details used in the proposed algorithm. Section 4 highlights the proposed Aprori-UB method .Section 5 gives the implementation details. Section 6 discusses the conclusion and future scope.

II. Related Work

Spatial datasets need to be preprocessed to construct the transaction database before mining spatial association rules according to the main idea of mining spatial association rules at present. Imam Mukhlash and Benhard Sitohang put forward the framework of spatial data preprocessing, including feature (spatial and non-spatial) selection based on spatial parameters, performing dimension reduction and selection of non-spatial attributes, performing data categorization based on non-spatial data parameters, performing join operations for spatial objects based on spatial parameters and transforming into output form [16].

The preparation and preprocessing of spatial datasets: The spatial datasets in the case included the elevation, the slope and the aspect with the spatial resolution of 100m and the land cover map. A spatial database is defined as a collection of inter-related geodspatial data that can handle and maintain a large amount of data which is shareable between different GIS applications.

The consistency with little or no redundancy and maintenance of data quality including updating. The self-descriptive analysis with metadata and high performance by database management system with database including security access control mechanism.

Spatial Data Mining (SDM) is a process of spatial support decision, which aims at extracting the implicit, unknown, potential, useful spatial and non-spatial knowledge from spatial data, including general geometry rules, spatial characteristics rules, spatial classification rules, spatial clustering rules, spatial association rules and so on [1]. Spatial association rule, termed as spatial association location pattern [2], is one of the most important branches in the SDM, which means a rule indicating certain association relationships among a set of spatial and nonspatial attributes of geographical objects. Because of the complexity of spatial data, the main idea of extracting spatial association rules is to mine spatial association rules in the transaction database categorized from spatial data using some mining algorithms.

A spatial database is a collection of spatially referenced data that acts as a model of reality. This database model represents a selected set or approximation of phenomena which are deemed important enough to represent the digital representation might be for some past, present or future time period .

The Apriori algorithm [3] is one of the most commonly used algorithms in mining association rules at present, and its typical application was market basket analysis to discover customer shopping patterns [4]. Apriori Algorithm can be used to generate all frequent itemset. A Frequent itemset is an itemset whose support is greater than some user-specified minimum support (denoted L, where k is the size of the itemset). A Candidate itemset is a potentially frequent itemset (denoted C , where k is the size of the itemset).

A. Generate the candidate itemsets in N1. Save the frequent itemsets in N2.

B. Steps

1) Generate the candidate itemsets in C from the frequent itemsets in L,k-1

2) Join L k-1 p with L q, as follows: insert nto C select p.item from L k-1 k p, L q ,where, p.itemk-1, p.itemk-1 , . . . , p.item = q.item

3) Generate all (k-1)-subsets from the candidate itemsets in Ck

4) Prune all candidate itemsets from C• where, some (k-1)-subset of the candidate itemset is not in the frequent itemset L k

5) Scan the transaction database to determine the support for each candidate itemset in Nk.

6) Save the frequent itemsets in L k.

III. Concept Used

X-tree has data nodes and normal directory nodes. A data node contains minimum bounding rectangles (MBRs) together with pointers to the actual data objects: (MBR, p) while the directory node consists of MBRs together with pointers to sub-MBRs. In addition, the X-tree introduces another type of nodes: super nodes. A super-node is large directory node of variable size (a multiple of the usual block size). Figure 7 shows an example of an X-tree structure with three kinds of nodes: directory node, leaf node, and super node.

Since the original X-tree was proposed [15], there have been several implementations of X-trees. One of them is the X+-tree [19]. The X+-tree allows the increase of the size of super-nodes in the X-tree to some degree. Technically, in order to avoid overlap, which is bad for performance, a super node might grow during the insertion. However, the linear scan of a large super node can be a problem. In the X-tree, the size of a super-node can be many times larger than size of a normal node. In the X+-tree, the size of super-node is at most the size of a normal node multiplied by a given user parameter MAX_X_SNODE. When the super-node becomes larger than the upper limit, the super-node has to be split into two new nodes.

The X-tree (eXtended node tree) is a new index structure supporting efficient query processing of high-dimensional data. The goal is to support not only point data but also extended spatial data and therefore, the X-tree uses the concept of overlapping regions. The X-tree therefore avoids overlap whenever it is possible without allowing the tree to degenerate; otherwise, the X-tree uses extended variable size directory nodes, so-called supernodes. In addition to providing a directory organization which is suitable for high-dimensional data, the X-tree uses the available main memory more efficiently

The X-tree may be seen as a hybrid of a linear array-like and a hierarchical R-tree-like directory. It is well established that in low dimensions the most efficient organization of the directory is a hierarchical organization.

The reason is that the selectivity in the directory is very high which means that, e.g. for point queries, the number of required page accesses directly corresponds to the height of the tree. This, however, is only true if there is no overlap between directory rectangles which is the case for a low dimensionality. It is also reasonable, that for very high dimensionality a linear organization of the directory is more efficient.

The reason is that due to the high overlap, most of the directory if not the whole directory has to be searched anyway. If the whole directory has to be searched, a linearly organized directory needs less space and may be read much faster from disk than a block-wise reading of the directory. For medium dimensionality, an efficient organization of the directory would probably be partially hierarchical and partially linear.

The problem is to dynamically organize the tree such that portions of the data which would produce high overlap are organized linearly and those which can be organized hierarchically without too much overlap are dynamically organized in a hierarchical form. The algorithms used in the X-tree are designed to automatically organize the directory as hierarchical as possible, resulting in a very efficient hybrid organization of the directory.

A. *Structure of the X-tree*

The overall structure of the X-tree is presented in Figure 1. The data nodes of the X-tree contain rectilinear minimum bounding rectangles (MBRs) together with pointers to the actual data objects, and the directory nodes contain MBRs together with pointers to sub-MBRs (cf. Figure 1).

The X-tree consists of three nodes: data nodes,normal directory nodes, and supernodes. Supernodes are large directory nodes of variable size (a multiple of the usualblock size).

The basic goal of supernodes is to avoid splits in the directory that would result in an inefficient directory structure. The alternative to using larger node sizes is highly overlapping directory nodes which would require accessing most of the son nodes during the search process. This, however, is more inefficient than linearly scanning the larger supernode.

The X-tree is completely different from an R-tree with a larger block size since the X-tree only consists of larger nodes where actually necessary. As a result, the structure of the X-tree may be rather heterogeneous as indicated in Figure 1[7]. Due to the fact that the overlap is increasing with the dimension, the internal structure of the X-tree is also changing with increasing dimension.

In Figure 1, three examples of X-trees containing data of different dimensionality are shown. As expected, the number and size of supernodes increases with the dimension. For generating the examples, the block size has been artificially reduced to obtain a drawable fan-out.

Due to the increasing number and size of supernodes, the height of the X-tree which corresponds to the number of page accesses necessary for point queries is decreasing with increasing dimension [7].

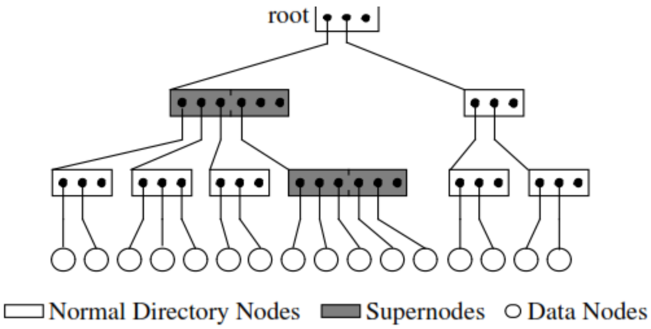


Figure1: Structure of a X-tree.

IV. PROPOSED WORK

The SAS algorithm puesodocode of the algorithm is:

a) Identify the correlated data in the spatial dataset.

b) Find all frequent item sets.

c) Generate scaled association rules from the frequent item sets

d) Identify the quantitative elements.

e) Sorting the item sets based on the frequency and quantitative elements.

f) Use Xtree to create a multidimensional spatial dataset.

g) Discard the infrequent item value pairs

h) Iterate the steps c to f till the required mining results are achieved.

Let I = {i1, i2 … i n items} be a set of items, and T a set of transactions, each a subset of I. An association rule is an

Implication of the form A=>B, where A and B are non-intersecting. The support of A=>B is the percentage of

The transactions that contain both A and B:

X tree psedocode :

Input: A set of M current model tree nodes M A set of current support tree nodes X.

Output: A list Z of feasible sets of points

```
1) X ← {} and X_curr ← X
2) IF we cannot prune based on the mutual compatibility of M:
3) FOR each s ∈ X_curr
4) IF s is compatible with M:
5) IF s is “too wide”:
6) Add s’s left and right child to the end of X_curr
7) ELSE
8) Add s to X.
9) IF we have enough valid support points:
10) IF all of v ∈ M are leaves:
11) Test all combinations of points owned by the model nodes, using the support nodes’ points as potential support.
Add valid sets to Z.
12) ELSE
13) Let m∗ be the non-leaf model tree node that owns the most points.
```

14) Search using m∗'s left child in place of m∗and S'instead of S.

15) Search using m∗'s right child in place of m∗and Sinstead of S.

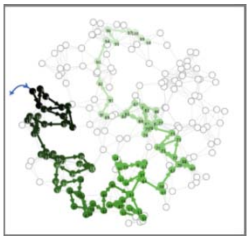


Figure 2 : spatial attributes

The main idea of the X-tree is to avoid overlap of bounding boxes in the directory by using a new organization of the directory which is optimized for high dimensional space. The X-tree avoids splits which would result in a high degree of overlap in the directory. Instead of allowing splits that introduce high overlaps, directory nodes are extended over the usual block size, resulting in so-called supernodes. The supernodes may become large and the linear scan of the large supernodes might seem to be a problem.

Additionally, oversize shelves are organized as chains of disk pages which cannot be read sequentially.We implemented the X-tree index structure and performed a detailed performance evaluation using very large [6]. The X-tree also provides much faster insertion times (about 4 times faster than ub-tree).

V. IMPLEMENTATION AND FUTURE WORK:

In order to show the performance of the proposed algorithm, we applied the algorithm to Diabetes Data Set which was obtained from UCI Machine Learning Repository[16].This dataset is multivariate, TimeSeries and has 20 attributes. After discovering rules, they have to be presented in understandable form to the user. Java programs since Java byte-codes are compiled or interpreted by the Java Virtual Machine resulting in performance penalty. The core of the X+-tree implementation in [15] is reused, with changes and additions made to data structure and functions.

Spatial Trend Detection: Spatial trends describe a regular change of non-spatial attributes when moving away from a start object o. The existence of a global trend for a start object o indicates that if considering all objects on all paths starting from the values for the specified attribute(s) in general tend to increase (decrease) with increasing distance. Our algorithm detects regions showing a certain global trend, and algorithm local-trends then finds within these regions some paths having the inverse trend (see figure 3).

The algorithm mine the frequent itemsets by using a divideand-conquer strategy as follows: SAS first compresses the database representing frequent itemset into a frequent-pattern tree, or X-tree, which retains the itemset association information as well. The next step is to divide a compressed database into set of spatial databases (a special kind of projected database), each associated with one frequent item. Finally, mine each such database separately, particularly, the construction of X-tree and the mining of X-tree (figure 6).

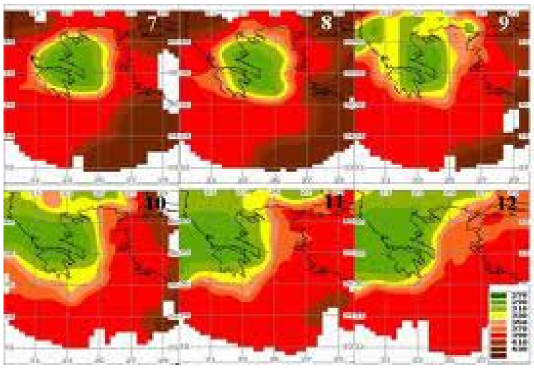
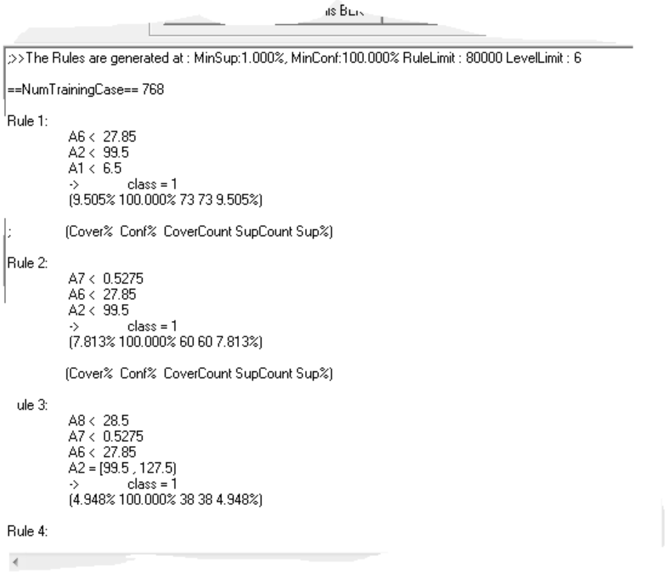
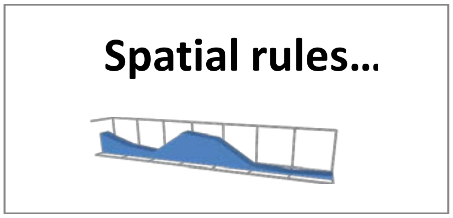


Figure 3 : trend detection phases.



```
;>>The Rules are generated at : MinSup:1.000%, MinConf:100.000% RuleLimit : 80000 LevelLimit : 6

==NumTrainingCase== 768

Rule 1:
            A6 <  27.85
            A2 <  99.5
            A1 <  6.5
            ->          class = 1
            (9.505% 100.000% 73 73 9.505%)

;           (Cover%  Conf%  CoverCount SupCount Sup%)

Rule 2:
            A7 <  0.5275
            A6 <  27.85
            A2 <  99.5
            ->          class = 1
            (7.813% 100.000% 60 60 7.813%)

            (Cover%  Conf%  CoverCount SupCount Sup%)

  ule 3:
            A8 <  28.5
            A7 <  0.5275
            A6 <  27.85
            A2 = [99.5 , 127.5)
            ->          class = 1
            (4.948% 100.000% 38 38 4.948%)

Rule 4:
```

Figure 4: Support and confidence of dataset with rules .



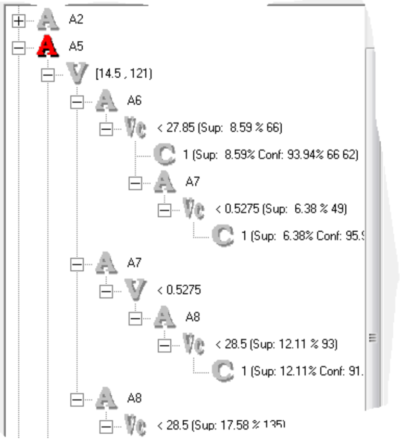
Spatial rules…

Figure 5:Rules generation

The experiment focused on evaluating aprori, Aprori-ub and SAS algorithms. Since we were interested in seeing the best performance, we used diabetes data set samples. The minsupp, minconfidence level and average rule error was compared in figure 8. The evaluation shows that our proposed SAS generated strong association rule with less rule generation error on spatial multidimensional dataset.

VI. CONCLUSION

In this paper we proposed a practical to find frequent patterns using X Tree. X-tree compresses both dense and sparse datasets by using numerical value representation. In this method we consider Fibonacci number characteristics to find CFP (closed frequent pattern) and then the MFP (maximal frequent pattern) our approach is efficient on both dense and sparse database.



```
⊞ A A2
⊟ A A5
   ⊟ V [14.5 , 121)
      ⊟ A A6
         ⊟ Vc < 27.85 (Sup:  8.59 % 66)
            C 1 (Sup:  8.59% Conf: 93.94% 66 62)
            ⊟ A A7
               ⊟ Vc < 0.5275 (Sup:  6.38 % 49)
                  C 1 (Sup:  6.38% Conf: 95.
      ⊟ A A7
         ⊟ V < 0.5275
            ⊟ A A8
               ⊟ Vc < 28.5 (Sup: 12.11 % 93)
                  C 1 (Sup: 12.11% Conf: 91.
      ⊟ A A8
         ⊟ Vc < 28.5 (Sup: 17.58 % 135)
```

Figure 6: X tree function calculation.

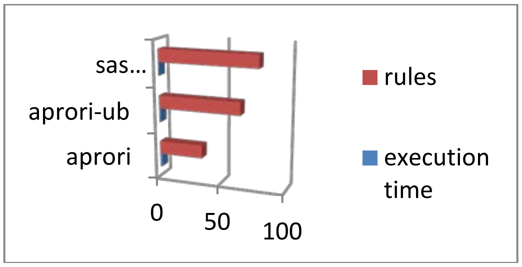


Figure 7 : Comparison of SAS with other algorithms

The algorithm will positively enhance the efficiency of judgment the relationship between spatial objects and further can be used in association analysis. The creation of maximal frequent patterns is done by intersecting the ordered list (OL) of similar type which reduces the search space.

Spatially, association is a relationship between spatial objects. Association analysis is one of the most widely research topics in data mining. The main focus of association rule mining is to generate hypothesis rather than to test them as is commonly achieved using statistical techniques (15). The concept of association rule, introduced by Agrawal, was used for analyzing market basket data to mine customer shopping patterns.

Spatial association algorithms find the frequent sets in spatial and non-spatial databases,and inter-relationship between different variables that are not explicitly stored in the spatial database. In many situations there is a need to discover spatial association rules, rules that associate one or more spatial objects. To confine the number of rules, the concept of minimum support and minimum confidence are used. The intuition behind this is that in large databases,there may exist a large number of associations between objects but most of them will be applicable to only a small number of objects, or the confidence of the rule may be low. However, a strong rule is a rule with large support, i.e., no less than the minimum support threshold, and a large confidence, i.e., no less than the minimum confidence threshold e.g. is_a (X, city) within (X,maharastra) adjacent_to (X,water) close_to(X, Karnataka)…(92%). The rule states that 92% of the cities within Maharastra and adjacent to water are close to Karnataka, which associates predicates is a, within, and adjacent_to with spatial predicate close_to. The quality of the rule is measured in the terms of the surprise associated with it. To calculate the surprise or interestingness associated with the mined rule the correlation and chi-square test technique is adopted.

We have generated scaled association rules with high support and confidence. Other future work in this field includes discovery algorithms with dynamic changes of µ level, improved performance strategies and new measures for rule management.

References

[1] Auroop R Ganguly and Karsten Steinhaeuser, (2008) ”Data Mining for climate change and impacts”,IEEE international conference on data mining workshops,ICDMW,15-19,Dec,2008,Italy.

[2] T. Cheng and J. Wang, (2006) “Applications of spatio-temporal data mining and knowledge discovery (STDMKD) for forest fire prevention”, ISPRS Commission VII Mid-term Symposium "Remote Sensing: From Pixels to Processes, Enschede, the Netherlands, 8-11 May 2006 .

[3] Diego Ruiz-Moreno, Mercedes Pascual, Michael Emch, Mohammad Yunus, (2010) “Spatial clustering in the spatio-temporal dynamics of endemic cholera“, BMC Infectious Diseases, Volume 10, 2010.

[4] Yang ping , Tang Xinming , Wang Shengxiao, (2008) “Dynamic cartographic representation of SpatioTemporal data“, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXVII. Part B2. Beijing 2008 .

[5] Z. Obradovic , D. Das, V.Radosavljevic, K.Ristovski, S.Vucetic, (2010) “Spatio-Temporal characterization of aerosols through active use of data from multiple sensors, “ISPRS TC VII Symposium – 100 Years ISPRS, Vienna, Austria, July 5–7, 2010

[6] Günther O., Noltemeier H.: ‘Spatial Database Indices For Large Extended Objects’,Proc. 7 th Int. Conf. on Data Engineering, 1991, pp. 520-527.

[7] Stefan Berchtold,Daniel A. Keim,Hans-Peter Kriegel., The X-tree:An Index Structure for High-Dimensional Data,Proceedings of the Twenty-second International Conference on Very Large Data-bases,Mumbai ,India .

[8] J.F. Roddick, K. Hornsby, and M. Spiliopoulou.YABTSSTDMR - yet another bibliography of temporal,spatial and spatio-temporal data mining research.In K.P. Unnikrishnan and R. Uthurusamy,eds, SIGKDD Temporal Data Mining Workshop,pages 167–175, San Francisco, CA, 2001. ACM.

[9] [Online] Available: http://www. http://www.cs.rpi.edu/~zaki/ software/

[10] [Online] Available: http://www. http://www.csc.liv.ac.uk/~frans/KDD/Software/FPgrowth/fpGrowth.html

[11] [Online] Available: http://www. http://www.csc.liv. ac.uk/~frans/KDD/Software/FPgrowth/FPtree.java

[12] [Online] Available: http://www. http://www.sigkdd.org/ kddcup/index.php?section=1998&method=data

[13] [Online] Available: http://www. http://www.kdnuggets.com/ software/associations.html

[14] [Online] Available: http://www. ttp://hen. wikipedia.org /wiki/Weka_machine_learning

[15] [Online] Available: http://www. http://en. wikipedia.org/ wiki/Weka_machine_learning#ARFF_file

[16] [Online] Available: http://www. http://www.cs.waikato. ac.nz/ml/weka/

[17] [Online] Available: http://www. http://sourceforge.net/ projects/weka/files/weka-3-7-windows-jre/3.7.4/weka-3-74jre.exe/download

[18] [Online] Available: http://www. www.sigkdd.org/kddcup/

[19] [Online] Available: http://www. http://kdd.ics.uci.edu/

[20] [Online] Available: http://www. http://www.kdnuggets.com/ datasets/

[21] Sujni Paul, (2010) “An Optimized Distributed Association rule mining algorithm in Parallel and distributed data mining with XML data for improved response time”, International Journal of Computer

Science and Information Technology, Volume 2, Number 2, April 2010.

[22] Yangming JIANG and Siwen BI, (2008) "Dynamic Object-Oriented Model and its Applications for Digital Earth", Digital Earth Summit on Geoinformatics, Nov,12-14,2008,Germany.

[23] ZHANG Ruiju et al, (2005) "An Object Oriented Spatio-Temporal Data Model", Proceedings of International Symposium on Spatio-Temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion, 27-29, Aug,2005, peking University, China.

[24] S. Nadi and M.R.Delavar, (2005) "Toward a General Spatio-Temporal Database Structure for GIS applications", Proceedings of International Symposium on Spatio-Temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion, 27-29, Aug,2005, peking University, China.

[25] Souheil Khaddaj,Abdul Adamu and Munir Morad, (2005) "Construction of an Integrated Object Oriented System for Temporal GIS", American Journal of Applied Sciences 2(12), 2005, pp.1584-1594,ISSN 1546-9239.

[26] Salvatore Rinzivillo and Franco Turini, (2005) "Extracting spatial association rules from spatial transactions", Proceedings of GIS'05 13 annual ACM international workshop on geographic information systems,Germany.

[27] Schluter T and Conrad S, (2010) "Mining Several kinds of Temporal association rules enhanced by Tree structures", Second international conference on Information,Process and Knowledge Management,2010, Saint Maarten.

[28] R. Agrawal and R. Srikant. "Fast algorithms for mining association rules in large databases. ",Research Report RJ 9839, IBM Almaden Research Center, San Jose, California, June 1994.

[29] Koperski, K., and J. Han. 1995. Discovery of spatial association rules in geographic information databases. In Proceedings of 4th International Symposium on Large Spatial Databases, SSD95, Maine: 47-66.

[30] Iwaki, Hideki, Masaaki Kijima & Yuji Morimoto (2001). "An economic premium principle in a multiperiod economy," Insurance: Mathematics and Economics 28,325-339.

AUTHORS PROFILE

Dr. M.N. Doja is Professor in the Department of Computer Engineering, Faculty of Engineering and Technology, Jamia Millia Islamia. He received his B.Sc. (Engg), M.Tech. and Ph.D. degrees from B.I.T, I.I.T. Delhi and Jamia Millia Islamia, New Delhi respectively. His areas of research are Software Engineering, Networks, Security, Simulation, Operating System and Soft Computing. of. Doja is a member of Academic Council, Board of Research Studies and Board of Studies of a number of universities including Ambedkar University Lucknow, NSIT New Delhi, A.M.U. Aligarh, Guru Gobind Singh Indraprastha University Delhi, NIT Jalandhar, Hamdard University New Delhi etc. He has been a member of a number of committee for various universities in various capacities. He has been expert member/member of various committee constituted by UGC and AICTE.

Sapna Jain is a Phd Fellow in the Jamia Hamdard University who has obtained her MCA (Masters of Computer Application) degree from Maharishi Dayanand University, ndia. Her area of research is Scalability of data mining algorithms.

Dr. M Afshar Alam is professor in Department of Computer Science, Jamia Hamdard,New Delhi.He has teaching experience of more than 17 years.He has authored 8 books and guided PhD research works.He has more than 30 publications in international/national/journal/conference proceddings. He has delivered special lectures as a resource person at various academic institutions and conferences He is a member of expert committees of UGC,AICTE and other national and international bodies.His research areas include software re-engineering,data miningbioinformatics and fuzzy databases.

Students' Perceptions of the Effectiveness of Discussion Boards:

What can we get from our students for a freebie point?

Abdel-Hameed A. Badawy
Electrical and Computer Engineering Department,
University of Maryland,
College Park, MD, USA

Abstract— **In this paper, we investigate how the students think of their experience in a junior (300 level) computer science course that uses blackboard as the underlying course management system. Blackboard's discussion boards are heavily used for programming project support and to foster cooperation among students to answer their questions/concerns. A survey is conducted through blackboard as a voluntary quiz and the student who participated were given a participation point for their effort. The results and the participation were very interesting. We obtained statistics from the answers to the questions. The students also have given us feedback in the form of comments to all questions except for two only. The students have shown understanding, maturity and willingness to participate in pedagogy-enhancing endeavors with the premise that it might help their education and others' education as well.**

Keywords- collaborative learning; cooperative learning; peer learning; teaching evaluation; pedagogy; discussion boards; Blackboard; pedagogy; survey; student feedback.[1]

I. INTRODUCTION

The students in both computer science and computer engineering have to take a C language [1] programming course that is worth four credits, which involves a lot of programming. This course requires a lot of effort from the teaching staff to help the students with their debugging issues and to clarify ambiguous, unclear or even mistakes in the requirements of the assigned programming projects. This c programming course has a password-protected website where lecture slides, project descriptions and other related documents can be retrieved. One of the courses that follow the C programming course is a regular three credits course that teaches computer organization, which also has a significant programming component in it as well. In some cases, projects can be as large as several hundreds of lines of code. Course management software is used in the management of the follow-up course. The course management software adopted by the University of Maryland at College Park (UMD) is called Enterprise Learning Management System (ELMS) [3] and is powered by Blackboard© version 8.0 at the time of conducting this study. Blackboard© version 9.1 was released recently [2]. We conducted a simple survey study to assess the perception of the students of the follow-up course to the use of the discussion boards of ELMS.

II. RELATED WORK

The volume of work, in the last few years in education and the theory of teaching and learning, is beyond any single person ability to follow. Many conferences and journals are concerned with Teaching and Learning. It is clear to us that we cannot be very thorough in our coverage of all related work to this work; nevertheless, we will try here to touch on some of the seminal works in the areas related to our work.

In terms of pedagogy, cooperative learning and collaborative learning are the two most closely related pedagogies to our theme in this paper. A great chapter that introduces cooperative, collaborative and peer learning appears in Wilbert J. McKeachie's "Teaching Tips" book [11]. Collaborative learning and cooperative learning are sometimes used synonymously even though in the literature they are different. Cooper discussed the differences and similarities between collaborative and cooperative learning in [18]. Gokhale examined collaborative learning techniques in a study and has shown that collaboration among students enhances learning and increases critical thinking [9]. Felder discusses effective techniques and methodologies for using collaborative learning in teaching [20].

Discussion boards are a commonly available feature in many online course management systems such as Blackboard. Jeong, of Florida State University, has published extensively on topics related to discussion boards in general [12, 13, 14]. He investigated how can online discussion boards engage all students and promote interaction among them [12]. He examined facilitating online discussions effectively [14]. He also designed computer-based tools and methods to analyze discussion boards and give instructors methods to assess, grade and evaluate discussion boards. Northover investigated whether or not online discussion boards are a friend or a foe. He also suggested best practices to develop effective situations that can be easily delivered and assessed [15]. Dringus and Ellis used data mining techniques as an assessment strategy for evaluating discussion forums [16].

A large volume of work has been conducted to investigate the use and effectiveness of technology in the classroom and how do they influence student learning and the ways the professors are teaching. The investigated techniques range from presentation software such as Microsoft's PowerPoint®, course management software, course websites, online lecture

[1] A shorter version of this paper appeared in [4, 17].

notes, and discussion boards to personal electronic devices or what is coined as mobile learning or handheld learning [7, 8, 10, 11, 12, 15, 17, 19, 21, 22].

TABLE I. DISTRIBUTION OF ACCESSES TO EACH DISCUSSION FORUM.

Forum	Accesses	% Total	Students	Staff	Ratio
Lab 1	4154	30%	2756	1398	66%
Lab 2	3332	24%	3241	91	97%
Lab 3	4557	33%	3985	572	87%
Lab 4	876	6%	876	0	100%
Technical	833	7%	700	133	84%
Total	13752	100%	11558	2194	84%

TABLE II. MESSAGES DISTRIBUTION IN EACH DISCUSSION FORUM.

Forum	Messages	% of Total	Students	Staff	Ratio
Lab 1	69	25%	44	25	64%
Lab 2	73	27%	69	4	95%
Lab 3	69	25%	61	8	88%
Lab 4	32	12%	32	0	100%
Technical	31	11%	24	7	77%
Total	274	100%	230	44	84%

There exists a discussion board that students can use for instructor created topics. In case of the course under consideration in this work, each programming project had its own discussion forum. There were four projects (Labs 1, 2, 3 and 4). In addition, there was a forum for technical questions (called Technical) related to using the computer resources for the course. The traffic on the discussion boards for the first project was overwhelming. Therefore, we decided to design a survey to be able to get the feedback of the students about their experience with the discussion boards, its effectiveness and its contribution to their learning.

The rest of the paper is organized as follows: Section III introduces motivational statistics about number of messages accessed or generated. Section IV addresses the survey and its questions, the participation statistics and the logistics of conducting it and finally the results. Section V (Appendix) contains all the short answers the student gave for the open-ended question.

III. Motivational Statistics

ELMS (Blackboard) has built-in usage statistics collection tools and utilities that were very useful in our study. We have used these existing tools in ELMS for all parts of the courseware to see statistics like how many posts happened during the semester and how many times the posts were accessed, read or replied to.

Tables I and II respectively show the usage statistics of the forums in terms of total number of accesses to each forum and the number of unique posts and responses in each forum. In addition, we have broken up each of these according to the total number of events belonging to the students as opposed to the total number of forum posts belonging to the instructional staff.

Each of the tables shows six columns. The headings of the columns are exactly the same for both tables. The first column shows the name of the different forums. There were five forums in this course. There exists one forum per programming laboratory (totaling four programming labs) and a technical questions forum where the students would inquire about any connectivity or accessibility related questions to each other or to the instructional staff. The second column shows the number of accesses and messages per forum. Column three shows the percentage of the accesses and messages per forum with respect to the total number of accesses and messages to all the forums. Columns four and five show the distribution of each forum accesses and messages with respect to who viewed or wrote them whether it is the students or the instructional staff. Column six shows the percentage of the students' accesses and messages to the overall number of accesses and messages per forum.

Examining the reported numbers in tables I and II, one cannot ignore the very large number of accesses to the forums, which is almost 14K accesses relative to the relatively small number of messages exchanged on the forum of less than 300. Using some simple math, the ratio of the number of accesses per posted message is 50 to 1 *i.e.* the average per message views are 50. We need to keep in mind that the total number of students enrolled in this class was 65 students and there were four instructional staff for this course.

We get the following statistics:

1) On average, the number of messages posted per student is about three.
2) On average, the number of messages read per student is 178.
3) On average, the number of messages posted per staff member is 11.
4) On average, the number of messages read per staff member is 549.
5) The ratio of the number of messages posted by all the students to the messages posted by all the staff members is about five.
6) The ratio of the number of messages read by students to the messages read by all the staff members is about five.

We can conclude several conclusions from the above averages and ratios. First, on average, a student read far more messages than what he or she writes or posts.

TABLE III. SURVEY QUESTION (1).

Do you think that ELMS is a helpful learning tool in this course?	**%**
Yes, it is a helpful learning tool. I love it.	64%
No, it is not a useful learning tool. I hate it.	2%
I neither love it nor hate. I am neutral.	34%

TABLE IV. SURVEY QUESTION (2).

Do you prefer courses that use ELMS over other courses with a regular website?	**%**
Yes, courses are better with ELMS.	63%
No, I like non-ELMS courses better.	6%
It does not really matter. I do not care	31%

TABLE V. SURVEY QUESTION (3).

How often do you post on the discussion boards?	**%**
Once Daily.	2%
Once a week.	19%
Once a month.	21%
Once a semester.	26%
Never posted.	13%
I only read but I do not post.	45%

TABLE VI. SURVEY QUESTION (4).

If you have questions, do you prefer to go to office hours or you try the boards first?	**%**
I prefer ELMS board posts.	36%
I prefer to talk to someone face to face.	43%
Depends on the time I have to figure out the answer.	36%
I just ask a classmate.	21%

This means that with careful monitoring for the on-going flow of messages and posts on the forums we can reach the students with crucial clarifications and answers to hairy questions that we know that many of the students will read.

TABLE VII. SURVEY QUESTION (5).

If you posted to the board and a fellow student answered your question, do you trust his answer?	**%**
Yes, sure I trust my classmates.	53%
No, they might be wrong.	9%
Only if someone from the instructional staff says it is a fine.	26%
Not on all issues I trust my classmates' answers.	21%

TABLE VIII. SURVEY QUESTION (6).

How many posts do you read?	**%**
I just read everything.	40%
It is a waste of time. I read nothing.	2%
I read posts depending on the title of the post.	51%
I read posts related to my questions only.	17%

TABLE IX. SURVEY QUESTION (7).

If there is a course with two sections one without an ELMS website and another with an ELMS website, which section would you enroll in?	**%**
The ELMS-based section.	43%
The non-ELMS section.	6%
It is not a factor at all.	51%

Another very important issue to notice here is that the staff members read three times more messages compared to the students, which makes perfect sense. The instructional staff is working hard to follow the discussions to make sure that the correct information is disseminated among the students and nobody is making some wrong, confusing, or misleading replies.

The statistics above suggest that the forums on ELMS are really a collaborative learning tool for the students. The students were the origin of most of the traffic on the forums. The students asked and replied to their own questions except in the rare cases where one of the instructional staff had to step in and correct or rectify a problematic issue. Clearly, the forums are a running archive for the students saving what the questions that were asked previously are and they were able to ask further questions. We can conclude here that the forums are a form of “student-directed online office hours” that are run by the students and monitored by the instructional staff.

TABLE X. SURVEY QUESTION (8).

Currently you cannot submit anonymous questions and/or answers to the boards. If there was that option, would you participate more by reading and/or writing?	**%**
I would have participated more.	30%
It would not matter to me.	55%
I would not trust the posts if they were anonymous.	15%

TABLE XI. SURVEY QUESTION (9).

Did you participate in this survey because of the freebie point?	**%**
Yes	89%
No	11%

TABLE XII. NUMBER OF COMMENTS PER SURVEY QUESTION.

Question	**# of comments**
Question 1	40
Question 2	35
Question 3	28
Question 4	25
Question 5	28
Question 6	25
Question 7	N/A
Question 8	N/A
Question 9	31

IV. THE SURVEY, PARTICIPATION, LOGISTICS AND RESULTS

A. Survey Participation and Logistics

The statistics we have shown so far show how much the forums helped the students. The total number of forum accesses and the other collected statistics are measures of the utility of the forum to the students. In order to increase the participation and reward the students who will participate in the survey we conducted, we gave each participating student a freebie point to be added to the total points a student scores in the course. Effectively, this point is 1% of the course grade.

Forty seven students completed the survey out of the 65 registered students for the class and 53 students attempted it, thus six students did not complete the survey they started. We consider this a great response from the students since it was a 72% completion rate and an 82% attempting rate. The goal at UMD for the end of semester course evaluations is 70% participation. The survey was open for participation for twenty-four hours only. The students of the class have scored a large participation turnout of over 90% in the campus-wide end of semester course evaluations, which clearly shows that we could have gotten better participation rate if the students were given more time to turn in the survey.

B. Survey Results

Tables (III through XI) summarize the results that we have obtained from our survey. The first cell on the top left columns of each table is the question the table results are addressing. Table (XII) shows that we have asked the students to give us their own comments after every question except for questions seven and eight. The number of comments that we got were in many cases was very large. In many instances, we got very thoughtful statements. We will discuss some of the interesting comments. We are including all of the comments in the Appendix in section V.

C. Discussion of the Results

In this subsection, we will go over some of the conclusions from the results reported in the tables. Table IV (Question 2) suggests that more than 60% of the students prefer ELMS over non-ELMS courses. Table V (Question 3) suggests that most of the students read the posts but do not write as much. Table VI (Question 4) suggest that ELMS is seen by students sometimes as a replacement of office hours. Table VII (Question 5) suggests that most of the students trust their fellow students for answering their questions. Table VIII (Question 6) suggests that about half the students read all posts and the other half reads posts depending on their titles. Table IX (Question 7) suggests that a very small percentage of the students prefer the non-ELMS sections, whereas close to half of the students prefer ELMS and a little above the half of the students do not bother whether the course is ELMS or not. Table X (Question 8) suggests that anonymity of the posts is not a factor for the students. Table XI (Question 9) suggest that most of the students found that the freebie point was a good push for them to take the survey.

D. Limitations of the Study

This is study is speculative at many of its conclusions, yet there is plenty to draw from it. The sample size of the survey is small (around sixty students). We should have worded more tightly to get precise answers. For example, we should have asked how ELMS is useful to the students’ learning. In some question, we allowed the students to check all that applied and thus in some questions the total of the percentage statistics is above 100%. This study should have happened along several years to remove any superfluous data or any jitter in the results. Two surveys should have been done in both courses of the sequence courses both of them not the junior level course only. Given all these limitations, we still think that the conclusion of the paper are relevant and useful for any computer science instructor who is teaching programming courses or courses with a heavy project and lab components.

E. Contributions

This paper aimed to articulate for faculty the importance of communication with the students. The students are willing to sacrifice their time to give us feedback. As faculty, our teaching style and pedagogy should be fully shaped and directed by the students. A continuous feedback loop should exist between student perceptions of different techniques and the adaptation of the technique(s). As faculty, we should disseminate our findings so that our fellow instructors can benefit from our experiences and experiments in the field of teaching and learning as well as in any other scientific discourse.

F. Selected Student Comments

Table XII summarizes the number of open-ended answers for the nine questions. The number of open-ended question range from 25 to 40 answers. We share in this section some of the student comments that are very thoughtful and interesting. Here are a selected set of the students' comments. Appendix I at the end of the paper contains a full listing of the students' comments organized per each survey question.

- "I do like giving feedback on pedagogy. Student feedback is hard to get/give in a big lecture hall. It benefits and improves the quality of the educational environment".
- "ELMS is alright, but if all the features were used, it could be (even) better".
- "I would have taken the survey without the point".
- "I also do feel that this survey is important".
- "Since there is no anonymity (in the boards), I am fairly certain no one would knowingly pass on false information (since their name associated with the post)".
- "If a teacher puts lecture notes and other materials on ELMS is helpful. Other teachers of mine has said they would use ELMS and then after 2 weeks they just gave up and just used emails".
- "Discussion boards are the most useful tool".
- "In the boards, I ask general questions to draw from the knowledge of the entire class rather than focusing all questions to TA's and the professor".
- "I think it would be helpful if all courses use ELMS".
- "(ELMS is) also nice for viewing grades and the syllabus and assignments".
- "Someone can post answers that are not perfect but still he/she knows better than I do".
- "If I can see the reasoning (in a classmate's answer of a question) and it makes sense, I can trust their answer well enough".
- "It (ELMS and its associated Discussion boards) makes classes without discussion sections easier to follow".
- "I do wish they would upgrade (the ELMS) GUI, it feels like a 90's web app".
- "Some classes do not use ELMS very well and others are very organized and utilize it".
- "(My issues with ELMS): you can't save your password to log into ELMS through google chrome which is painful".
- "(ELMS is) a great idea, but the interface is terrible".
- "When I have a question that is not answered by the discussion boards, I assume it is a question only I have and I will then go to a teacher to ask it".
- "Almost always, the face to face benefits are better than ELMS".
- "It is usually easier and more convient to post a message to the discussion board".

ACKNOWLEDGMENT

The author would like to acknowledge Dr. Michelle Hugue of the Computer Science Department, University of Maryland, College Park. The author also wants to acknowledge all the staff at the UMD's Center for Teaching Excellence (CTE) especially Henrike Lehnguth, Dr. Dave Eubanks and Prof. Spencer Benson. Without their feedback and support this study would not have materialized. Ms. Mary Maxson of the R. H. Smith School of Business' Information Technology for helpful feedback and suggesting to check a technology survey done at the School of Business in UMD. The author would also like to thank the anonymous reviewers for their feedback and insightful comments.

REFERENCES

[1] B. W. Kernighan and D. M. Ritchie, "The C Programming Language", Prentice-Hall, NJ, 1988.

[2] "Blackboard inc. website", http://www.blackboard.com/ Last accessed July 23rd 2012.

[3] "The University of Maryland collaborates with Blackboard Inc. to offer a single, campus-wide teaching and learning system", http://www.umd.edu/umnews/blackboard.html appeared online July 2006. Last accessed July 23rd, 2012.

[4] A. Badawy, M. Hugue, "Determining the Effectiveness of Discussion Board Interactivity", University of Maryland's 4th Innovation in Teaching and Learning Conference, College Park, MD, April 2010.

[5] C. Elam, T. Stratton, and D. Gibson, "Welcoming a New Generation to College: The Millennial Students", Journal of College Admission, vol.195, pp.6, 2007.

[6] M. E. Wilson, "Teaching, learning, and millennial students" New Directions for Student Services Volume 2004, Issue 106, pp. 59-71.

[7] Y. Leung and M. Ivy "How Useful are Course Websites? A Study of Students' Perceptions," Journal of Hospitality, Leisure, Sport & Tourism Education, Vol. 2, No. 2, 2003.

[8] M. McSporran, "Online Learning: Which Strategies do New Zealand Students perceive as most Valuable?", In Proceedings of the 21st ASCILITE Conference, 2004.

[9] A. A. Gokhale, "Collaborative Learning Enhances Critical Thinking" Journal of Technology Education, Volume 7, Number 1, ed. Mark Sanders, Fall 1995.

[10] S. J. Coopman, "A critical examination of Blackboard's e–learning environment," First Monday, Volume 14, Number 6, 2009.

[11] W. J. McKeachie, M. Svinicki and B. Hofer, "McKeachie's Teaching Tips: Strategies, Research, and Theory for College and University Teachers," 12th Edition, Boston: Houghton Mifflin Company, ISBN: 0-618-51556-9, 407 pages, 2006.

[12] A. Jeong "The Sequential Analysis of Group Interaction and Critical Thinking in Online Threaded Discussions", American journal of distance education, Volume 17, Number 1, pp. 25-43, 2003.

[13] A. Jeong,."Computer-Based Tools & Methods for Assessing Group Discussions and Critical Thinking Processes", presented at the Association for Educational Communications & Technology, Dallas TX, November 2002.

[14] A. Jeong, "Facilitating Online Discussion", presented at the 16th Annual Conference on Distance Teaching & Learning, Madison WI, August 2000.

[15] M. Northover ,"Online discussion boards--Friend or foe", In Proceedings of the 19th Annual Conference of the Australian Society for Computers in Learning in Tertiary Education, pp. 477-483, 2002.

[16] L. P. Dringus, and T. Ellis, "Using Data Mining as a Strategy for Assessing Asynchronous Discussion Forums", Computers & Education, Volume 45, Issue 1, pp. 141-160, August 2005.

[17] A. Badawy, M. Hugue, "Evaluating Discussion Boards on BlackBoard© as a Collaborative Learning Tool: A Students' Survey and Reflections", in proceeding of the 2010 IEEE International Conference on Education and Management Technology (ICEMT 2010), Cairo, Egypt, November 2010.

[18] J. L. Cooper, P. Robinson, and D. B. Ball (Eds.), "Small Group Instruction in Higher Education: Lessons from the Past, Visions of the Future", Stillwater, OK: New Forums Press, 2003.

[19] E. Perry, "The Impact of Online Lecture Notes on Learning Outcomes of Beginning Thermodynamics Students", In Proceedings of the American Society for Engineering Education Annual Conference, 2007.

[20] R. Felder, "Effective Strategies for Cooperative Learning", Journal of Cooperation & Collaboration in College Teaching, Volume 10, Number 2, pp. 69–75, 2001.

[21] R. A. Bartsch, and K. M. Cobern, "Effectiveness of PowerPoint Presentations in Lectures", Computers & Education, Volume 41, Issue 1, pp.77-86, August 2003.

[22] S. Perkins, and G. Saltsman, "Mobile Learning at Abilene Christian University: Successes, Challenges, and Results from Year One", Journal of the Research Center for Educational Technology [Online], Volume 6, Number 1, (4 March 2010).

V. APPENDIX I

We archive here the reflections of the students to each of the survey questions included in the survey. Some little editing occurred to maintain privacy and to hide names in some cases.

1) *Student Comments for Question 1*

The question was: Do you think that the ELMS website is a helpful learning tool in this course?

- Good for teachers who do not have their own web pages.
- A place for discussions and lecture notes is good.
- I think ELMS is very helpful for courses for the freshman or sophomore. But for a small class, I never use ELMS.
- I have found the discussion board to be the most useful tool on blackboard. It helps to be able to ask general questions to draw from the knowledge of the entire class rather than focusing all questions to TAs and the professor.
- I would say that I am neutral but I think that the discussion board is a very useful tool for any class.
- Yes, but it would be more helpful if grades were kept up to date and accurate.
- It is often slow, and it not kept particularly organized.
- Always knew all the information I needed could be found somewhere in ELMS.
- When used well by the teachers (all useful documents posted, grades posted, etc.), it is a very valuable resource for information. Using the documents posted and the lecture notes are crucial to studying for exams in the course, I've found, as everything is there to be utilized, even if you don't happen to have your book with you at all times (especially then!). If you missed something important in class, ELMS lets teachers post the information in a place accessible at any time.
- The course documents are useful references, and I like being able to see my grades. I do not use the discussion board much but I probably would if I did not know other students in the class.
- Its information resources are much more useful than both the lectures and the textbook.
- Some classes do not use ELMS very well and others are very organized and utilize it. Also, you cannot save your password to log into ELMS through Google chrome, which is painful.
- Yes, there are lots of helpful links and material all compiled together in one site to make finding everything easier.
- I like the departmental grades server and forums better.
- A great idea, but the interface is terrible. We may as well have a website with frames and an associated forum site! Computer Science is supposed to do this sort of thing better than the humanities.
- ELMS, in general, tends to be slow and difficult to navigate. Combined with mediocre organization because of old things from different semesters it is very annoying.
- I believe that for this tool to be extremely effective, we need updates for the current semester to be present. Mainly, this pertains to having the projects sent to ELMS as opposed to our emails.
- It is just as effective as a class website with lectures and other material like that. There are plenty of tools that help the class, like the discussion board.
- Having everything in one-place saves a lot of time.
- I find courses that use ELMS pretty helpful since there is a place with all of the course information, unlike other courses where I simply have a syllabus. The only reason I did not say I love it, is because sometimes the site can be rather disorganized. It can be difficult to find what you are looking for sometimes without looking through every module.
- I like how many of my courses can be integrated into one website
- I like it because of the discussion boards.
- It's very helpful for discussion purposes
- Like the forum section and having, all class assignments and materials in one place.
- For some classes it is important, for others it is not. It is only helpful when you are updating the site often, have a live discussion board, and keep important documents and grades online.
- I think ELMS is great for posting lectures slides and practice material, but I am not a big fan of the discussion board. I think some sort of separate forum would be much more useful.
- For this course, I would rather have everything be on the grades server and through email.
- It is not life changing or anything but, it is a good tool to use. I would give it an eight out of 10.
- The ability to have course information organized in one central location is extremely helpful. Also having the grade book updated makes it easy to keep track of how you are doing in the class.
- The professor actually utilizes ELMS and most of its features to make it worthwhile.
- I think ELMS course website is a helpful learning tool. However, it is great if the user name and password can be saved automatically, so I do not need to enter username and password every time I visit my courses in ELMS.
- Good to have stuff be posted and in a structured way.
- Everything is organized in a readable fashion.
- I know the instructor of this course thus, I know that if the information she had on ELMS was not here it would be on her webpage. Other teachers however if they did not have ELMS would probably email you everything and if you lost the email then you would be out of luck. So in those cases it can be helpful
- The tools for organization are not always clear, but having all (or the vast majority) of necessary online materials clumped in one place is very useful.
- ELMS has great resources, easy to check grades and online tests are awesome
- It's a useful learning tool but having information on both the instructor website and on ELMS makes it difficult to know where to go to get things
- ELMS is all right, but if all the features were used, it could be better.
- Course notes and the discussion board are really useful, as well as sample exams.

2) *Student Comments for Question 2*

The question was:

Do you prefer courses that use ELMS over other courses with a regular website?

- ELMS is annoying.
- Having a webpage is essential, though whether it is through ELMS or through a class-specific webpage does not really matter.
- It makes classes without discussion sections easier to follow.
- Especially for Computer Science classes, it does not matter because most CS teachers have their own class website and we have our own grades server so essentially all the functions of ELMS are available through the CS department specific tools.
- Only slightly better. A good website can be very good as well.
- Most other professors fail at using ELMS properly.
- ELMS lets you have a discussion board.

- I get distracted when I have to open up multiple sites for a course. It is nice to have it all in one frame where I can also access my other courses.
- Helps to have a central location for information.
- All depends on how much work the teachers put into the page. If a teacher puts lecture notes and other materials on the page then it is helpful. Other teachers of mine have said they would use ELMS and then after 2 weeks just stopped using it and resort to emails.
- It is nice to have all class websites on one page versus about 3 or 4 pages.
- Some courses benefit from ELMS more than others do. This course benefits because there are general questions about projects. Questions for this course cover more than just course material. For example, programming questions that do not directly apply to course material but are still required for projects can be asked and answered freely. Courses where questions are more material specific would be harder to pose questions to the class.
- If instructors do not use ELMS, but create their own websites with enough information it is usually okay, but often not using ELMS means not having any online source for the class.
- It is easy to bookmark classes that have individual web pages while with ELMS you always have to log in each time you access something.
- I think ELMS is a straightforward, easy to use tool that helps me access material quickly.
- In particular, I think courses with active discussion boards are the best.
- They are better, because you can get all the information you need, view your grades, etc. all in one place.
- The features I use in ELMS are also available in classes with regular web sites.
- I prefer courses that use ELMS since it has many tools such as assignments, related coursework, discussion board, and other tools and so on.
- I prefer courses that have ELMS.
- If teachers are able to maintain an updated website on their own, that is great. If not, then ELMS provides an okay way to do it.
- A website specifically designed for the course will always be better than something that tries to be generic and work for every class.
- I always know where to find information as long as the teacher updates it.
- As long as everything is in one place, it is better than not having it.
- I like being able to go to one single website and look up information for all my classes instead of having to go to several different websites.
- Some classes use ELMS well, while other barely put any materials on ELMS or use it only to distribute email material.
- Online materials are always a plus.
- My course choices are sharply limited, and it depends strongly on the professor's proactive website upkeep.
- ELMS would be even better if teachers updated the grades section so people can know how they stand in the class.
- Of course, a website is good, but it does not need to be ELMS.
- As long as, I can access my grades, but not on ELMS, I am happy.
- Courses without ELMS pages - or courses that don't utilize ELMS well - definitely lack the benefit of being able to have all the useful, relevant and important information and material easily accessible by students at any time. Being able to access information I may have missed, or misheard, or did not understand, is VERY helpful. I very much prefer classes that have this benefit.

3) *Student Comments for Question 3*

The question was: How often do you post on the discussion boards?

- I look for questions and answers that relate to my tasks. However, I will post if I have questions or I know the answer, so it does not depend on the timeframe.
- I never post question because I prefer to ask a TA or professor through e-mail or in person if I got a question.
- I do not like to post because it shows my name!
- If I have the answer, I answer quickly, and if I have found answers to questions, I previously asked on the discussion boards.
- I did not post many posts because for programming project one and two, I did them relatively early compared to other students, and I did not look at the discussion posts afterwards.
- The discussion boards' questions are answered too slow and may not be accurate. The teacher almost never comes on it.
- The kinds of questions I have are usually ones that are not allowed to be answered...so.
- I think the discussion boards could be vastly improved. It just feels clunky and outdated. A forum layout is much better to work with. Especially with programming discussions.
- People are rude on the discussion boards at times.
- Mostly when required for classes.
- I am sort of a shy person. I answered someone's question once, but I feel as a retard for some of the questions I want to ask. It is a great idea, but a lot of times, having an attached drawing or something for an explanation would be a good idea.
- Most of time, I rely on others asking good questions and learning from their questions instead of asking my own.
- Simple technical questions may be posted. Other than that, I prefer face to face.
- I will post if I am truly confused about something.
- In general, this is a lot easier than being spammed with fifty e-mails about student questions that apply to more people.
- Definitely good when there are no office hours available.
- Would rather attend office hours.
- I only post occasionally when I have questions but not too often. Usually I just look at other people's questions.
- Semi-daily when nearing the end of projects, otherwise almost never.
- I check the message boards a lot, but most of the questions I have are already answered there and I therefore do not need to post messages.
- I rarely post my own questions because one such question usually already exists.
- Mostly look for answers, rarely post
- The discussion board is a great resource.
- Most of my classes that I have taken that have used ELMS; do not extensively use the discussion boards.
- I would post anonymously.
- I have posted before but I do not post often in THIS class. For ELMS in general, I took a class entirely based on using ELMS, and I posted every day in those discussion boards. I do not know if that is relevant or not.

4) *Student Comments for Question 4*

The question was: If you have questions, do you prefer to go to office hours or you try the boards first?

- It is clearer, if I can talk with someone face to face, but sometimes, information in the discussion board really help me, and I can see others have the same troubles.
- Almost always, the face-to-face benefits are better than ELMS. For example, you get immediate, certain, and complete information when you sit down with a teacher and hash out a specific topic that is difficult.
- It is usually easier and more convenient to post a message to the discussion board.
- If I am nearby the TA office hours already, I will check in with them, else the discussion board is much more convenient.
- It depends on the question.
- Usually, I will ask a classmate first. If I still cannot figure it out, I might try office hours or the boards.
- It is complicated sometimes to get your point across through a message. I also like talking face to face because I never know what I can and cannot put online in terms of code.
- Better to talk in person on campus.
- I don't think a discussion board can ever be as good as talking to a TA
- ELMS is more convenient than going to office hours because it is all electronic, and some office hours are during student class time.
- Usually the answer is either not allowed to be explained, or it is too complicated to explain over discussion boards. Asking in person is usually better for me.
- I cannot really attend the office hours; I have classes at those times.
- I look for questions all over the place, and ELMS having a discussion board for that purpose definitely saves me a lot of time and effort compared to if it were not there.
- I look on all fronts.
- When I have a question that is not answered by the discussion boards, I assume it is a question only if I have and I will then go to a teacher to ask it.
- It depends on the question.
- Usually, the TAs help better than just a post on ELMS.
- The discussion board posts are the fastest method of getting the information but office hours goes more in-depth.
- If it is a simple question, I like ELMS. If it is more of a difficult question that can only be answered face to face then I will go that option.
- I prefer to go to office hours, but it is a long drive for me, so I check ELMS first.
- I prefer talking to someone if I actually have a question.
- If I have question I usually try to contact a TA first.

- I feel the forum questions are generally, "what is this" rather than the more detailed/intricate "how do I do this", which requires a face-to-face discussion.

5) *Student Comments for Question 5*

The question was: If you posted to the board and a fellow student answered your question, do you trust his answer?

- I believe that the students who actually take the time to answer questions will know the answer correctly.
- I trust everyone!
- Unless they say something along the lines of "I'm not sure", I figure if they bother to post the answer they got it to work for themselves.
- I trust the answer of my classmates. Although there would be some wrong answers, other classmates suggest correct answer.
- That is true. My classmate might give the wrong answer since his/her posts based on his/her knowledge that is not 100% correct; however, sometimes he/she posts other sources or his/her reviews for the previous tasks is a great information.
- Since there is no anonymity, I am fairly certain none would knowingly pass on false information.
- I trust the answer if it makes sense.
- I cannot take everything for granted.
- An incorrect answer would be corrected by another student or by the instructor.
- If I can see their reasoning and it makes sense, I can trust their answer well enough. If it sounds completely farfetched or I am not sure of its accuracy, I will probably ask someone else what they think of the answer before trusting it.
- I usually just try it and see if it works, and then I will know whether I can trust the person.
- The good thing is, with a public board even if someone does make a mistake another student or the instructor can always come along to clarify.
- Also, it depends on which classmate it is.
- Sometimes I trust an answer. If I know the classmate and how smart they are or if I know the classmate has been wrong before and is not the best student then probably not.
- Usually I do unless it sounds too outrageous to be true, which has not been the case so far.
- I do trust my fellow students, but I would feel much more confident about the answer if someone from the instructional team approves.
- I trust if none has corrected the individual or has confirmed its correctness even more.
- It depends on the way in which the answer is presented and how well the answer is supported, in short, on the presentation of the answer.
- I trusted the TAs and the professor.
- I would prefer to verify a classmate's answer before putting it in practice, time permitting.
- Most of the time, I would assume they are correct unless something seems off to me.
- They are usually right and it is not hard to double check.
- Someone who can post answers that are not perfect but still they know better than I do.
- Most of the time students only post when they know the answer, but it is always assuring to know the instructional team says it is the right answer.
- Answers are tested by trying them to see if they are right. Usually they help to see another perspective.
- I would compare to my own answer.

6) *Student Comments for Question 6*

The question was: How many posts do you read?

- I read everything, but everything I read does not actually help me at all.
- Similar questions or interesting posts.
- I typically look for things that I am having trouble with and look for similar title names.
- I read most things; in case there is, something I might have misunderstood that is clarified.
- I skipped the post saying "Please, give us an extension." :)
- I read everything on the topics I am confused about.
- A number of topics I expected to be inane have proved (marginally) useful, and I'd rather not miss information I didn't know because I didn't expect to find it there.
- I try to at least skim every post in the discussion boards - any little bit of information may be useful, and even if I do not delve into the question at the time, I may remember it later and find it useful then.
- For example, I will not look back to posts of a project after I finished it.
- point epends on how much free time I have.
- It always help to read everything
- I will read most of the posts, unless it is about specific questions that I already know the answers to them.
- In my opinion, the majority of the posts I have read in the past are made up of either not so intelligent questions or non-helpful answers, so they are not really of much value.
- I try to read all the posts everyday!
- I have them emailed to me automatically.
- Discussion board is always fun to read.
- If the post is dealing, with nothing I need then I do not read it.
- I only read posts that are relevant to the information I am looking for.
- If it is something I know will not matter, I do not read it. But in general, I like to read almost everything.
- That question was pointless.
- I usually take a peek at every post.
- I read posts posted by either the instructor or the TAs.
- Not exactly, I read all the posts that relate to my concerns, and other questions that I find helpful or can answer them.

7) *Student Comments for Question 9*

The question was: Did you participate in this survey because of the freebie point?

- I am not going to say "false" to the previous question - who does not want an extra point? – I most likely would have taken the survey with or without being given an extra point - It is not like it is all that hard to do.
- I think ELMs is pretty good. I do wish they would upgrade the GUI, it feels like a 90's web application.
- HELL YEA!
- I participated because of the point, but I would have given feedback either way.
- Thanks for the extra point but I do feel like it was valuable to give the students input on ELMS.
- If points were not an incentive, I'm not sure I'd get this done in time because of what else is going on this week, but I hope to broaden ELMS (or an improved version of ELMS) usage in order to promote smarter learning.
- Extra points are attractive.
- It did not hurt, but I do like giving feedback on pedagogy. Student feedback is hard to get/give in the big lecture hall.
- Nevertheless, I DID answer them honestly!
- Cool survey.
- Probably would have done it eventually anyways
- I think it would be helpful if all courses have ELMS support. It's also nice for viewing grades and the syllabus and assignments.
- I do hope my feedback helps.
- Mostly, but I probably would have done so anyway if I thought it would help someone out.
- POINT!!!
- If the instructor had said that I should do it but had not offered the point, I still would have done it. If it were something that was not recommended by the instructor then I might not have done it.
- Point!
- A simple cost-benefit analysis suggests that an extra point is easily worth answering a few questions.
- While I did participated for the extra credit, I answered all questions truthfully and to the best of my ability.
- It is an incentive; incentives such as this always get more people to do surveys. It is a matter of deciding what my time is worth. Is 5 minutes of my time worth nothing to me or is it worth a few points. I like to think my time is worth something.
- However, I also do feel that this survey is important, when I started at Maryland, only one or two of my classes used ELMS as opposed to now, when I never have a class that does not use it.
- True, however I answered all questions according to how I feel and did not rush through it. :)
- I need the extra points as the instructor said.
- I need every point I can get. Don't hesitate to offer another one of these, because, I am sure my classmates would welcome it as would I.
- Nevertheless, it is also my benefit and improved the quality of the education environment.
- Hey, at least I am honest.
- It is nice to get the point, although I would have taken the survey without the

point.
- Extra points, yay!

AUTHORS PROFILE

Abdel-Hameed A. Badawy, an ABD PhD Candidate at the Electrical and Computer Engineering Department, University of Maryland, College Park, MD. Mr. Badawy obtained a B.Sc. in Electronics Engineering with a specialization in Computing Systems and Automatic Control systems from Mansoura University, Egypt in 1996 with Distinction and Honors. He was ranked the first on his graduating class in his specialization. He obtained a graduate certificate in software development from the Egyptian Cabinet sponsored, Information Technology Institute, Giza, Egypt in 1997. He was a Teaching Assistant at the Systems and Controls department at Mansoura University from September 1996 till August 1999. He obtained his M.Sc. in Computer Engineering at the University of Maryland in August 2002 and defended his PhD proposal and advanced to candidacy in August 2006. Mr. Badawy is a Teaching Assistant Training and Development (TATD) Fellow at the Electrical and Computer Engineering Department for the academic years 2010-2011 and 2011-2012. Mr. Badawy is currently a Center for Teaching Excellence (CTE) Graduate Lilly Fellow. Mr. Badawy is a student member of IEEE, ACM. HE is also a member of the Science and Engineering Institute (SCIEI), a member of the International Association of Computer Science and Information Technology (IACSIT), an associate editor for the August 2010 issue of the Journal of Learning, reviewer for the International Journal of Advanced Computer Science and Applications (IJACSA), reviewer for the Society of Imaging Informatics in Medicine (SIIM) Journal of Digital Imaging (JDI), reviewer for the International Journal of Computer and Electrical Engineering (IJCEE), and technical program committee member for several conferences. Mr. Badawy's published research has won best student paper awards at AIPR 2010. He won the prize of excellence at the University of Maryland, Graduate Research Interaction Day 2004 (GRID). He also won a best poster award at GRID 2012. Mr. Badawy's research interests span computer architecture, compiler optimizations, machine intelligence applications in medicine including but not limited to medical imaging, and engineering and computer science Education.

The Impacts of ICTs on Banks

A Case study of the Nigerian Banking Industry

Matthew K. Luka
Department of Electrical and Information Engineering
Covenant University Ota, Ogun State, Nigeria

Ibikunle A. Frank
Department of Electrical and Information Engineering
Covenant University Ota, Ogun State, Nigeria

Abstract— **ICT has taken the center stage in almost every aspect of human endeavor. ICT help banks improve the efficiency and effectiveness of services offered to customers, and enhances business processes, managerial decision making, and workgroup collaborations, which strengthens their competitive positions in rapidly changing and emerging economies. This paper considers the impacts and trends of ICTs on the banking industry of the 21st century. Four (4) parameters, namely: productivity, market structure, Innovation and value chain were used for benchmarking. Case studies of the IT platform employed by two Nigerian banks were included to for a more informed inference.**

Keywords- Banking industry; CBN; Customers; Economic growth; ICT; productivity.

I. Introduction

One of the modern yardsticks used for rating a modern business enterprise is its ICT infrastructural layout. This is an indication of the importance of ICT for business establishments. Banks in particular adopt information and communication technology to improve the efficiency and effectiveness of services offered to customers, improve business processes, as well as to enhance managerial decision making and workgroup collaborations. This helps strengthen their competitive positions in rapidly changing/emerging economies. Environmental, organizational, and technological factors are creating a highly competitive business environment in which customers are the focal point [1]. Furthermore, these factors can change quickly, sometimes unpredictably. Thus, the growth of any enterprise is tied to retaining loyal customers, improving productivity, cutting costs, increasing market share, and providing timely organizational response. ICT is a major enabler for dealing with these issues. Because the pace of change and the degree of uncertainty in today's competitive environment are accelerating geometrically. Organizations are operating under increasing pressures to produce more, using fewer resources. in order to succeed (or even merely to survive) in this dynamic world, companies must not only take traditional actions such as lowering costs, but also undertake innovative activities such as changing structure or processes and continuously revising competitive strategies.

ICT affects all processes associated with modern day banking. From the daily routines of preparing payroll and order entry, to strategic activities such as the acquisition of a company, ICT surfaces as a key element. In View of the importance of ICT in the banking industry, a number research works have been carried out. In [2], an evaluation of the response of Nigerian banks to the adoption of ICT was presented. In [3], a technical model that to ascertain the impact of ICT on the Nigerian banking sector as a function of banking reforms was proposed.

Some benchmarks for evaluating the impact of ICT in the banking industry were outlined in [4]. These benchmarks will be used to evaluate the impact of ICT on the Nigerian banking industry.

II. The Nigerian Banking Industry

The Nigerian banking industry is regulated by the Central Bank of Nigeria (CBN). The major players in the industry are the 22 commercial (deposit) banks and 906 Micro-finance institutions. Other financial institutions that complement banking services include 5 discount houses, 5 development finance institutions, 731 bureau de change, 102 Primary Mortgage Institutions, and 82 finance companies [5], [6]. The Nigerian banking Industry has been undergoing major changes, reflecting a number of underlying developments. Advancement in communication and information technology has facilitated growth in internet-banking, ATM Network, Electronic transfer of funds and quick dissemination of information.

Structural reforms in the banking sector have improved the health of the banking sector. The reforms recently introduced include the enactment of the Securitization Act to step up loan recoveries [7], establishment of asset reconstruction companies, initiatives on improving recoveries from Non-performing Assets (NPAs) and change in the basis of income recognition has raised transparency and efficiency in the banking system. Spurt in treasury income and improvement in loan recoveries has helped Nigerian Banks to record better profitability. Reforms have compelled banks to improve the utilization of ICT. The recently introduced punitive 'handling charge' on cash based transaction by the CBN is a pointer to the ever increasing role of ICT in the Nigerian banking industry.

III. Research Methodologies

The aim of this study is to ascertain the level of use of ICT infrastructures and their impacts on customer service; which invariably determines growth of banks. Considering ICT as a growth enabler, the extent of deployment by banks and customers' perception of its relevance are the basis of the research.

A random sampling technique was used to issue questionnaires to customers in the selected banks. Four commercial banks were selected on the basis of sufficient branch networks. A total of 400 questionnaires were given out to customers at the bank premises. About 280 of the questionnaires were returned to the researchers, a response rate of 70%. The four banks visited are: Guaranty Trust Bank plc, First Bank of Nigeria plc, Zenith Bank international and United Bank for Africa (UBA). The response were measured with a 5 pointer likert - type rating, where strongly agree (SA) = 5; Agree (A) = 4; Neutral (N) = 3; Disagree = 2; Strongly Disagree = 1. See appendix 1 for more details on the questionnaire. Two case studies are included to validate the findings.

IV. Data Analysis And Interpretation

The major dimensions of the banking industry for which the impact of ICT has a critical consequence include: productivity, innovation dynamics, market structure, and value chain characteristics.

A. Impacts of ICT on Productivity

ICT has productivity increasing effects on labor productivity and total factor productivity of companies. ICT-induced productivity effects vary significantly between sectors and among countries. The banking industry is one of the sectors that enjoy the largest productivity growth effect of ICT.

Table 1 show the effect of ICT on the productivity of the banks as perceived by customers. About 85.4% of the respondents agree that ICT is helping the cahiers to be more productive. The use of computers and peripherals simplifies the task of getting customers' data and counting money to effect transaction. This enables a single cashier to serve thousands of customers in a day which would have cost the bank enormous staff strength and large building. However, about 80% of those interviewed agree that the bank needs to improve its services. This is indication of the fact that ICT investment does not lead to productivity growth at firm-level by itself. It depends on how the technology is actually used in business processes, i.e. on a company's ability to innovate its work processes and business routines with support of ICT. Thus, banks need to multiplex ICT investments with complementary investment in working practices, human capital, and firm restructuring to optimize its impact on productivity.

TABLE 1. EFFECT OF ICT ON PRODUCTIVITY

Question	SA	A	N	D	SD
Computers are helping the tellers in their work	40.0	45.4	5.5	9.1	0
The bank needs to improve its services	51.0	29.0	11.0	8.0	1.0

B. Impact of ICT on Innovation

A technological change such as the massive diffusion of ICT represents an interesting case for an analysis with respect to firms' innovation strategies. For example, it is said that industry leaders often reject important inventions and fail to bring them to the market [8, 9]. Entrepreneurial companies are more likely to exploit these opportunities. Entrants frequently introduce products or production processes based on a new technology, which can challenge incumbents or even drive them out of the market [10]. This was the scenario that played out in the Nigerian banking industry with the emergence of new generation banks that introduced innovative products and services. Innovation in this context aims to reduce the cost of banking while making the process of transaction easier and more convenient.

About 67.5% of the respondents disagree that the amount the bank charges on transaction is okay. This indicates that banks need to come up with innovative products that will reduce the cost of banking operations; which can be passed down to the customer in the form of reduced charges. About 65% of the respondents enjoy information update about the bank through SMS and email alerts. This enhances customer royalty and confidence. 78% of those interviewed agree that they prefer to use the ATM than coming into the banking hall. This is due to extensive publication that has encouraged the use of ATM. Thus banks can encourage the use of other ICT media such as the internet and POS which enhances cashless banking.

TABLE 2 IMPACT OF ICT ON INNOVATION

Question	SA	A	N	D	SD
The amount the bank charges on transactions is okay.	10.0	12.5	10.0	40.0	27.5
I enjoy information update about the bank through SMS and email alerts	30.0	35.0	10.0	12.0	13.0
I prefer to use the ATM than coming into the banking hall.	48.0	30.0	9.0	7.0	6.0

C. Impact on Market Structure

Innovations enabled by ICT changes the cost structure of companies. Hence, innovations have a significant impact on the market structure in which companies operate. Radical changes in technology traditionally lead to emergence of new products or change the production processes of existing products. In either case, companies face a large degree of uncertainty regarding future demand or cost of service delivery [5]. Furthermore, during times of technological change, mergers reflect the process of assets reallocation toward more efficient firms [11]. The mergers that were recently evidenced in the Nigeria banking industry were actually a result of the consolidation exercise of 2004 and the technological change that dawned on the industry.

Technological change forces firms to adopt new modes of production and, consequently, to reorganize its assets. If a company fails to reorganize internally, it will probably disappear from the industry and its assets will be reorganized externally. New technology spreads faster if such asset reallocation works smoothly [12]. The diffusion of ICT is technological change that has greatly revolutionized the banking sector.

Table 3 below indicates that 75% of the respondents agree that the banks have improved the quality of service rendered. This is necessary for the bank to retain its customer as well as attract potential ones. 79% of the customers agreed that they enjoy prompt and efficient service for which 87% of the respondents are willing to recommend the bank to others.

TABLE 3 IMPACT OF ICT ON MARKET STRUCTURE

Question	SA	A	N	D	SD
The quality of service has improved in this bank	50	25	10	14	1
I enjoy fast, efficient and prompt service	49	30	11	10	0
I can recommend this bank to someone	55	32	8	5	0

D. *Impact on Sector Value Chain*

Empirical findings suggest that some of the main effects of ICT diffusion are organizational changes and the redefining of organizational boundaries [12]. Thus, it is relevant to assess if the diffusion of ICT in the banking industry had any impact on the restructuring process. The impact on value chain reflects in re-shaping firm boundaries and changing the constellations of value chains are enormous.

From table 4, only 47% of the respondents agree that value added services/special accounts encourage them to patronize the bank. This customer perception needs to be improved upon by more extensive publications on these value added service so as to complement the impact of ICT. The number of branches a bank has is another value chain that enhances the impact and level of deployment of ICT.

TABLE 4 IMPACT OF ICT ON BANKING SECTOR VALUE CHAIN

Question	SA	A	N	D	SD
Special services/account types encourage me to patronize this bank	27	20	5	28	20
The number of branches this bank has motivated me to chose it	30	35	5	13	17

V. CASE STUDIES

To further study the impact of ICT on growth of the Nigerian banking industry, a case study of two large banks and their choice of an ICT platform will examined [13, 14].

A. *First Bank of Nigeria (FBN)*

First Bank of Nigeria Plc (FBN) was established in 1894 and has distinguished itself as a leading banking institution and a major contributor to the economic advancement and development of Nigeria. With 339 branches, the Bank maintains the largest branch network in the banking industry in Nigeria.

At the turn of the bank's century, FBN found itself in a unique position as, despite its size and reputation, there were challenges to maintain the leadership position in a market that was as dynamic as it was competitive. It was at this point that the bank launched its business transformation initiative called 'Century II'. Century II clearly identified IT as an enabler for the bank going forward. The Key Business Drivers for an ICT platform were:

- Need to Integrate Banking Operations:

The bank's 300+ branches were operating mainly as silos; information was hard to compile and disseminate, which affected decision-making.

- Urgency to Meet Regulatory Requirements:

FBN needed to adhere to the regulatory requirements imposed by the Central Bank of Nigeria as well as the common business practices followed by Nigerian banks. Since no two banks work in exactly the same way, the bank-specific requirements were also important. The central bank's increasingly proactive role in regulating the industry to bring it up to speed with international trends meant that the bank had to remain agile in order to survive and come out a winner.

Need for Innovation and Faster Time to Market With sophistication of customer requirements and increased competition, the bank's critical requirement was to not only to meet the existing demands of the customer but also to stay agile and meet the changing requirements going forward.

One of the pillars of Finacle's value proposition to FBN was its new generation solution architecture, designed to help the bank build an agile business through innovative offerings to the market and a significantly superior speed of response to customer, competitive and regulatory requirements. The other was Finacle's proven track record of 100% successful implementations across the globe, which offered the bank the attractive proposition of minimized risk. FBN piloted on Finacle in six months and since then has rolled out the solution to over 170 branches, on time and within budget. The benefits of the solution include:

- Time-to-market Advantage:

FBN's unique requirements were catered to using Finacle's Extensibility toolkit, the infrastructure that enabled the bank to customize its specific requirements without touching the source code. This provided significant time-to-market advantage to the bank and enabled them to design and launch new product offerings quickly.

- 24/7 Operability:

Regular version upgrades over the years have provided increased and more sophisticated functionality to the bank as the relationship has progressed. The new generation flexible architecture of Finacle has ensured 24/7 operability, with close to 100% uptime, a feature of immense importance in a country not known for failsafe network connectivity.

- Scalability:

Finacle's technological superiority and functional richness were important factors but its proven ability to scale up to FBN's explosive growth plans was the clincher. Finacle successfully met FBN's expectations of the solution being able to "scale up and be the vehicle of growth to meet the emerging global challenges in the financial arena."

- Streamlined Operations:

The new generation architecture of Finacle - fully web-enabled, with powerful and unique capabilities such as Straight Through Processing (STP), workflow, scalability and true 24/7 banking across multiple delivery channels has enabled the Bank to streamline its operations.

B. *United Bank for Africa (UBA)*

United Bank for Africa PLC (UBA) is the product of a merger of two of Nigeria's top five banks, UBA and Standard Trust Bank Plc (STB). Today, consolidated UBA is largest financial services institution in sub- Saharan Africa (excluding South Africa) with a balance sheet size in excess of 400 billion naira (approx. US$ 3 bn), and over two million active customer accounts. With over 400 retail distribution outlets across Nigeria, UBA also has a presence in New York, Grand Cayman Island and aspires to expand within Sub-Saharan Africa.

UBA is the first successful merger transaction in the history of the Nigerian banking sector and was born out of a desire to lead the sector to a new era of global relevance by championing the creation of the Nigerian consumer finance market and leading a private/public sector partnership aimed at accelerating the economic development of Nigeria.

The Nigeria banking industry is going through so tremendous flux. The Central Bank's mandate of a minimum N25 billion capitalization by December 2005 resulted in the Nigerian market witnessing consolidation activity on a large scale. Though the UBA-STB merger was consummated during the ongoing consolidation era, it was a strategic move by the bank to become a large regional player, with an increased reach and synergies in terms of larger customer base and complementary product portfolio.

In its determination to continue to leverage on a robust IT infrastructure designed to achieve excellent service delivery to its teeming clientele, UBA opted for Finacle universal banking solution, comprising core banking, corporate e-banking, alerts, CRM and treasury solutions from Infosys in October 2005. The relationship between Finacle and UBA dates back to 5 years ago when STB changed from its existing Globus system to Finacle. Finacle core banking solution helped power STB's rapid growth at the turn of the millennium and its emergence as one of Nigeria's leading new generation banks. In addition STB is credited to have spearheaded the deployment of ATM's and internet banking in the Nigeria market riding on Finacle.

To power ahead in the dynamic post-consolidation banking landscape of Nigeria, UBA requires a technology partnership that transcended a typical customer-vendor relationship. From the STB experience, what emerged was the impeccable delivery track record of the Infosys implementation team. Recall that the bank (STB) completed a 65-branch roll out in quick time, less than 6 months, and a far cry from the 18-24 month implementation cycles prevalent in the country then. UBA also needs to capitalize on an integrated channel strategy that incorporated e-banking and CRM, among others.

VI. Discussions On Findings And Conclusion

The results of the research indicate that investment on ICT system and infrastructures has become a key element in productivity and growth in the banking industry. Increased investment in ICT-Capital has accelerated growth in industry. Also, ICT facilitates the absorption of high and medium skilled labor. This has a positive effect on the labor output of the banking industry. The case studies indicate that ICT also enables banks offer a broad variety of services to customers, coordinate branch activities, meet up with changes in government regulations and policies as well as adjust to market demands and competition.

However, only 25.4 million of Nigerians, representing 30% of the adult population have bank accounts. This leaves about 70 % of the adult populations unbanked. Thus to justify investment on ICT, banks need to draw out explicit ways to reach the unbanked. One way of achieving this objective is to increase the geographical outreach of the financial system through the use of non-bank agents; a method that will involve investment in innovative ICT products and services.

In sum, the business environment is becoming ever competitive and dynamic, invariably then, banks require solutions that can scale up to their growth plans and provide them the much-needed agility to create a clear differentiation in the market.

Thus, banks need to employ ICT in such a way that meets the desired qualities of flexibility and scalability, providing them with a competitive advantage to stay ahead and provide new and improved products and services to delight their customers.

It must however be noted that ICT investment does not lead to productivity growth at firm-level by itself. It depends on how the technology is actually used in business processes, i.e. on a company's ability to innovate its work processes and business routines with support of ICT. Thus, only if ICT investment is combined with complementary investment in working practices, human capital, and firm restructuring will it have an impact on performance.

The finding of this study indicates that basic ICT infrastructures such as computer and peripherals, local area networks, and ATMs are crucial to the operations of banks. However, the case studies indicate that to meet the ever increasing sophistication of customers, new government policies and stay competive in a fast changing economy, a scalable, flexible and robust ICT solution is essential.

References

[1] Efraim Turban, Dorothy Leidner, Ephraim McLean, James Wetherbe,"Information Technology for Management: Transforming Organizations in the Digital Economy", 3rd edit., John Wiley & Sons, Inc., pp.10-15. ISBN 978-0-471-78712-9.

[2] Akinlolu Agboola,"Information And communication Technology (Ict) In Banking Operations In Nigeria –An Evaluation Of Recent Experiences", African Journal of Public administration and Management Vol XVIII, No. 1 January 2007.

[3] Osabuohien, Evans S.C, " Ict And Nigerian Banks Reforms: Analysis Of Anticipated Impacts In Selected Banks " Global Journal of Business Research, Vol.2, No.2, 2008.

[4] Rambøll Management, "ICT and e-Business impact in the Banking Industry", A Sectoral e-business Watch study Report Version 4.0 September 2008, available at http://ec.europa.eu/enterprise.

[5] Wikepedia Encyclopedia, "List of banks in Nigeria ", retrieved on march 20th, 2012 from http://en.wikipedia.org/wiki/List_of_banks_in_Nigeria

[6] Central bank of Nigeria, "List of Financial institutions" retrieved on March 20th, 2012 from http://www.cenbank.org/Supervision/Inst-DFI.asp

[7] The Senate Federal Republic Of Nigeria, "A Bill For An Act To Establish The Asset Management Corporation Of Nigeria For The Purpose Of efficiently Resolving The Non-Performing Loan Assets Of Banks In Nigeria And For Related Matters", Asset Management Corporation Of Nigeria Bill, 2010. Retrived on March 20th, 2012 from: http://www.proshareng.com/admin/upload/reports/2705.pdf

[8] Arend, R. J, "Emergence of entrepreneurs following exogenous technological change", Strategic Management Journal, Vol. 20, pp. 31-47, 1999

[9] Clayton M.Christensen, "The Innovator's Dilemma: when new technologies cause great firms to fail" Harvard Business School Press Boston, 1997

[10] Peter Yannopoulos," Defensive and Offensive Strategies for Market Success ", international Journal of Business and Social Science Vol. 2 No. 13 [Special Issue - July 2011].

[11] Jovanovic, B. and Rousseau, P. L. "General Purpose Technologies", Handbook of Economic Growth edited by Philippe Aghion & Steven Durlauf Edition 1, Vol. 1, No.1, Elsevier.

[12] Brynjolfsson E., Malone T. W., Gurbaxani V., and Kambil A. "Does Information Technology Lead to Smaller Firms?", Management Science, Vol. 40, No. 12, pp. 1628-1644, 1994.

[13] Infosys Finnacle, "First Bank of Nigeria Scaling Up for Explosive Growth ", case studies, retrieved on March 10th, 2012, from http://www.infosys.com/finacle/customers/case-studies.

[14] Infosys Finnacle, "United Bank for Africa Marching Towards Leadership ", case studies, retrieved on March 10th, 2012, from http://www.infosys.com/finacle/customers/case-studies.

Security Analysis of Image Cryptosystem Using Stream Cipher Algorithm with Nonlinear Filtering Function

Belmeguenaï Aïssa
Laboratoire de Recherche en
Electronique de Skikda
Université 20 Août 1955- Skikda
BP 26 Route d'El-hadaeik
Skikda, Algeria

Derouiche Nadir
Laboratoire de Recherche en
Electronique de Skikda
Université 20 Août 1955- Skikda
BP 26 Route d'El-hadaeik
Skikda, Algeria

Mansouri Khaled
Département d'Electronique
Université Badji Mokhtar
Annaba, Algeria

Abstract— **In this work a new algorithm for encryption image is introduced. This algorithm makes it possible to cipher and decipher images by guaranteeing a maximum security. The algorithm introduced is based on stream cipher with nonlinear filtering function. The Boolean function used in this algorithm is resilient function satisfying all the cryptographic criteria necessary carrying out the best possible compromises. In order to evaluate performance, the proposed algorithm was measured through a series of tests. Experimental results illustrate that the scheme is highly key sensitive, highly resistance to the noises and shows a good resistance against brute-force, Berlekamp-Massey Attack and algebraic attack.**

Keywords- cipherImage; cryptosystem; key-stream; nonlinear filtering function; stream cipher.

I. Introduction

In this paper, we are interested in the security of the data images, which are regarded as particular data because of their sizes and their information which is two-dimensional and redundant natures. These characteristics of the data make the classical cryptographic algorithms such as DES, RSA, and ... are inefficient for image encryption due to image inherent features, especially high volume image data. Many researchers proposed different image encryption schemes to overcome image encryption problems [1], [2], [3], [4]. In this work, we present a new algorithm for encryption and decryption images by using a stream cipher algorithm with filtering the linear feedback shift registers (LFSRs). The main advantages of such systems are their extreme speed and the change of the key of encryption for each symbol of the plaintext. In term of application, it is still the type of encryption preferentially and quasi-exclusively used in the industrial world (in particular in telecommunications and governmental). It allows implementations in hardware much easier, economic (less complexity). These algorithms are thus used in a privileged way in the case of communications likely to be strongly disturbed because they have the advantage of not propagating the errors [5]. This type of encryption is much faster than block ciphers.

The Boolean function used in this scheme is resilient function satisfying all the criteria cryptographic necessary to carry out a maximum security and can resist to certain attacks [6], [7], [8], [9].

II. Non Linear Filtering Function

This system was proposed by Siegenthaler [10] to increase the linear complexity of the binary sequence produced by linear feedback shift register (LFSR). A single register (LFSR) is used, length L, producing a binary sequence in maximum period. Certain stages of this register (LFSR) are combined by a nonlinear function g .

Such function is called filtering function. The sequence produced by the function which will constitute the key-stream, combined with the clear text. We refer to [11], [12] for further details. The linear complexity of the key-stream is at most $\lambda(s)=\sum_{i=1}^{d}\binom{L}{i}$, where d is the degree algebraic of g .

A. Linear Feedback Register

Linear feedback shift register produce a sequence $s=s_0,s_1,...,$ satisfying the linear recurrence relation $s_n=\sum_{i=1}^{L}c_i s_{n-i}$, $n\geq L$ where L is the length of the LFSR, $c_i\in F_2$ for $i=1,...,L$ and $s_i\in F_q$, $i\geq 0$.

The L stages, $S_n=(s_n,...,s_{n+L-1})$, is called a state of the shift register and we note $S_n=(s_n)_{n=0}^{\infty}$ the state sequence. We define the feedback polynomial to be $p(X)=1+c_1X+c_2X^2+...+c_LX^L$.

The first output symbols $s_0,s_1,...,s_{L-1}$, are initially loaded into the LFSR, these symbols are called the initial state. This is also the secret key of the LFSR.

The sequences $S=S_0,S_1,...$ produced by linear feedback register have many interesting properties such as a long periodicity. If the feedback polynomial p is primitive the period is $2^L - 1$.

B. Non Linear Boolean Function

Nonlinear Boolean function purpose in key-stream generators is to hide the linearity introduced by the LFSRs. A Boolean function is function $g : F_2^n \rightarrow F_2$.

The function g can be represented uniquely by a multivariate polynomial over F_2 of the form:

$$g(x_1,...,x_n) = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{1 \le i \prec j \le n} a_{ij} x_i x_j + ... + a_{12..n} x_1 x_2 ... x_n .$$

Where the coefficients a_0, a_i, a_{ij} ,.., $a_{12..n}$ belong to F_2. The degree of this polynomial is called the algebraic degree or simply degree of g , and it is denoted by $\deg(g)$. The functions of degrees at most one are called affine functions.

III. Algorithm Description

The fundamental objective of our contribution is to propose a cryptosystem images which allows two people, called traditionally *Alice* and *Bob* (for example), to transfer from the images through a not very sure channel so that a third nobody, pirate can't understand what is exchanged. It is supposed that *Alice* wishes to send in a way made safe by network a plain-image $imag$ of $n \times m$ pixels with *Bob*.

Initially *Alice* transforms the plain-image into binary flows of bits which one calls flow of bits of the plain-image. Then, starting from a secret key k , *Alice* generates the key-stream Y same size as the flow of bits of the plain-image for this session (see algorithm B). Lastly, *Alice* calculates the binary flow of the cipher-image and sends it to *Bob* as shown in the figure 1. *Alice* and *Bob* must exchange the secret key k as a preliminary. *Bob* then receives the binary flow of the cipher-image C , and of dimensioned sound, will use the secret key k to generate the key-stream Y , then, he calculates the binary flow of the deciphered image X . *Bob* put the binary flow of the deciphered image X in the form of an image of $n \times m$ pixels and stores it in $imgdech$. *Bob* can then visualize $imgdech$.

If *Alice* wishes to send a new image to *Bob*, he will use a new secret key k_1 for this new session.

A. Encryption and Decryption Image Algorithm

Encryption

Alice ciphers the plain-image $imag$ while passing by the following stages:

```
1.  To read the plain-image $imag$ of $n \times m$ pixels;
2.  To transform the plain-image into binary values and to store them in $X$ ;
3.  $N \leftarrow$ the size of $X$ ;
4.  for $i = 1$ to $N$ to make ;
5.      To generate the key-stream $Y(i)$ by using the algorithm B ;
6.  End to make ;
7.  for $i = 1$ to $N$ to make
8.      $C(i) = xor(X(i), Y(i))$ ;
9.  End to make ;
10. The binary flow of the cipher-image $C$ is sent.
```

Decryption

Bob deciphers the binary flow of the cipher-image C while passing by the following stages:

```
1.  $N \leftarrow$ the size of $C$ ;
2.  for $i = 1$ to $N$ to make ;
3.      To generate the key-stream $Y(i)$ by using the algorithm B ;
4.  End to make ;
5.  for $i = 1$ to $N$ to make;
6.      $Z(i) = xor(C(i), Y(i))$ ;
7.  End to make ;
8.  To put the binary flow of the deciphered image $Z$ in the form of an image of $n \times m$ pixels and to store it in $imgdech$ ;
9.  To post the deciphered image $imgdech$ .
```

B. Key-Stream Calculation Algorithm

Inputs:

- $imag$: *plain-image;*
- $s_0, s_1, ..., s_{L-1}$ *are initially loaded into the LFSR;*
- g : *filtering function with a 13 variables.*

Results:

- s : *binary sequence produced by LFSR ;*
- Y : *Key-stream produced by* g .

Treatment:

```
1.  To read $N$ , the size of $X$ ;
2.  To introduce the secret key, the value of initialization of LFSR $s_0, s_1, ..., s_{L-1}$ ;
3.  for $i = 1$ to $N + L - 1$ to make;
4.      To generate the binary sequence $s(i)$ produced by LFSR ;
5.  End to make ;
6.  for $i = 1$ to $N$ to make;
7.      To generate the key-stream $Y(i)$ produced by function $g$ ;
8.  End to make.
```

IV. The Proposed LFSR And Filtring Function

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

The realization a stream cipher system which is as resistant as possible to the known attacks requires having an important mathematical tool which makes it possible to generate robust and unforeseeable key-stream on the formal level but also in the field of the implementation.

We considered the linear feedback shift registers of length 521bits to produce a binary sequence. The feedback polynomial of LFSR is chosen to be the primitive polynomial $p(x)=1+x^{48}+x^{521}$and the initial state of LFSR is never allowed to be the all zero state. It follows that LFSR produces a maximum-length sequence of period$T=2^{521}-1$.

The filtering function g that we used here is drawn from [13]. This function must be a high algebraic degree, balancedness, good correlations immunity, high non linearity and preferably to have good algebraic immunity to resist certain attacks.

Let $G^0(x_1,...,x_9,x_{10})$ be a function on F_2^{10} proposed for standard LILI-128 (called function f_d [14]) is 3-resilient of algebraic degree 6 and nonlinearity $NG^0=480$ with algebraic immunity 4.

Let$G^1(x_1,...,x_9,x_{11},x_{12})=G^0(x_1,...,x_9,x_{11}\oplus x_{12})$.

Let $F(x_1,...x_{12})=x_{12}\oplus x_{11}+G^0(x_1,...x_{10})$ and

$H(x_1,...x_{12})=x_{12}\oplus x_{10}+G^0(x_1,...x_9,x_{11}\oplus x_{12})$.

We construct a function g in 13-variables in the following way, $g(x_1,x_2,...,x_{13})=$

$(1\oplus x_{13})F(x_1,...,x_{12})\oplus x_{13}H(x_1,...x_{12})$ is 5-resilient function, of algebraic degree 7 and nonlinearity $Ng=2^{12}-2^7$ with algebraic immunity 6. This function is optimal for the compromise between the degree and the order of resiliency, we have $7+5\leq 13-1$. This function satisfies all the cryptographic criteria necessary carrying out the best possible compromises.

V. Simulation And Results

Simulation was carried out using MATLAB V 7.5. The proposed crypto-data hiding methodology was tested in different images. However, we present the results for the four bringing images, illustrated figures. 2.a, 3.a, 4.a and 5.a. They were ciphered with the same key of size 521-bit.

We first, we applied our cryptosystem to different images, we have the following results: From the original images illustrated by the figures 2.a, 3.a, 4.a and 5.a, we applied our Encryption algorithm with a secret key 521 bits in order to obtain the cipher-images illustrated by the figures 2.b, 3.b, 4.b and 5.b. We notice that initial information is not any more visible. From the cipher-images illustrated by the figures 2.b, 3.b, 4.b and 5.b, we apply the algorithm of decryption algorithm (the rebuilding of the original images) with the same key 521 bits in order to obtain the deciphered images illustrated in figures 2.c, 3.c, 4.c and 5.c. Difference between plain images and its corresponding decrypted images shown in figures 2, 3, 4 and 5, and their histograms are shown in figure 6 are prove that, there is no loss of information, the difference is always 0.

VI. Security Analysis

A good encryption procedure should be robust against all kinds of cryptanalytic, brute-force (exhaustive research) and principal attacks (Berlekamp-Massey Attack, algebraic attack). In this section, the performance of the proposed image cryptosystem is analyzed in detail. We discuss the security analysis of the proposed image encryption scheme including some important ones like key sensitivity analysis, key space analysis, statistical attacks etc. to prove the proposed cryptosystem is secure against the most common attacks.

A. Key Space Analysis

For secure image encryption, the key space should be large enough to make the exhaustive research attack infeasible. Since the algorithm has a 521 bits key, the intruder needs 2^{521} tests by exhaustive research. An image cipher with such as a long key space is sufficient for reliable practical use.

B. Berlekamp-Massey Attack

For a filtering function of degree$d$, the linear complexity $\lambda(s)$ of the resulting key stream is upper bounded by $\sum_{i=1}^{d}\binom{L}{i}$. Moreover, it is very likely that the $\lambda(s)$ of the key stream $(Y_i)_{i\geq 0}$ is lower bounded by $\binom{L}{d}$ and that its period remains equal to 2^L-1. The Berlekamp-Massey attack [15] requires $2\lambda(s)$ data and has a complexity of $\lambda(s)^2$. Using the parameters L = 521; d = 7, linear complexity $\lambda(s)$ is between $2.0125e^{15}$and$1.9854e^{15}$, it is sufficiently large. This complexity completely excludes to use the Berlekamp-Massey attack.

C. Algebraic Attack

The complexity $C(L,d)$ of the algebraic attack on the stream cipher system with a key of size L bits and equations of d degree is given by $C(L,d)=\left(\sum_{i=0}^{d}\binom{L}{i}^{w}\right)=L^{w.d}$, where w corresponds to the coefficient of the method of the solution most effective by the linear system and d is equal to algebraic immunity of the filtering function. We employ here the expression of Strassen [16] which is $w = \log_2(7) \approx 2.807$.

In our cryptosystem the secret key is 521 bits and the algebraic immunity of the filtering function is equal to 6. This leads to algebraic attack with a complexity which is $5.7145e^{45}$, which is sufficiently large. It is not easy to make a linear approximation of the filtering function within the framework of algebraic attack.

D. Noise Analysis

We also tested the resistance our cryptosystem to the noise by adding to the cipher-images a noise. From the cipher-images illustrated in the figures 2.b, 3.b, 4.b and 5.b we added a noise of the same size of plain-images. The results are given in the figure 2.d, 3.d, 4.d and 5.d. From the images 2.d, 3.d, 4.d and 5.d, we apply the decryption algorithm presented in section A; we have the results illustrated in figure 2.f, 3.f, 4.f and 5.f. The noise added to ciphers-images 2.b, 3.b is a matrix containing pseudo-random values drawn from a uniform distribution on the unit interval, generates with function “rand”.

The noise added to ciphers-images 4.b and 5.b is a matrix containing pseudo-random values drawn from a normal distribution with mean zero and standard deviation one, generates with function “randn”. In two cases examined, we can note that the deciphered images presented in figures 2.f, 3.f, 4.f and 5.f are identical to the original images (see 2.a, 3.a, 4.a and 5.a), there is no difference pixel with pixel has indeed between the deciphered images and plain-images because of reversibility of our technique of encryption. Figures 2.e, 3.e, 4.e and 5.a are representing difference image between cipher-images and cipher-images with additive noise.

E. Sensitivity Analysis

Thus, we tested our cryptosystem to the sensibility to the keys, for example, we cipher the images 2.a, 3.a, 4.a and 5.a with the secret key $K_1 = 521$ bits and, we decipher it with different key; $K_2 = 521$ bits. The result is given by figure 7.

F. Correlation Coefficient Analysis

Table 1 gives the correlation coefficient results. In table 1, we denoted respectively by Cor_1, Cor_2, Cor_3, and Cor_4 correlation coefficient between plain-images and encrypted images, correlation coefficient between plain-images and their decrypted images, correlation coefficient between encrypted images and decrypted images with different key; K_2, and correlation coefficient between plain-images and decrypted images with different key; K_2. It is observed that the correlation coefficient is a small correlation between plain-images and encrypted image, encrypted images and decrypted images with different key; K_2, and plain-images and decrypted images with different key; K_2.

G. Entropy Analysis

Table 2 gives entropy results. In table 2, we denoted respectively by E_1, E_2, E_3, and E_4 entropy values: of plain-images, encryptions images, decrypted images and decrypted images with different key; K_2. The entropy values of encryptions images, decrypted images with different key; K_2 obtained are very close to the theoretical value of 8. This means that information leakage in the encryption process is negligible and the encryption system is secure upon the entropy attack.

H. Histogramm Analysis

In the experiments, the original images and its corresponding encrypted images are shown in figure 2, 3, 4 and 5, and their histograms are shown in figure 8. It is clear that the histogram of the encrypted image is nearly uniformly distributed, and significantly different from the respective histograms of the original image. So, the encrypted image does not provide any clue to employ any statistical attack on the proposed encryption of an image procedure, which makes statistical attacks difficult.

These properties tell that the proposed image encryption scheme has high security against statistical attacks. In the original image (i.e. plain image), some gray-scale values in the range [0, 255] are still not existed, but every gray-scale values in the range [0, 255] are existed and uniformly distributed in the encrypted image. Some gray-scale values are still not existed in the encrypted image although the existed gray-scale values are uniformly distributed. Different images have been tested by the proposed image encryption procedure.

VII. CONCLUSION

In this Work, a new algorithm based encryption scheme for image data was introduced; simulations were carried out for different images. The visual test indicates that the encrypted image was very different and no visual information can be deduced about the original image for all images. In addition, this method is very simple to implement, the encryption and decryption of an image.

Here the security aspects like key space, Berlekamp-Massey attack, algebraic attack, noise analysis, statistical attacks and sensitivity with respect to key, are discussed with examples. It is seen that the present cryptosystem is secure against the statistical attacks, brute force attack, Berlekamp-Massey attack, algebraic attack and to resists the additive noises.

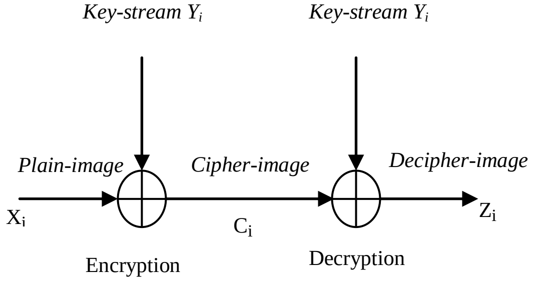


Figure 1. Principal encryption and decryption

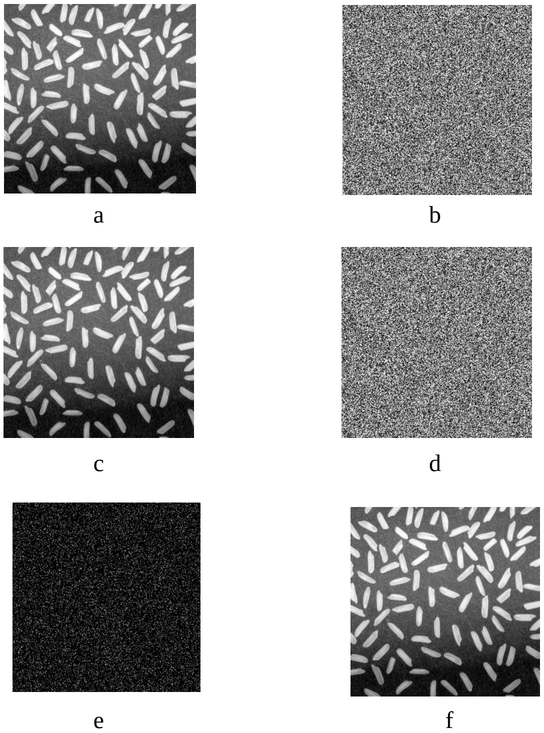


Figure 2. (a) Plain-image, (b) Cipher-image, c) Decipher image, d) Cipher-image with noise added, e) Difference image between image (b) and image (d), f) Decipher image (d).

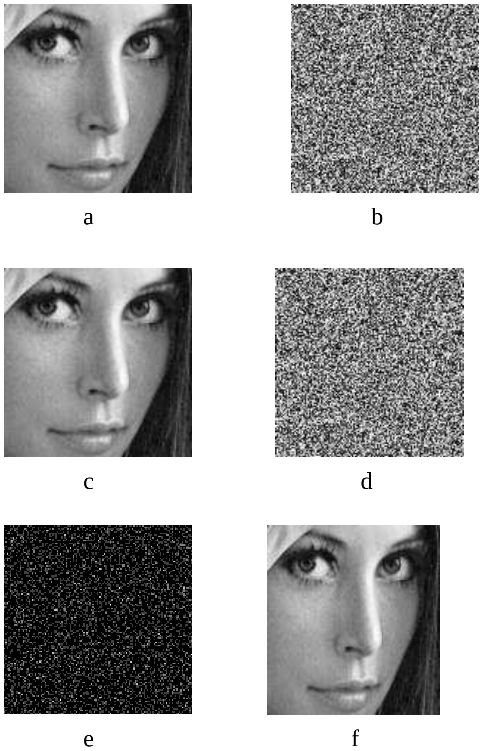


Figure 3. (a) Plain-image, (b) Cipher-image, c) Decipher image, d) Cipher-image with noise added, e) Difference image between image (b) and image (d), f) Decipher image (d).

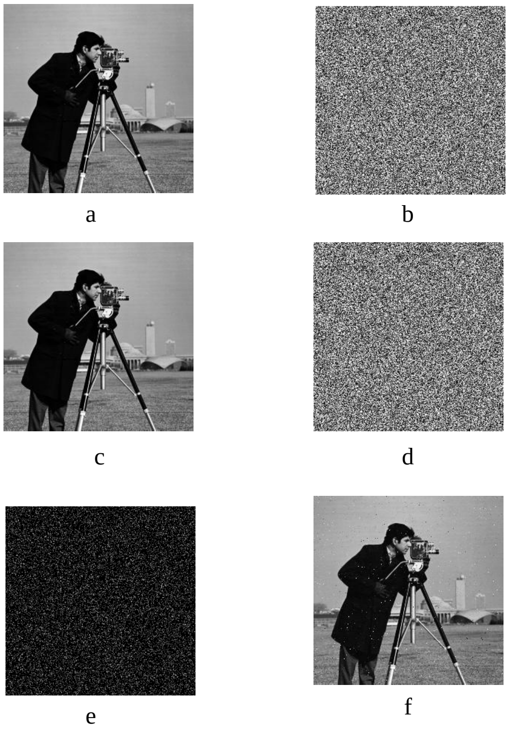


Figure 4. (a) Plain-image, (b) Cipher-image, c) Decipher image, d) Cipher-image with noise added, e) Difference image between image (b) and image (d), f) Decipher image (d).

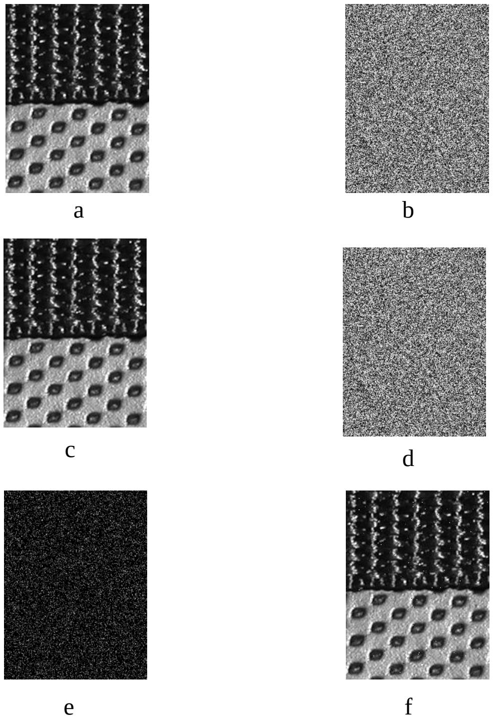


Figure 5. (a) Plain-image, (b) Cipher-image, c) Decipher image, d) Cipher-image with noise added, e) Difference image between image (b) and image (d), f) Decipher image (d).

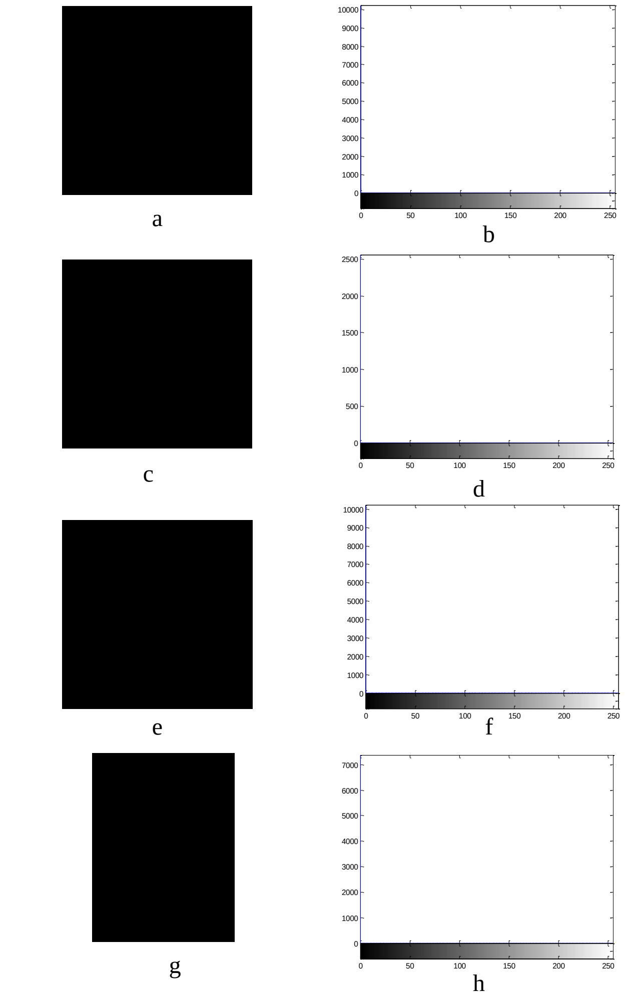
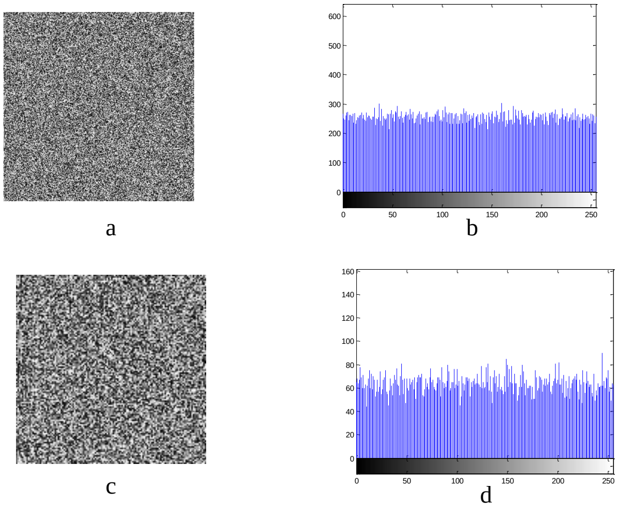


Figure 6. Frame (a), (c), (e) and (g) respectively show the difference between original images shown in figures 2.a, 3.a, 4.a and 5.a, and their decrypted image shown in fig 2.c, 3.c, 4.c and 5.c. Frame (b), (d), (f) and (h) respectively show their histogram.



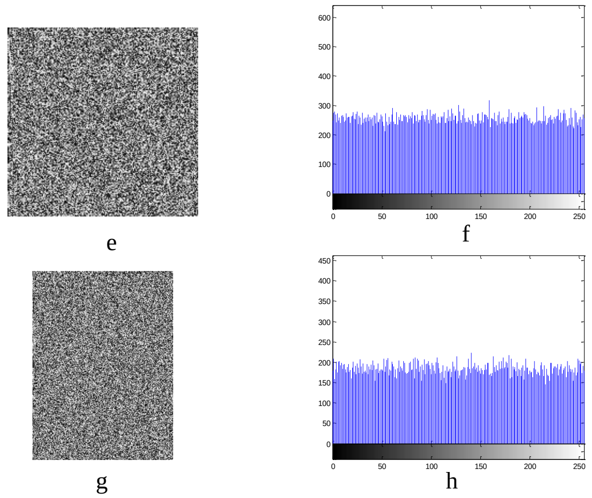


Figure 7. Sensitivity analysis: Frame (a), (c), (e) and (g) respectively, show decrypted image with wrong key (K_2) of the encryption images shown in figures 2.b, 3.b, 4.b and 5.b. Frame (b), (d), (f) and (h) respectively, show histogram of images ((a), (c), (e) and (g).

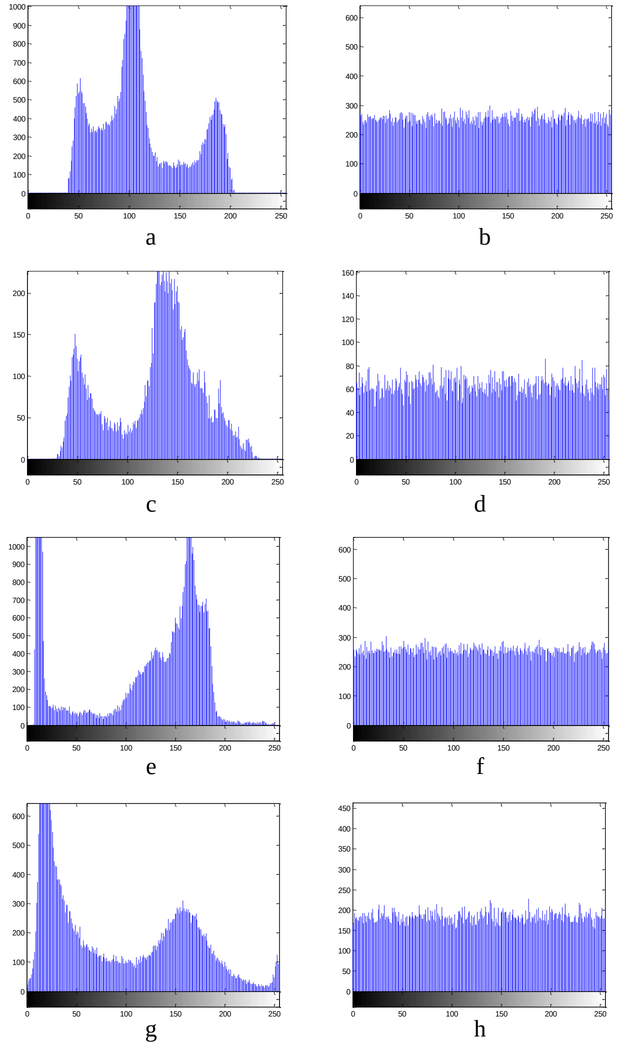


Figure 8. Histogram analysis: Frame (a), (c), (e) and (g) respectively, show the histogram of the plain images shown in figures 2.a, 3.a, 4.a and 5.a. Frame

(b), (d), (f) and (h) show the histogram of the decrypted image shown in figures 2.c, 3.c, 4.c and 5.c.

TABLE I. CORRELATION COEFFICIENTS

Cases	E_1	E_2	E_3	E_4
Image 2.a	7,0115	7,9973	7,0115	7,9973
Image 3.a	7,2631	7,9904	7,2631	7,9894
Image 4.a	7,0097	7,9977	7,0097	7,9972
Image 5.a	7,4864	7,9962	7,4864	7,9958

TABLE II. IMAGES ENTROPY

Cases	COR_1	COR_2	COR_3	COR_4
Image 2.a	0,0975	1	-0,0055	-0,0032
Image 3.a	-0,0050	1	-0,0022	-0,0018
Image 4.a	-0,0068	1	-0,0046	0,0024
Image 5.a	-0,0066	1	-0,0022	-0,0030

REFERENCES

[1] M. Sharma and M. K. Kowar, "Image encryption techniques using chaotic schemes: a review", International Journal of Engineering Science and Technology, vol. II, no. 6, 2010, pp. 2359–2363.

[2] A. Jolfaei and A. Mirghadri, "An applied imagery encryption algorithm based on shuffling and baker's map," Proceedings of the 2010 International Conference on Artificial Intelligence and Pattern Recognition (AIPR-10), Florida, USA, 2010, pp. 279–285.

[3] A. Jolfaei and A. Mirghadri, "A novel image encryption scheme using pixel shuffler and A5/1," Proceedings of The 2010 International Conference on Artificial Intelligence and Computational Intelligence (AICI10), Sanya, China, 2010.

[4] L. Xiangdong, Z. Junxing, Z. Jinhai and H. Xiqin, "Image scrambling algorithm based on chaos theory and sorting transformation," IJCSNS International Journal of Computer Science and Network Security, vol. 8, no. 1, 2008, pp. 64–68.

[5] C. Carlet, "On the cost weight divisibility and non linearity of resilient and correlation immune functions", Proceeding of SETA'01 (Sequences and their applications 2001), Discrete Mathematics, Theoretical Computer Science, Springer p 131-144, 2001.

[6] T. Siegenthaler, "Decrypting a class of stream ciphers using cipher text only", *IEEE Transactions on Computers*, C-34(1):81–85, January 1985.

[7] C. Ding, G. Xiao, and W. Shan, "The stability theory of stream ciphers", Lecture Notes in Computer Science, Number 561, Springer Verlag, August 1991.

[8] N. Courtois and W. Meier, "Algebraic attacks on stream ciphers with linear feedback", Advances in cryptology– EUROCRYPT 2003, Lecture Notes in Computer Science 2656, pp. 346-359, Springer,2002.

[9] N. Courtois, "Fast algebraic attacks on stream ciphers with linear feedback", advances in cryptology–CRYPTO 2003, Lecture Notes in Computer Science 2729, pp. 177-194, Springer, 2003.

[10] T. Siegenthaler, "Cryptanalysis representation of nonlinearly filtered ML-sequences*"*, In : Advances in cryptology- EUROCRYPT' 85, Lectures Notes in Computer science 219,pp 103-110,Springer Verlag, 1986.

[11] P.van Oorschot A. Menezes and S. Vantome, "Handbook of applied cryptography", Available: htt.www.cacr.math.uwaterloo.ca/,1996.

[12] G. Ars, "Une application des bases de Gröbner en cryptographie", DEA de Renne I, 2001.

[13] E. Pasalic, S. Maitra, T. Johansson and P. Sarkar, "New constructions of resilient and correlation immune Boolean functions achieving upper bounds on nonlinearity*",* In Workshop on Coding and Cryptography - WCC 2001, Paris, January 8–12, 2001. Electronic Notes in Discrete Mathematics, volume 6, Elsevier Science, 2001.

[14] L. Simpson, E. Dawson, J. Golic, and W. Millan, "LILI-128 key-stream generator*"*, In Selected Areas in Cryptography, 7th Annual International Workshop, SAC2000, volume 2012 of Lecture Notes in Computer Science, pages 248–261. Springer-Verlag, Berlin, Heidelberg, New York, 2001.

[15] E.R Berlekamp. "Algebraic coding theory", *Mc Grow- Hill, New- York*, 1968.

[16] V. Strassen, "Gaussian elimination is not optimal", *Numerische Mathematik*, 13:354-356, 1969.

A Multi-Stage Optimization Model With Minimum Energy Consumption-Wireless Mesh Networks

S.Krishnakumar
Research Scholar, CSE Dept., SRM University
Chennai, India

Dr.R.Srinivasan
CSE Dept, SRM University
Chennai, India

Abstract—**Optimization models related with routing, bandwidth utilization and power consumption are developed in the wireless mesh computing environment using the operations research techniques such as maximal flow model, transshipment model and minimax optimizing algorithm. The Path creation algorithm is used to find the multiple paths from source to destination.A multi-stage optimization model is developed by combining the multi-path optimization model, optimization model in capacity utilization and energy optimization model and minimax optimizing algorithm. The input to the multi-stage optimization model is a network with many source and destination. The optimal solution obtained from this model is a minimum energy consuming path from source to destination along with the maximum data rate over each link. The performance is evaluated by comparing the data rate values of superimposed algorithm and minimax optimizing algorithm. The main advantage of this model is the reduction of traffic congestion in the networ**k.

Keywords-optimization; breakthrough; transportation; aximization; superimposed; transshipment.

I. Introduction

A. Formulation of Linear programming problem (LPP)

Let s be the source node and N be a set of neighbor nodes of source. Let x_{ij} be the rate of transmission of packets over the link (i,j) and c_{ij} be the capacity of the link (i,j). Then LPP form of maximal flow problem is

Maximize $\sum_{j \in N} x_{sj}$ It represents the amount of flow passing from the source to the sink.

Subject to the constraints

$\sum_{j} x_{ij} = \sum_{k} x_{jk}$ For all j (flow conservation condition)

(Total incoming flow = Total outgoing flow)

$x_{ij} \leq c_{ij}$ (capacity constraint)

$x_{ij} \geq 0$ (non-negativity restrictions)

Even though this LPP can be solved using simplex method, we are using a simple and efficient algorithm called maximal flow algorithm [20] to find the maximal flow.

1) Maximal Flow Algorithm

Step 1: For all links set the residual capacity equal to the initial capacity and label source node 1 with [∞, -]. Set i = 1 and go to step 2.

Step 2:Determine S_i as the set unlabeled nodes j that can be reached directly from mode i by arcs with positive residuals. If S_i is non-empty go to Step 3 otherwise go to Step 4

Step 3: Determine k in S_i such that $C_{ik} = \max \{C_{ij}\}$ where j belongs to S_i and C_{ij} represents capacity of the link (i,j). Set $A_k = C_{ik}$ and label node k with (A_k, i). If the sink node has been labeled (i.e., k = n) and a breakthrough path is found, go to step 5. Otherwise, set i=k, and go to step 2.

Step 4: (Backtracking)If i=1, no further breakthroughs are possible; go to step 6. Otherwise, let r be the node that has been labeled immediately before the current node i and remove i from the nodes that are adjacent to r. Set i = r, and go to step 2.

Step 5:(determination of residue network). Let Np = [1, k_1, k_2….., n) define the nodes of the path breakthrough path from source 1 to sink n. Then the maximum flow along the path is computed as

$Fp = \min \{A_1, A_{k1}, A_{k2} \ldots An\}$.

The residual capacity of each arc along the breakthrough path is decreased by F_p in the direction of the flow and increased by F_p in the reverse direction. Reinstate any nodes that were removed in step 4. Set i = 1, and return to step 2 to attempt a new breakthrough path.

Step 6: (**Solution**)

(a) Given that m breakthrough paths have been determined, compute the maximal flow in the network as $F = F_1 + F_2 + \ldots\ldots.. + F_m$

(b) The optimal flow over the link (i,j) is computed as follows:

Let a=initial capacity – final residue over the link (i,j) and b= initial capacity – final residue over the link (j,i)

If a>0, the optimal flow from i to j is a. Otherwise, if b>0 the optimal flow from j to i is b.In the next section we develop an optimization model which utilizes the capacity of the link effectively. The advantage of this model is the elimination of congestion problem in the network.

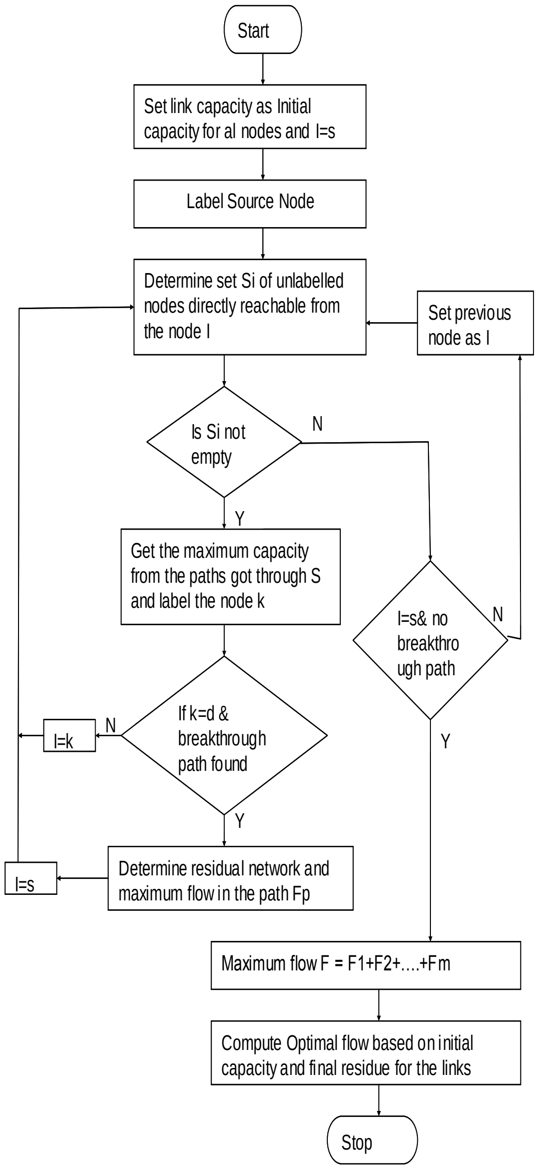

Start
Set link capacity as Initial capacity for al nodes and I=s
Label Source Node
Determine set Si of unlabelled nodes directly reachable from the node I
Set previous node as I
Is Si not empty
N
Y
Get the maximum capacity from the paths got through S and label the node k
I=s& no breakthro ugh path
N
If k=d & breakthrough path found
N
I=k
Y
Y
Determine residual network and maximum flow in the path Fp
I=s
Maximum flow F = F1+F2+....+Fm
Compute Optimal flow based on initial capacity and final residue for the links
Stop

Figure1: Maximum Flow Algorithm

II. Optimization Model For Capacity Utilization

Using Shannon's theorem, we can calculate capacity of a link from its bandwidth. We assume that the nodes have infinite energy to transmit any number of packets in order to estimate the maximum data rate over each link. Then we apply maximal flow algorithm to find the maximum rate of transmission of packets over each link and the maximal flow in the network.

A. Input

1. A network with source and destination
2. Capacity of each link in the network.

B. Procedure

Maximal flow algorithm

C. Output

1. Maximum number of packets that can be transmitted from source in one second.

2. Maximum rate of transmission of packets over each link.

Maximum flow algorithm for many sources and many destinations:

A maximal flow problem may have several sources and sinks. The objective is to find the maximum flow between the number of sources and destinations. We can reduce the problem of determining a maximal flow in a network with multiple sources and multiple sinks to an ordinary maximal flow problem [2, 5, and 7].

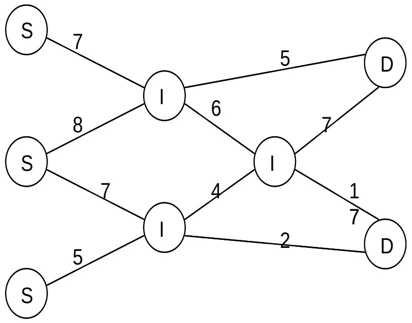


Figure 2: Many sources to many destinations

Firstly, we are converting this multiple sources and multiple sinks into only one source and one destination. For this, We are creating two nodes as Super source(S') and Super Destination(D'),then adding the edge(S',Si) with capacity C(S',Si)=MAX The MAX value is the maximum capacity of all the links or infinite capacity will be allocated as the capacity (link) of the super source and the super destination. Then we are connecting such that all the source nodes are get connected with the Super source and all the destination nodes are get connected with the Super destination [3, 17]. Now this situation is assumed as the data passed form single source and destination. Then we can implement the maximum flow algorithm.

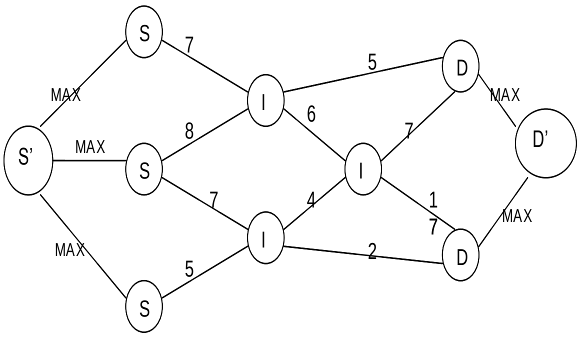


Figure2.1: Superimposed Sources and destinations

SUPERIMPOSED ALGORITHM:

Step 1: creating the links for the network randomly by the input .The links will be generated as N * N link matrix in which 1's and 0's will be generated where '1' denotes link existence and '0' non- existence of links.

Step 2: Randomly generating link capacity for the created links and storing in data

Step3: creating two nodes as Super source(S') and Super Destination (D'), then adding the edge(S', Si) with capacity C(S', Si) =MAX

Step4**:** Implement The Maximum Flow Algorithm.

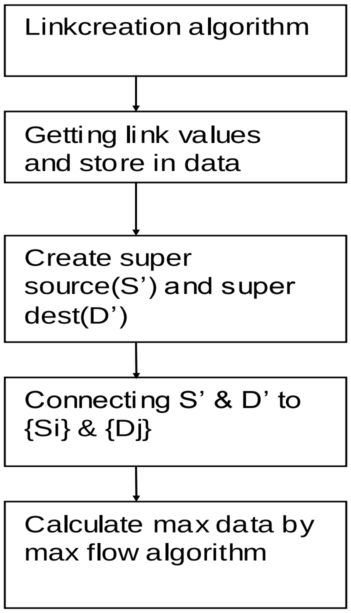

Linkcreation algorithm
Getting link values and store in data
Create super source(S') and super dest(D')
Connecting S' & D' to {Si} & {Dj}
Calculate max data by max flow algorithm

Figure 2.2: Finding break paths

III. Transportation Model

Let there be m source nodes and n destination nodes. Let S_i be a supply from node i and d_j be a demand from node j. Let c_{ij} be the energy required to transmit a packet from node i to node j and x_{ij} be the number of packets that can be transmitted from node i to node j. Let P_i be the power level of node i.

Objective of transportation problem is to minimize the total energy consumption given by

$$\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}$$

Subject to the constraints

$x_{ij} + x_{i2} + \ldots\ldots\ldots + x_{in} = S_i$; i=1, ……….,m (supply constraint)

$x_{ij} + x_{2j} + \ldots\ldots\ldots + x_{mj} = d_j$; j=1, ………,m (demand constraint)

$\sum_j c_{ij}\ x_{ij} \le P_i$ for all i (power constraint)

And

$x_{ij} \ge 0$ for all i and j

Procedure for Solving Transportation Problem:

A. MODI Method (Modified Distribution Method) [20]

Cell:Each cell represents a shipping route, which is an arc on the network and a decision variable in the Linear Programming formulation.

Step 1 Formulate the transportation table

Step 2 Construct the initial basic feasible solution by using any of the following methods

(i) NWC rule
(ii) LCM
(iii) VAM

To get the optimal solution (with smaller number of iterations) quickly, use VAM.

Step 3 Test the optimality

Make sure that there are m+n-1 non-zero allocations (Non-degenerate basic feasible solution). These allocations should be in independent positions.

(i) Determine

u_i i=1…m
v_j j=1…n
Such that for each occupied cell (r,s)
$c_{rs} = u_r + v_s$

This can be done by choosing arbitrarily one of the $u_i = 0$ or $v_j = 0$. For more convenience, choose the row with the most allocations. If i^{th} row has the most allocations (among rows) then take $u_i = 0$.

(ii) Calculate the cell evaluations (net evaluations)

$\Delta_{ij} = c_{ij} - (u_i + v_j\ )$ for all unoccupied cells (empty cells).

Note that $\Delta_{ij} = 0$ if (i,j) is an occupied cell.

(iii) If $\Delta_{ij} \ge 0$, then the present solution is optimal.

(iv) If at least one $\Delta_{ij} < 0_n$ for at least one cell then the present solution is not optimal.

Step 4 (Iteration towards optimal solution)

(i) Choose the cell which has the most negative Δ_{ij}value and mark * in it (The corresponding variable is entering non-basic variable)

(ii) Draw a closed path consisting of horizontal and vertical lines beginning and ending at * cell and having its corners at the allocated cells.

Mark + sign at * cell and + and – signs alternatively at other corner cells of the path. The cell with + signs are called donor cells, which has the least allocation in the leaving basic variable.

(iii) Add the value of leaving basic variable to the allocation for each recipient cell. Subtract this value from the allocation for each donor cell.

This gives improved feasible solution to step 3 for testing optimality.

Remark

In Step 4, (i) if the entering non-basic variable has a tie, then select a variable which has lower cost i.e., if Δ_{ij} and Δ_{rs} are most negative, then select (i,j) cell, if c_{ij} is small.

B. Unbalanced transportation problem

A transportation problem is said to be unbalanced if the total supply is not equal to total demand.

SOLUTION TO THE UNBALANCED TRANSPORTATION PROBLEM

Convert the unbalanced transportation problem into a balanced transportation problem by the following techniques.

(i) Total supply > Total demand (Surplus of supply)

Add a dummy destination node to distribute the surplus (excess) amount of supply and let zero be the cost of transportation to this dummy destination.

i.e., add a dummy column at end of the transportation table and take the excess amount of supply (Total supply – Total demand) as the demand at this destination. Take zero as the unit transportation cost for the cells in this column.

(ii) Total supply < total demand (Slackness or Supply)

Add a dummy source node to produce the slackness of supply in order to saturate excess amount of demand i.e., Add a dummy row at the end of the transportation table. Supply at the dummy source = Total demand – Total supply. Take zero as the unit transportation cost for the cells in this row.

C. Maximization Type Transportation Problem

A maximization type transportation problem can be converted into usual minimization type transportation problem by subtracting each of the costs from the highest cost given in the problem to obtain only the optimal solution. For calculating the total transportations cost, use the original cost given in the problem.

D. Transshipment Model

A transportation problem in which the supply may not be sent directly from sources to destinations, i.e., the supply may pass through one or more sources or destinations before reaching its actual destination, is called as transshipment problem.

The nodes of network with both input and output links act as both sources and destinations, and are referred to as transshipment nodes. The remaining nodes are either pure supply nodes or pure demand nodes. The transshipment model can be converted into a transportation model by computing the amount of supply and demand at different nodes as follows:

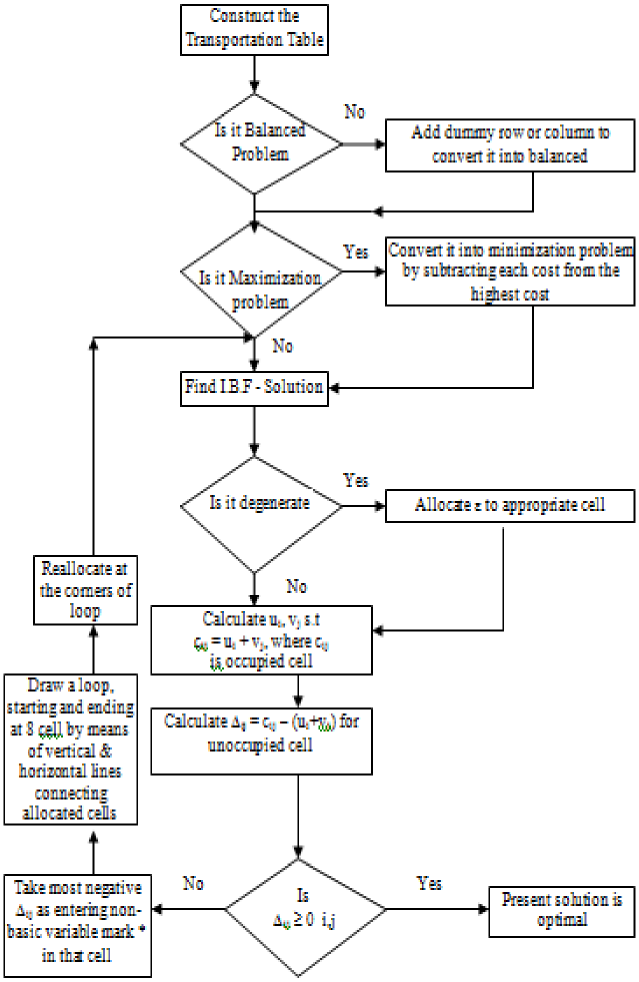

Construct the Transportation Table
Is it Balanced Problem
No
Add dummy row or column to convert it into balanced
Is it Maximization problem
Yes
Convert it into minimization problem by subtracting each cost from the highest cost
No
Find I.B.F - Solution
Is it degenerate
Yes
Allocate ε to appropriate cell
No
Reallocate at the corners of loop
Calculate u_i, v_j s.t $c_{ij} = u_i + v_j$, where c_{ij} is occupied cell
Draw a loop, starting and ending at 8 cell by means of vertical & horizontal lines connecting allocated cells
Calculate $\Delta_{ij} = c_{ij} - (u_i+v_j)$ for unoccupied cell
Take most negative Δ_{ij} as entering non-basic variable mark * in that cell
No
Is $\Delta_{ij} \geq 0$ i,j
Yes
Present solution is optimal

Figure: 3 Flowchart for solving a Transportation Model

Let P = total supply (or demand)

Supply at a pure supply node = Original supply

Supply at a transshipment node = Original supply + P

Demand at a pure demand node = Original Demand

Demand at a transshipment node = Original Demand + P

Assume that transportation cost (energy consumption) the same node is zero i.e., $c_{ii} = 0$ for all i.

E. Energy Optimization Model

In wireless mesh computing environment some of the nodes are sources (eg. sensors), some are sinks (eg. Pagers), some are both source and link (eg. Computers) and others are only junctions (eg. Routers and bridges).Since we consider a network with one source and one destination and rest of the nodes to forward packets, we apply transshipment model to develop an optimization model in power consumption.

Let e be the energy required by a host to transmit a message to another host who is d distance away. Then $e = rd^c$ where r and c are constants for the specific wireless mesh system. Hence energy consumption is proportional to the distance between nodes. Here distance between nodes is calculated by number of hops between the nodes.

Here supply at a node is the number of packets that can be transmitted by a node and demand is the number of packets that can be received by a node. Hence, supply and demand at a node are depending on the available battery energy in the node.

We assume that transshipment nodes (i.e., nodes forwarding packets) possess sufficient energy to forward packets. We can apply transshipment model to get minimum energy consuming path and the maximum number of data that can be transmitted.

1) *Input*
   a) *A network with source and destination*
   b) *Residual power level at each node*
   c) *Distance between nodes*
2) *Procedure*
   Transshipment model algorithm
3) *Output*
   a) *Minimum energy consuming path*
   b) *Maximum number of packets that can be transmitted*
   c) *Total energy consumption.*

IV. OPTIMIZATION

We consider a wireless network containing multiple sources and multiple destinations. In this chapter we discuss about the multistage optimization model which is a combination of the data rate and energy optimization that have been developed in the earlier sections.

The objective of this model to find the minimum energy consuming path from source to destination and the maximum data rate over each link in the minimum energy consuming path[2,3,7]

Transmission of messages is continued along minimum energy consuming path for a period of time T. The parameter T can be determined using the time taken by the nodes for recharging their battery and dynamically changing speed of topology of the network

A. *Minimax optimizing algorithm:*

The Minimax algorithm is the algorithm for integrating the maximization and minimization [2, 7 and 17]. In this algorithm, the maximization (maximal flow model) is compared with the minimization (transshipment algorithm) and the break paths are selected and calculating the maximum data rate with minimum energy consumption.

ALGORITHM:

Step 1: creating the links for the network randomly by the input .The links will be generated as N * N link matrix in which 1's and 0's will be generated where '1' denotes link existence and '0' non-existence of links.

Step 2: Randomly generating link capacity for the created links and storing in data and tempdata and getting the Sources {Si} and Destinations {Dj} from the input.

Step 3: Calculate the maximum Data rate through superimposed algorithm.

Step 4: Check whether the iterations are complete for source and destination {SDij}, then go to step9 or go to step 5.

Step 5: If there is any path, calculate the maximum data rate using maximal flow algorithm in tempdata or go to step 7.

Step 6: Send the path (links) to the transshipment algorithm and get the energy E {SDij} consumed in that path. Go to step 5.

Step 7: Select the Break paths with energy based on data rate.

Step 8: Update the value in data and update data value to temp data.

Step 9: Calculate the total data rate of each {SDij}.

Step 10: Compare the data rate values of superimposed algorithm and the Minimax Algorithm and plot it in graph.

Selecting Break paths:

Step 1: Tabulate the energy and the data rate of each path for each source and destination {SDij}.

Step 2: Calculate average energy as X, where n is the number of nodes.

Step 3: Convert the data rate value to Average value as Y and tabulate it.

Step 4: Arrange the rows according to data rate, in decreasing order.

Step 5: Select the first two paths.

V. RESULTS

Experiment Inputs-

Number of nodes	: 11
Number of sources and Destination {Si, Dj}	:3, 4
Source Nodes {si}	: 2, 3, 4
Destination Nodes {dj}	: 5, 6, 7,8.
Paths {Pij}	: 2→5, 2→6, 2→8, 3→6, 3→7, 4→5, 4→7

Input: no of nodes, {Si, Dj {si},{dj},{Pij}.

11, {3,4}, {2,3,4},{5,6,7,8},{1,1,0,1,0,1,1,0,1,0,1,0}

The multistage optimization model is implemented using a Language C++, in the Windows 2000 operating system.

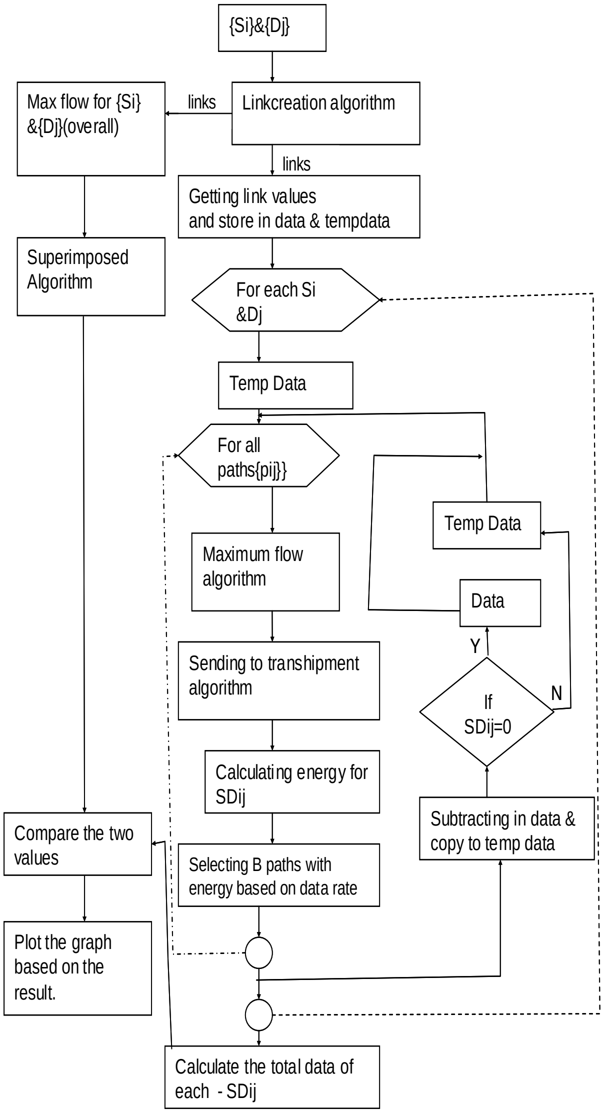

{Si}&{Dj}
Max flow for {Si} &{Dj}(overall)
links
Linkcreation algorithm
links
Getting link values and store in data & tempdata
Superimposed Algorithm
For each Si &Dj
Temp Data
For all paths{pij}}
Temp Data
Maximum flow algorithm
Data
Y
Sending to transhipment algorithm
If SDij=0
N
Calculating energy for SDij
Subtracting in data & copy to temp data
Compare the two values
Selecting B paths with energy based on data rate
Plot the graph based on the result.
Calculate the total data of each - SDij

Figure 4: Minimax optimizing algorithm

VI. Comparison:

Analysis of performance of the model by comparing the simulation results of this Minmax model with the simulation results of superimposed algorithm related with power consumption and Bandwidth utilization in the Wireless environment. Packet loss attribute is not included for this model, as it occurs in both the models

A. *1. Superimposed algorithm data rate value and Minimax optimizing algorithm data rate value with respect to number of nodes:*

In the figure 5, we are comparing the values of superimposed algorithm data rate value and Minimax optimizing algorithm data rate value with respect to number of nodes. For the increase in number of nodes, the data rate value of both the algorithm is increased.

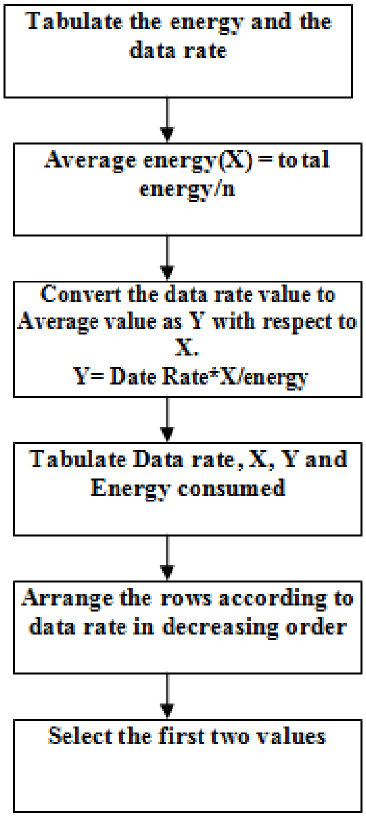

Tabulate the energy and the data rate
Average energy(X) = total energy/n
Convert the data rate value to Average value as Y with respect to X.
Y= Date Rate*X/energy
Tabulate Data rate, X, Y and Energy consumed
Arrange the rows according to data rate in decreasing order
Select the first two values

Figure 4.1: Selecting Break paths:

But the data rate value of Minimax optimizing algorithm is greater than the superimposed algorithm.

B. *Superimposed algorithm data rate value and Minimax optimizing algorithm data rate value with respect to number of sources and destinations:*

In the figure 5.1, we are comparing the values of superimposed algorithm data rate value and Minimax optimizing algorithm data rate value with respect to number of sources and destinations.

For the increase in number of sources and destinations, the data rate value of both the algorithm is increased. But the data rate value of Minimax optimizing algorithm is greater than the superimposed algorithm.

C. *Superimposed algorithm data rate value and Minimax optimizing algorithm data rate value with respect to number of paths:*

In the figure 5.2, we are comparing the values of superimposed algorithm data rate value and Minimax optimizing algorithm data rate value with respect to number of paths.

For the increase in number of paths, the data rate value of both the algorithm is increased. But the data rate value of Minimax optimizing algorithm is greater than the superimposed algorithm

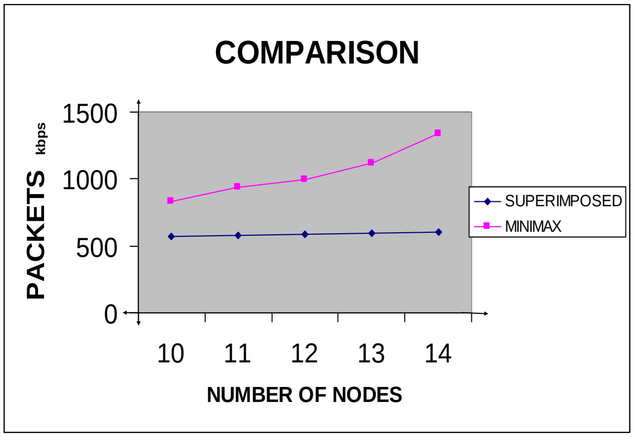


Figure 5: Data rate Vs Number of nodes

Table 1 Data rate Vs Number of nodes

Number of nodes	Superimposed algorithm Data rate (packets per second)	Minimax optimizing algorithm. Data rate (packets per second)
10	573	834
11	580	939
12	588	994
13	592	1114
14	603	1335

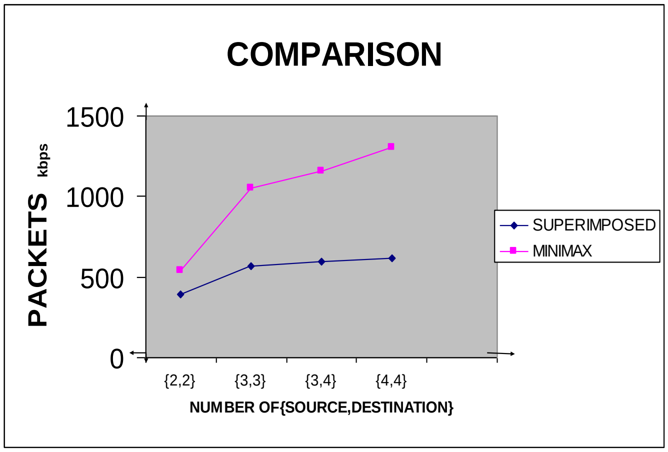


Figure 5.1 Data rate Vs Number of {source, destination}

Table 2 Data rate Vs Number of {source, destination}

Number of {SOURCE,DESTINATION}	Superimposed algorithm. Data rate (packets per second)	Minimax optimizing algorithm. Data rate (packets per second)
{2,2}	396	540
{3,3}	567	1048
{3,4}	597	1160
{4,4}	620	1302

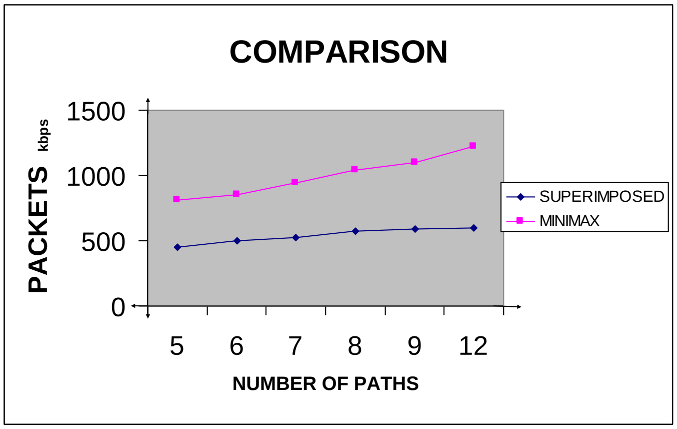


Figure 5.2: Data rate Vs Number of Paths

Table 3 Data rate Vs Number of Paths

Number of paths	Superimposed algorithm Data rate (packets per second)	Minimax optimizing algorithm Data rate (packets per second)
5	449	813
6	497	850
7	521	942
8	577	1039
9	588	1099
12	596	1220

VII. CONCLUSION

The optimization models related with routing, bandwidth utilization and power consumption are developed using the OR techniques like maximal flow algorithm, transshipment model and minimax optimizing algorithm

VIII. SUGGESTIONS FOR FURTHER WORK

1) Determination of frequency of running the model by exploiting the recharging capability of the nodes and the speed of dynamically changing topology of the network.

2) Analysis of performance of the model by comparing the simulation results of this model with the simulation results of other routing protocols related with power consumption and Bandwidth utilization in the wireless mesh computing environment.

3) The comparison metrics, with other existing methodologies, other parameters and with more demonstration.

REFERENCES

[1] R.Ahuja, T. Magnati and J. Orlin, (1993), "Network Flows Theory, Algorithms and Applications", Prentice Hall, Upper Saddle River, N.J.

[2] Thomas H.cormen, Charles E.leiserson, Ronald L.rivest,(2001), "Introduction to Algorithms",3rd ed., Prentice Hall, New Delhi.

[3] N. Bambos (June 1998), "Toward Power-sensitive Network Architectures in Wireless Communications: Concepts, Issues, and Design Aspects", IEEE Personal Communications Magazine, pp. 50-59.

[4] M. Bazaraa, J. Jarvis, and H. Sherali (1990), "Linear Programming and Network Flow", 2nd ed., Wiley, New York.

[5] Benjie Chen, Kyle Jamieson, Hari Balakrishnan, and Robert Morris (July 2001), "Span: An energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks", in 7th Annual Int. Conf. Mobile Computing and Networking 2001, Rome, Italy.

[6] D. Bertsekas, R. Gallager (2000), "Data Networks", 2nd ed, Prentice Hall of India, New Delhi.

[7] Philips, Solberg, Ravindran (1976), "Operation Research – techniques and Practice" ,John Wilen & sons, New York.

[8] B. Brumitt (Oct 2000), "Ubiquitous computing and the roles of Geometry", IEEE Pers. Commun.pp 47-53.

[9] J.-H. Chang and L. Tassiulas (Sept. 1999), "Routing for maximum system lifetime in wireless ad-hoc networks," in *Proceedings of 37-th Annual Conference on Communication, Control, and Computing*, Monticello, II.

[10] J-H Chang and L. Tassiulas (March 2000). "Energy Conserving Routing in Wireless Ad-hoc Networks." Proceedings of IEEE Infocom 2000. Tel Aviv, Israel.

[11] A. Croll, E. Packman (2001), "Managing Bandwidth", Pearson Education Asia.

[12] K. Edwards and Rebecca Grinter (September 2001), "At Home with Ubiquitous Computing: Seven Challenges", *Ubiquitous Computing 2001*, Atlanta, GA.

[13] J.M. Rabaey (2001), "Wireless beyond the 3rd generation facing the energy challenge", proc. 2001 Int's symp. Lower power electronics and design (ISLPED 01), ACM Press, New York, PP 1-3.

[14] T.S. Rappaport (1996), "Wireless Communications: Principles and practice", prentice Hall, New Hersey.

[15] V. Rodoplu, et al (June 1998), "Minimum energy mobile wireless networks", Proceedings of IEEE ICC, Atlanta, GA, Vol.3, PP 1633-1639.

[16] M. Satyanarayanan (Aug 2001), "Pervasive Computing: Vision and Challenges", IEEE Pers. Commun., PP 10-17.

[17] Fredrick S. Hiller, Gerald, J.Lieberman,(1990), "Introduction to operation research", McGraw-Hill edition.

[18] M. Steenstrup (1995), "Routing in Communications Networks", Prt.Hall-Inc.

[19] H.A. Taha (2001), "Operations Research: An introduction", 6th ed. Prentice Hall of India, New Delhi.

[20] R. Want et. al., (Jan-Mar' 2002), "Disappearing Hardware", IEEE Pervasive computing, PP 36-47.

[21] Adrian Deaconu, Eleonor Ciurea, Corneliu Marinescu (Oct 2010) "A Study on the feasibility of the inverse maximum flow problems and flow modification techniques in the case of non-feasibility", ACM WSEAS Transactons on Computers, Vol 9 issue 10, PP 1098 – 1107.

Time-Domain Large Signal Investigation on Dynamic Responses of the GDCC Quarterly Wavelength Shifted Distributed Feedback Semiconductor Laser

Abdelkarim Moumen, Abdelkarim Zatni, Abdenabi Elyamani, Hamza Bousseta
M.S.I.T Laboratory, Department of Computer Engineering high school of technology, Ibnou Zohr University.

Abdelhamid Elkaaouachi
Department of Physics, Faculty of Sciences, Ibnou Zohr University.

***Abstract*—A numerical investigation on the dynamic large-signal analysis using a time-domain traveling wave model of quarter wave-shifted distributed feedback semiconductor lasers diode with a Gaussian distribution of the coupling coefficient (GDCC) is presented. It is found that the single-mode behavior and the more hole-burning effect corrections of quarter wave-shifted distributed feedback laser with large coupling coefficient can be improved significantly by this new proposed light source.**

Keywords-component; Distributed feedback laser; optical communication systems; Dynamic large signal analysis; Time domain model.

I. INTRODUCTION

Long-haul modern Fiber-Optic Telecommunication Systems need optical source with high quality: high output optical power, low threshold current and reduced spatial hole burning effects, the longitudinal side mode are undesirable due to the presence of fiber dispersion [1][2][3][4]. The distributed feedback semiconductor lasers diode (DFB) have attracted great attention as the most favorable candidate. But the main disadvantage of this laser was the mode degeneracy and high threshold [1][6]. A phase shift along laser cavity can be introduced to remove the mode degeneracy [2]. Experimental results and numerical simulations have shown that the quarterly wavelength shifted distributed feedback laser (the phase shift is located at the center of the cavity and its value is fixed at 90^0) oscillates at the Bragg wavelength, presenting the smallest threshold current and the high gain selectivity when compared to other λ Phase-Shifted DFB laser diodes [1][2]. However, presences of the phase shift in the grating of DFB laser generally causes spatial no-uniformity and more interaction between the photon and carrier densities, especially for at high injection currents, this phenomenon, called spatial hole burning effect, reduce the performances of the λ Phase-Shifted DFB lasers diodes [2][4]. Recently, $\lambda/_4$ Phase-Shifted DFB with Gaussian distribution of the coupling coefficient (GDCC QWS) is proposed [1] to overcome the influence of spatial hole burning effect by maintaining uniform internal filed along the laser cavity and reduce the threshold current, extensive studies have verified that stable single-mode and high power operation can be achieved in GDCC lasers with large coupling coefficient $\kappa L = 2.5$, the studies is conducted by the Transfer Matrix Model (TMM) [1]. However, the relative important characteristics of this structure as the dynamic response have not been considered in their investigation.

In this letter the study consists in comparing the performance of the new proposed light source (GDCC QWS DFB) and conventional lasers having same total coupling coefficient in order to show the superiority of the GDCC configurations. The transient responses of the devices under analysis will be analyzed by using the time-domain multimode algorithm that is capable of including the longitudinal variation of the optical-mode and photon density profiles, the parabolic model of material gain is assumed [1]. In addition the spontaneous emission noise, the no uniform carrier density resulting from the hole burning effects as well as that the refractive index distribution are also taken into account, As a result this model may be applied to multi-sections lasers, such as phase-tunable lasers, tunable lasers and lasers designed to compensate for spatial hole burning (Subject of this paper). The model may also be applicable to tunable DFB laser amplifiers, the noise properties of DFB laser amplifiers and to bistable DFB switches.

The paper is organized as follows: the time-domain model is briefly described in section II. Simulation resultants of structures under analysis are presented and discussed in section III. Finally, a brief conclusion is drawn.

II. TIME-DOMAIN MODEL (TDM)

For the phase-shifted distributed feedback semiconductor laser diode, the electric field in the laser cavity is given by [8]:

$$\boldsymbol{\psi}(\boldsymbol{z},\boldsymbol{t}) = \left[\boldsymbol{R}(\boldsymbol{z},\boldsymbol{t})\boldsymbol{e}^{-\boldsymbol{i}\boldsymbol{\beta}_0\boldsymbol{z}} + \boldsymbol{S}(\boldsymbol{z},\boldsymbol{t})\boldsymbol{e}^{+\boldsymbol{i}\boldsymbol{\beta}_0\boldsymbol{z}}\right]\boldsymbol{e}^{-\boldsymbol{i}\boldsymbol{\omega}_0\boldsymbol{t}} \qquad (1)$$

Where β_0 is the propagation constant at Bragg frequency and ω_0 is the reference frequency. R and S are the slowly varying complexes fields components include the amplitude and phase information of the forward and reverse wave in the waveguide respectively.

The fields R and S can be derived from the Maxwell's equations using the slowly varying amplitude approximation. Time-dependent coupled equation can be written as [8][11]:

$$\frac{1}{c_g}\frac{\partial R}{\partial t}+\frac{\partial R}{\partial z}=$$

$$\left[\left(\frac{1}{2}\left(\frac{\Gamma g(z,t)}{1+\varepsilon P}-\alpha_l\right)-i\delta\right)R+i\kappa(z)S\right]e^{-i\varphi(z)}+\xi(z,t) \qquad (2)$$

$$\frac{1}{c_g}\frac{\partial S}{\partial t}-\frac{\partial S}{\partial z}=$$

$$\left[\left(\frac{1}{2}\left(\frac{\Gamma g(z,t)}{1+\varepsilon P}-\alpha_l\right)-i\delta\right)S+i\kappa(z)R\right]e^{+i\varphi(z)}+\xi(z,t) \qquad (3)$$

Where c_g is the group velocity, Γ is the optical confinement factor, $\kappa(z)$ is the coupling coefficient between the forward and backward propagation waves, α_l is the waveguide loss (includes the absorption in both the active and cladding layer as well as any scattering), $\varphi(z)$ is the phase shift at z position, ε is the gain compression coefficient (non-linear coefficient to take into account saturation effects) and ξ is the spontaneous emission term contributed to the forward and backward propagation components, the stochastic property of the noise term (ξ) is described by a random process with zero mean value and correlation function as described in [8][9][10] satisfying the correlation:

$$\begin{cases}\langle\xi(z,t),\xi^*(z',t')\rangle=\frac{\beta KBN^2}{c_gL}\delta(t-t')\delta(z-z')\\ \langle\xi(z,t),\xi(z,t)\rangle=0\end{cases} \qquad (4)$$

Where β is the spontaneous coupling factor, K is the Peterman Coefficient and $\frac{BN^2}{c_gL}$ is the bimolecular recombination per unit length contributed to spontaneous emission.

$g(z,t)$ is the material gain, given by the parabolic formula:

$$g(z,t)=A_0\Delta N(z,t)-A_1[\Delta\lambda+A_2\Delta N(z,t)]^2 \qquad (5)$$

In the above equation, A_0 is the differential gain, A_1 and A_2 are parameters used in the parabolic model assumed for the material gain, ΔN and $\Delta\lambda$ are the change of the carrier density and lasing wavelength defined as:

$$\begin{cases}\Delta N(z,t)=N(z,t)-N^0\\ \Delta\lambda=\lambda-\lambda_0\end{cases} \qquad (6)$$

N is the carrier density, N^0 is the carrier concentration at transparency $(g=0)$, λ the oscillating wavelength and λ_0 is the peak wavelength at transparency. Using the first-order approximation for the refractive index, one obtains:

$$n(z,t)=n_0+\Gamma\frac{dn}{dN}N(z,t) \qquad (7)$$

Where n_0 is the refractive index at zeros carrier injection and $\frac{dn}{dN}$ is the differential index.

The δ in equations (2) and (3) represent the mode detuning (derivation from Bragg condition) defined as:

$$\delta(z,t)=\frac{2\pi}{\lambda}n(z,t)-\frac{2\pi n_g}{\lambda\lambda_B}(\lambda-\lambda_B)-\frac{\pi}{\Lambda(z)} \qquad (8)$$

Where λ_B is the Bragg wavelength, λ is the lasing-mode wavelength, n_g is the group refractive index and Λ is the pitch (period) of the grating.

The carrier concentration $N(z,t)$ and the stimulated photon density are coupled together through the time-dependent carrier rate equation in the active layer which is shown here as [14]:

$$\frac{dN}{dt}=\frac{I}{qV}-\frac{N}{\tau}-BN^2-CN^3-\Gamma c_g g(z,t)\frac{P}{1+\varepsilon P} \qquad (9)$$

Where I is the injection current, q is the modulus of the electron charge, V is the volume of the active layer, τ is the carrier life time, B is the radiative spontaneous emission coefficient and C is the Auger recombination coefficient and P is the photon density, which is related to the magnitude of travelling wave amplitudes as:

$$P(z,t)=|R(z,t)|^2+|S(z,t)|^2 \qquad (10)$$

In the TDM simulation the Large-signal spatiotemporal response of the laser is obtained by solving directly in the time domain the coupled wave equation (2)-(3) and the carrier rate equation (9) with axially-varying parameters. A finite-difference time-domain algorithm is applied to these equations with uniform intervals of time and space, to take the spatial hole burning and the carrier induced refractive index fluctuation into consideration, the laser cavity is divided into a large number of Subsections $(M=5000)$ with length $\Delta z=\frac{L}{M}=c_g\Delta t$, L is the length of the cavity. In each section the material and structure parameters are kept constant, also the reflectivity at the end facet supposed to be zero. The numerical method followed here is similar to the one developed in [14].

The time-domain model is applicable to various types of semiconductor laser diodes. In this letter we apply the numerical model to compare to performance of the conventional quarterly wavelength shifted distributed feedback laser and GDCC QWS DFB laser having the same total coupling coefficient. In the proposed quarterly wavelength shifted distributed feedback laser the λ phase-shifted is located at the centre of the cavity and the coupling coefficient κ is a function of the longitudinal coordinate z , κ change continuously along the laser cavity as follows:

$$\kappa(z)=\kappa_0 e^{-G((z-\frac{L}{2})/L)^2} \qquad (11)$$

Where κ_0 the average value of the coupling coefficient, this parameter is introduced in order to allow a straightforward comparison between the characteristics of the GDCC QWS DFB and the conventional QWS DFB. The parameters definitions of these structures are summarized in table I, their distribution of the coupling coefficient are presented in the figure 1.

TABLE I. SUMMARY PARAMETRS DEFINITIONS OF STRUCTURES

Acronym	G	$\kappa_0 L$
Conventional QWS DFB	0	2,50
GDCC QWS DFB	1	2,7098

TABLE II. SUMMARY MATERIAL AND STRUCTURAL PARAMETRS

SYMBOL	PARAMETRS	VALUE
τ	Carrier lifetime	$4.10^{-9}s$
B	Bimolecular recombination	$10^{-16}m^3s^{-1}$
C	Auger recombination	$3.10^{-41}m^6s^{-1}$
N^0	Transparency carrier density	$1{,}5.10^{24}m^{-3}$
ε	Non-linear gain coefficient	$1{,}5.10^{-23}m^3$
A_0	Differential gain	$2{,}7.10^{-20}m^6$
A_1	Gain curvature	$1{,}5.10^{19}m^{-3}$
A_2	Differential peak wavelength	$2{,}7.10^{-32}m^4$
α_l	Internal absorption	4.10^3m^{-1}
n_g	Group index	3,7
c_g	Group velocity	$2{,}7.10^{-32}m^4$
L	Cavity length	$500\ \mu m$
D	Active layer thickness	$0{,}12\ \mu m$
w	Active layer width	$1{,}5\ \mu m$
V	Volume for active region	$90\ \mu m^3$
Λ	Grating period	$227{,}039\ \mu m$
λ_B	Bragg wavelength	$1550\ nm$
λ_0	Peak wavelength at transparency	$1565\ nm$ [1]
Γ	Optical confinement factor	0,35
φ	Phase shift	90^0
Ω	Residue corrugation phase at left facet	0^0

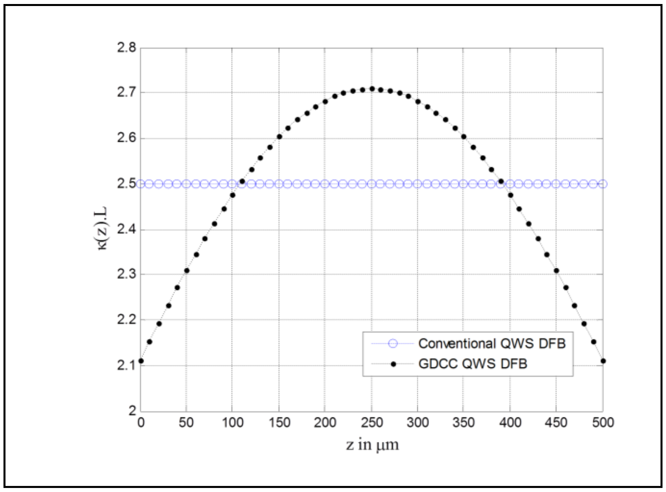


Figure 1. Normalized coupling coefficient configurations used for the numerical simulations.

RESULTS AND DISCUSSION

Modest injection levels ($I = 20mA$)

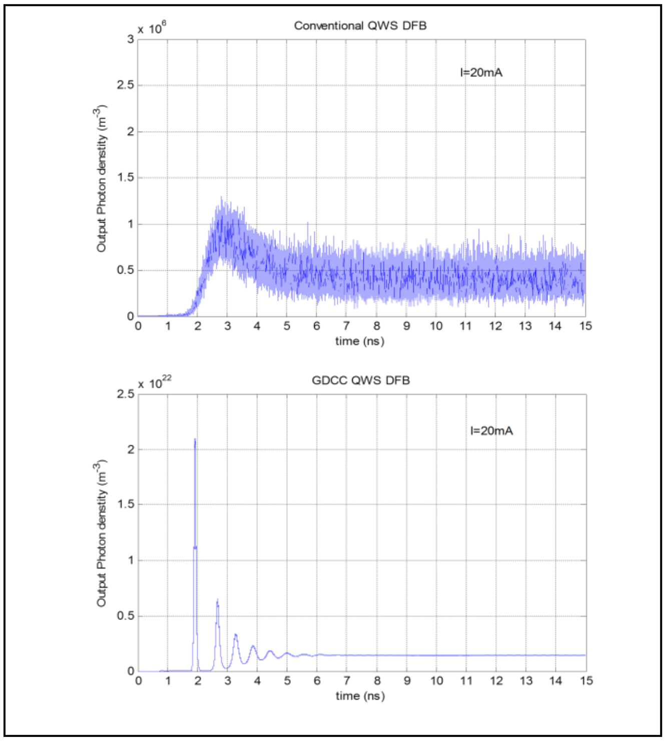


Figure 2. Transient response($current\ from\ 0mA\ to\ 20mA$) of output photon density for the Conventional QWS DFB and the GDCC-QWS DFB.

When the biasing currents trends toward the threshold (Figure 2) the cavity is the seat of a spontaneous emission noise in the case of conventional QWS. Therefore this biasing current is inadequate to initiate the laser effects; the threshold current of the conventional structure is more than $20mA$, while the

[1] According to the results obtained by the static study published in the [1]

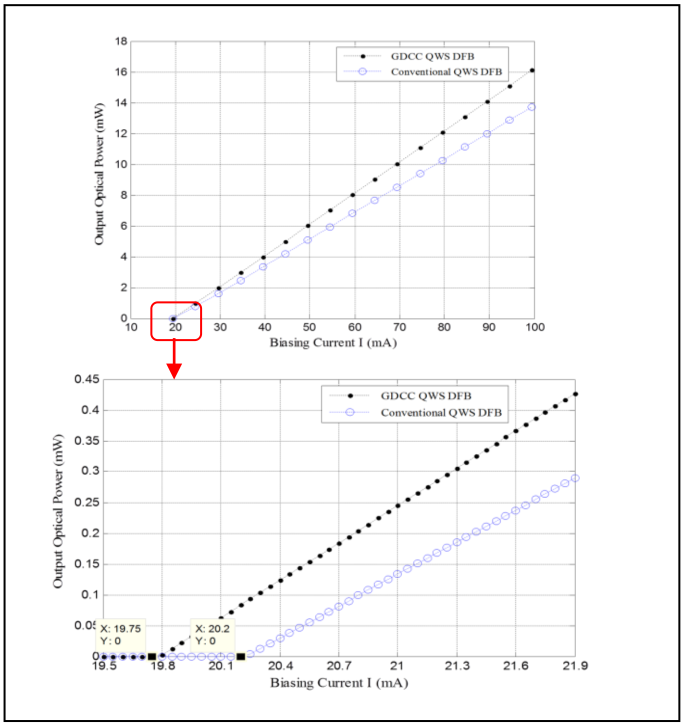


Figure 3. Emitted optical power versus current injection for the Conventional QWS DFB and the GDCC QWS DFB laser.

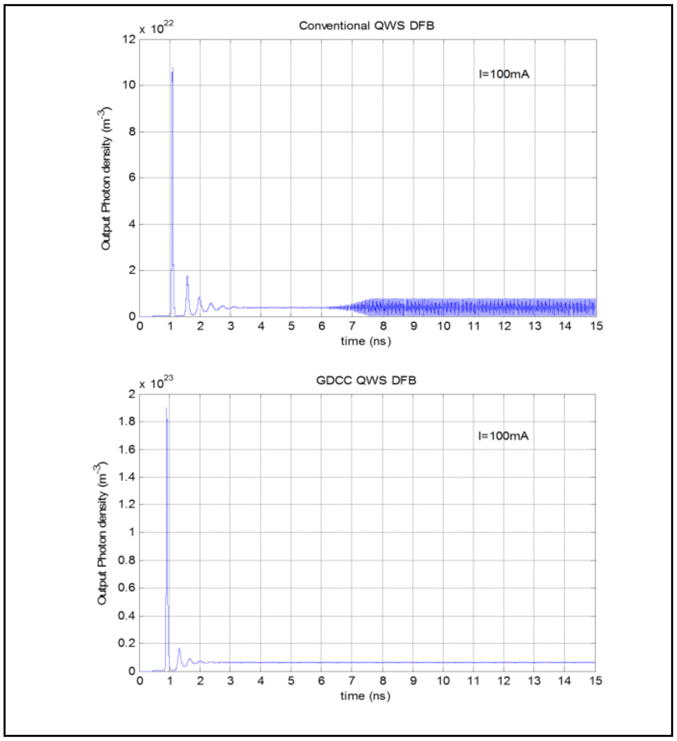


Figure 4. Transient response ($current\ from\ 0mA\ to\ 100mA$) of output photon density for the Conventional QWS DFB and the GDCC-QWS DFB.

turn-on-transient after the laser has switched and a typical oscillations are obtained in case of GDCC QWS DFB, this optical source has a low threshold current compared to the conventional QWS DFB, this first main advantage can be verified by evaluation of the optical power versus the current injection.

From the emitting photon density at the facet, the output optical power can be evaluated. Figure 3 summaries results obtained for the conventional QWS DFB and GDCC QWS DFB LDs with the biasing current as parameter. Compared with the standard QWS DFB, it seems that the use of a smaller coupling coefficient near the facet has increased the overall cavity loss (case of GDCC QWS DFB Laser structure). The figure also shows that the GDCC QWS DFB laser structure has à relatively smaller value of threshold current $I_{th} = 19{,}75mA$ and a relatively larger output power under the same biasing current.

High injection current $(I = 100mA)$

In the figure 4, the damping of transient in GDCC QWS DFB is better than for the conventional device. After some relaxation oscillations, other differences occur between the conventional and GDCC QWS; the output photon density starts to oscillate in strong amplitude as the consequence of the beating between two modes in the case of conventional QWS DFB. This is confirmed by taking a sample of the emission spectrum in two different moments:

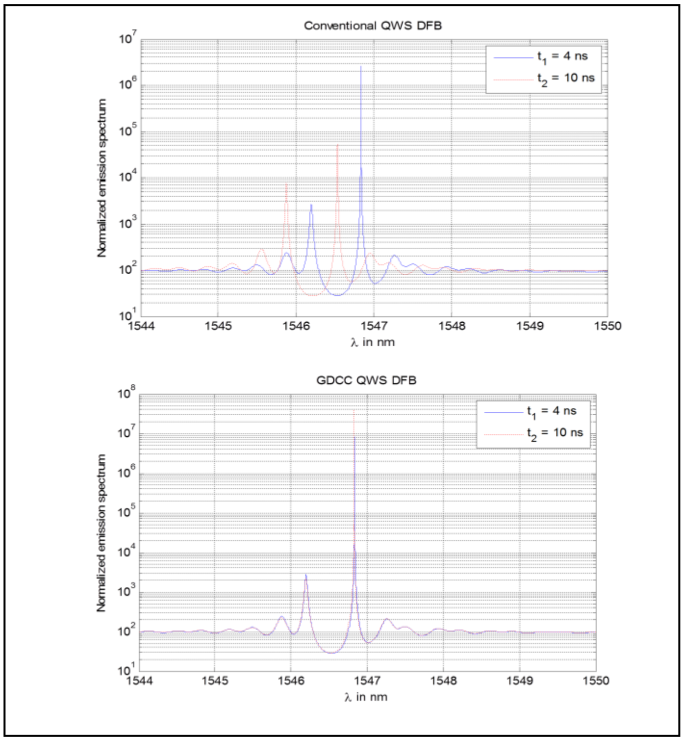


Figure 5. The Normalized emission spectrum in two different times, for the Conventional QWS DFB and the GDCC-QWS DFB.

The spectral characteristics of the GDCC QWS DFB laser structures with the time changes are shown in the figure 5, distinct peaks which correspond to different oscillating modes are observed along the spectrum; the spectral amplitude of the dominant lasing mode found near $1546{,}90\ nm$ shows no sign of reduction and remains at a high value near 10^6. Compared with the standard QWS DFB structure, the GDCC QWS DFB laser structure shows no server mode competition and an SMSR at least $25\ dB$ is maintained throughout of the time range.

In the case of the conventional QWS DFB, it can be seen that all peak wavelengths shift towards the shorter wavelength, and reduction of the spectral amplitude difference between the lasing mode and the side mode which is located at shorter wavelength side. At time $t_2 = 10\ n$, the side mode suppression ratio (SMSR) is reduced to less than $25\ dB$.

The variation of the longitudinal profiles of carrier density and refractive index can also indicate the occurrence of a multimode operation in DFB structures. As an illustration, we have plotted in the figure 5 the longitudinal profiles of refractive index in two distinct instants (t_1, t_2) and the statistic longitudinal standard deviation of carrier density in the period $[t_1, t_2]$ given by:

$$\boldsymbol{\sigma}(\boldsymbol{N_t}) = \sqrt{\boldsymbol{E}(\boldsymbol{N_t^2}) - (\boldsymbol{E}(\boldsymbol{N_t}))^2} \qquad (12)$$

The beating between two modes observed in the case of conventional QWS DFB (Figure 5) is caused by the longitudinal hole burning effects. This phenomenon alters the lasing characteristics of the QWS DFB LD by changing the refractive index along the cavity (Figure 6 especially in the case of conventional structure). Under a uniform current injection, the light intensity inside the laser structure increases with biasing current. For strongly coupled laser devices, most light concentrate at the centre of the cavity. The carrier density at the centre is reduced remarkably as a result of stimulated recombination. Such a depleted carrier concentration induces an escalation of nearby injected carriers and consequently a spatially varying refractive index results Figure 6. This figure also shows the temporal instability of the carrier density especially near the facets of the cavity Conventional, which explains the strong amplitude oscillations observed for output photon density in the figure 4.

III. Conclusion

With the help of a traveling wave model of semiconductor laser diodes, the dynamic analysis of Quarterly Wavelength Shifted Distributed Feedback Semiconductor Lasers with the Gaussian distribution of the coupling coefficient (GDCC) has been investigated and compared to conventional structures, to conduct this study we have developed a simple algorithm to calculate the large-signal dynamic response of DFB lasers by solving the time-dependent coupled wave equations and the carrier rate equation in the time domain. The spontaneous emission noise, longitudinal variations of carrier (hole burning) and photon densities as well as that of the refractive index are taken into consideration. The TDM was applied to GDCC QWS DFB and to Conventional QWS DFB, which is characterized by its uniform coupling coefficient, was shown to have a largest threshold current has the smallest output optical power. At high injection current, the conventional QWS structure is subject to mode beating and its output photon density starts to oscillate in strong amplitude as the result of the interference between the involved modes caused by the LSHB. Although the GDCC QWS DFB laser maintains a steady-state regime in which the output power becomes stabilized (no mode beating), no remarkable change in the spectral output in time, the damping of transient is better than for the conventional device. We may conclude that this new proposed light source can be used to extend the transmission distance in optical communication systems.

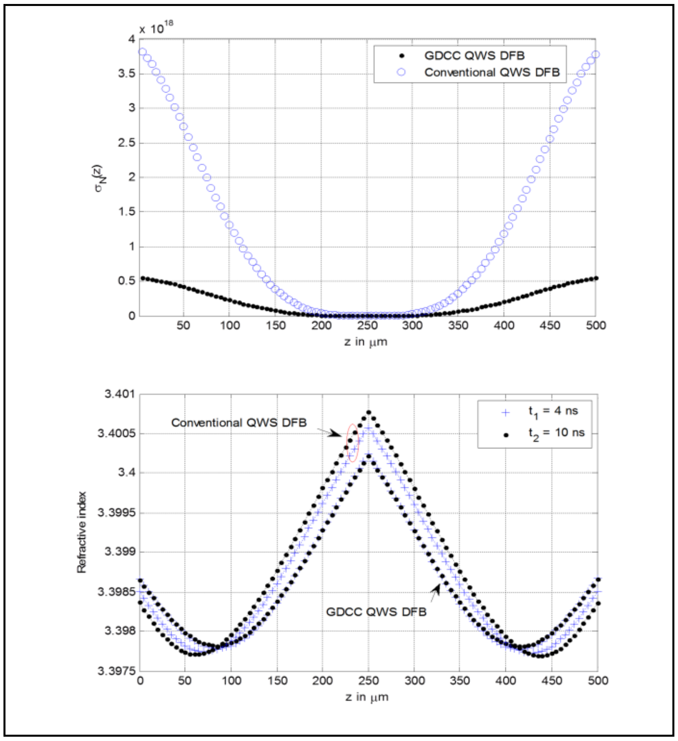


Figure 6. Statistic longitudinal standard deviation of carrier density in the period $[t_1, t_2]$ and the longitudinal distribution of refractive index in t_1 and t_2

Acknowledgment

This work was supported by the CNRST.

References

[1] A. Moumen, A. Zatni, A. Elkaaouachi, H. Bousseta, A. Elyamani, “A Novel Design of Quarter Wave-Shifted Distributed Feedback Semiconductor Laser for High-Power Single-Mode Operation,” *Journal of Theoretical and Applied Information Technology* , vol. 38, No. 2, May 2012,

[2] Ghafouri-shiraz, " Distributed feedback laser diodes and optical tunable filters," *Birminghman, UK: WILEY*, 2003

[3] A. Zatni; J. Le Bihan, "Analysis of FM and AM responses of a tunable three-electrode DBR laser diode," *IEEE Journal of Quantum Electronics*, vol. 31, pp. 1009-1014, 1995.

[4] C. Ferreira Fernandes,"Hole-burning corrections in the stationary analusis of DFB laser diodes," *materials sciences & engineering B*, B74, pp. 75-79, 2000

[5] A.Zatni, "Study of the short pulse generation of the three quarter wave shift DFB laser (3QWS-DFB)," *Annals Of Telecommunications*, vol. 60, pp. 698-718, 2005

[6] Carlos. A. F. Fernandes; Jose B. M, Bovida "optimisation of an asymmetric three phase-shift distributed feedback semiconductor laser structure concerning the above-threshold stability, " *The European Physical Journal Applied Physics*, vol. 49, pp. 1-9, 2010

[7] T. Fessant, "Threshold and above-threshold analysis of corrugation-pitch-modulated DFB lasers with inhomogeneous coupling coefficient," *IEE Proc., Optoelectron*, vol. 144, pp. 365-376, 1997.

[8] A. Zatni, J. Le Bihan, A. Charaia and D. Khatib, "FM and AM responses of a three-electrode DBR laser diode, " *The 1st International Conference on Information & Communication Technologies: from Theory to Applications - ICTTA'04*, pp. 167-168, 2004

[9] Xin-Hong Jia, Dong-Zhong, Fei Wang, Hai-Tao Chen, "detailed modulation response analyses on enhanced single-mode QWS-DFB lasers with distributed coupling coefficient," *Optics communications* vol. 277, pp. 166-173, 2007

[10] L. M. Zhang, S. F. Yu, M. C. Nowell, D. D. Marcenac, J. E. Caroll, and R. G. S. Plumb, "Dynamic analysis of radiation and side-mode suppression in a second-order DFB Laser Using Time-domain large signal traveling wave model," *IEEE journal of quantum electronics*, vol. 30, No. 6, pp. 1389-1395, 1994.

[11] Jacques W. D. Chi, Lu Chao; M. K. Rao, "Time-Domain Large-Signal Investigation on Nonlinear interactions between An Optical Pulse and Semiconductor Waveguides," *IEEE Journal of Quantum Electronics*, vol. 37, No. 10, octobre 2001

[12] Thierry Fessant, "Enhanced Dynamics of QWS-DFB Lasers by Longitudinal Varaiation of their Coupling Coefficient," *IEEE photonics technology lettres*, vol. 9, No. 8, agust 1997

[13] Jing.-Yi. Wang and Michael Cada, "analysis and optimum design of distributed feedback lasers using coupled-power theory," *IEEE journal of quantum electronics*, vol. 36, pp. 52-58, 2000.

[14] A. Zatni, D. Khatib, M. Bour, J. Le Bihan "Analysis of the spectral stability of the three phase shift DFB laser (3PS-DFB), " *Annals of Telecommunications*, vol. 59, pp. 1031-1044, 2004

[15] F.Shahshahani,V.Ahmadim K. Mirabbaszadeh "concave tapered grating design of DFB laser at high power operation for reduced spatial hole-burning effect," *materials science and engineering* , vol. B96, pp. 1-7, 2002.

[16] Thierry Fessant, "Influence of a Nonuniform Coupling Coefficient on the Static and Large Signal Dynamic Behaviour of Bragg-Detnued DFB Lasers," *IEEE photonics technology lettres*, vol. 16, No. 3, March 1998

[17] G.-X. Chen, W. Li, C.-L.Xu, W.-P. Huang, S.-S. Jian, "Time and Spectral Domain Properties of Distributed Feedback-Type Gain Clamped Semiconductor Optical Amplifiers," *IEEE photonics technology lettres*, vol. 18, No. 8, April 2006

[18] M. G. Davis, R. F. O'Dowd, "A Transfer Matrix Method Based Large-Signal Dynamic Model For Multielectrode DFB Lasers," *IEEE Journal of Quantum Electronics*, vol. 30, No. 11, November 1994

AUTHORS PROFILE

Abdelkarim. MOUMEN received the MSc degree in electrical and electronics system engineering from faculty of sciences University Ibnou Zohr in 2008; he is currently working the PhD at the centre of doctoral studies (Ibnou Zohr CED). His research interests include design, characterization, modelling and optimization of optoelectronic components and fibre optic communications systems.

Abdelkarim. ZATNI was educated at the Telecom Bretagne University France; He obtained a PhD at the National School of Engineers of Brest France in 1994. He has been teaching experience for 20 years. He is currently a Professor and the Head of computer science department in Ibnou Zohr University at Higher School of technology Agadir, Morocco; He conducts his research and teaches in computer science and Telecommunications.

Mutual Exclusion Principle for Multithreaded Web Crawlers

Kartik Kumar Perisetla
Department of Computer Science and Engineering
Lingaya's Institute of Management and Technology
Maharishi Dayanand University
Faridabad, India

***Abstract*— This paper describes mutual exclusion principle for multithreaded web crawlers. The existing web crawlers use data structures to hold frontier set in local address space. This space could be used to run more crawler threads for faster operation. All crawler threads fetch the URL to crawl from the centralized frontier. The mutual exclusion principle is used to provide access to frontier for each crawler thread in synchronized manner to avoid deadlock. The approach to utilize the waiting time on mutual exclusion lock in efficient manner has been discussed in detail.**

Keywords-* *Web Crawlers; Mutual Exclusion principle; Multithreading; Mutex locks.

I. INTRODUCTION

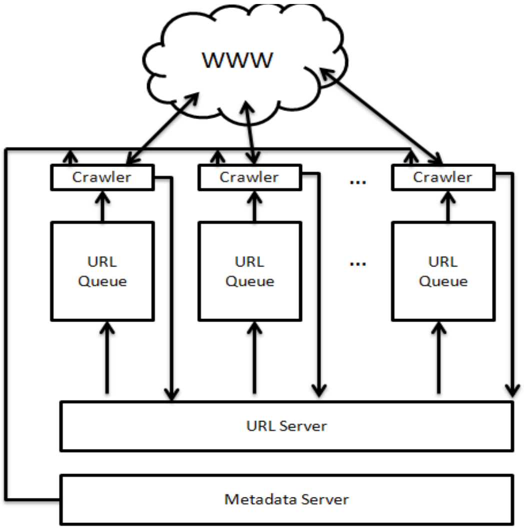
Web crawlers are programs that exploit the graph structure of the World Wide Web. The most important component of a search engine is an efficient crawler. World Wide Web is growing very rapidly; it is pertinent for search engines to opt for efficient and fast crawler processes to provide good results on search. Crawlers are also called as robots or spiders. Crawlers employed by search engines usually operate in multithreaded manner for high speed operation. When started multithreaded crawlers initialize a data structure, usually queue that holds the list of URLs to be visited by that crawler thread. These queues are filled constantly by a program employed within URL server which constantly monitors the count in each queue so that load on each crawler thread is balanced. The Load Balancing aspect is important to ensure efficient utilization of resources i.e. crawler threads. [1, 3]

Each thread start with a URL usually called a seed from their queue maintained in their local address space; they fetch the web page corresponding to that URL from World Wide Web, parse the page, extract the metadata and add links in this page to the frontier set which consists of the unvisited URLs. The data extracted consisting of body text, title, link text called as metadata are added into the metadata server. This metadata is further used by indexers for ranking the pages thus crawled. This ranked page set is then used by search engines as search results.

II. PROBLEM FORMULATION

A. *Problem statement and comparison model*

Traditional crawlers operate with a URL queue. The main drawback in this case is that each of them maintains a URL queue in local address. Initially, each thread holds 50 URLs to be visited. And each of them is to be monitored by single URL server program for adding new URLs to the queue as URLs are popped by crawler. Consider scenario where crawlers are operating in multithreaded manner and they access centralized URL frontier to fetch URL. Due to this, there might be cases of infinite waiting for crawler threads. To avoid such conditions and to provide synchronization among threads, mutual exclusion lock is used. Our focus is on comparison of operation model of multithreaded crawlers with synchronization lock and multithreaded crawlers without synchronization lock. We will analyze the behavior of these models and draw a conclusion based on performance.



B. *Experiment model*

We are considering a thread generator program capable of generating multiple crawler threads at a specified rate. Each crawler thread is capable of accessing the same centralized URL frontier, a database. The rate at which thread is generated can be easily controlled within the experimental setup to record observations. We will refer a model as "Non-mutex" when multiple threads operate without synchronization lock and we will refer a model as "Mutex" when multiple threads operate with synchronization lock to access shared resource. HTTP (Hypertext transfer protocol) is widely used for transfer of hypertext over the internet. Each thread fetches the page as a result of HTTP request and HTTP response actions. Each web server, according to robot exclusion protocol has a file named “robot.txt” that specifies which of the pages that are changed since robots last visited. But here we are ignoring that file meant for robots. In order to indicate the benefits of mutual exclusion lock in terms of performance we have also implemented the thread generator program without the mutual exclusion lock, hence in this case it is possible for more than one crawler thread to access the URL frontier at the same time. [5]

III. Mutual Exclusion Locks For Crawler Threads

We are considering a thread generator program generates the crawler threads at a specific rate that can be tuned to different values so as to record the observations for the experiment. Mutual Exclusion principle states that multiple processes or threads intending to access the same resource will access it mutually exclusively, that is only one at a time. This can be achieved by using a binary semaphore as mutual exclusion lock, ‘mutex’. Mutual exclusion for crawler threads applies in similar manner. When a crawler thread need to access the shared resource i.e. URL frontier, it check for the availability of the mutex lock. If it is in released state then it locks it and access the frontier. By that time if any other crawler threads need to access frontier it must wait until the lock is released by thread that holds the lock. Only one thread can access the URL frontier at a time hence providing controlled access and avoiding deadlock. Each thread fetches the URL to be visited from the URL frontier and establishes the connection with the web server. [6]

Pseudo code for mutex locks implementation:

```
while(mutex.isLocked())
//wait here until lock is released
Mutex.lock()
{//acquire the lock
//do processing here}
Mutex.ReleaseLock()
//release the lock
```

IV. Crawler Architecture

A. Structure

Crawler thread is the thread generated by a program. Thread runs in background mode in operating system. Crawler thread is responsible for fetching the web pages from worldwide web over HTTP. For non-mutex model, each crawler thread holds data structure for holding the raw data fetched from single source.

For mutex model, each crawler thread holds holds data structure probably a stack to hold raw data from multiple sources as discussed in latter sections. Also, in mutex model the thread generator program is responsible for providing the mutex lock to all crawler threads generated by it.

The data structure to hold the raw data is filled when HTTP response is received and it is flushed when the raw data is pushed into the raw fetched data store or the database for parsing. The threads generated by thread generator can be called as connections as each represent a connection with the web server. For example: 50 Crawler threads per sec.

B. Operation

As each thread is created it fetches a URL to be visited from the URL frontier, sends a HTTP request to the web server and waits for the HTTP response containing raw text of the page requested.

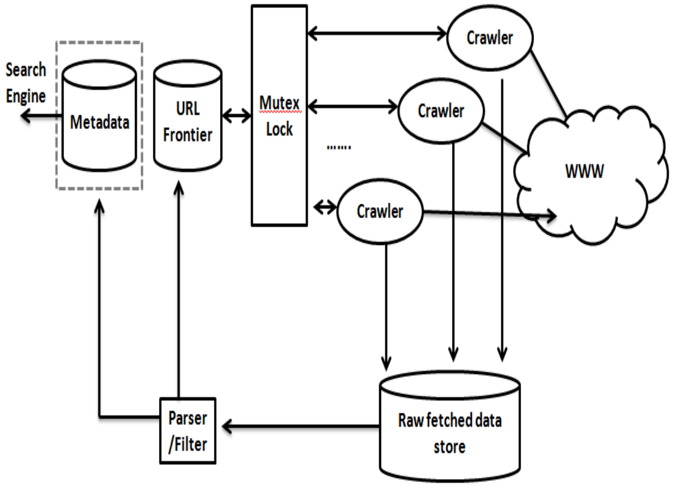


Figure 2. Mutithreaded Crawlers using the Mutual exclusion lock

By the time this thread is fetching the URL from frontier, all other threads wait for mutex lock to be released. Once the thread release the lock, another thread which was waiting for the lock acquires it. The next thread which gets this lock is dependent on how operating system manages the priority for providing the lock to next waiting thread.

The raw text thus received from HTTP response i.e. raw data is added to the ‘raw fetched data store’. And then this thread repeats its action from fetching the URL. All threads will terminate when there is no URL in URL frontier. The raw data fetched is to be processed to extract metadata and links from pages. Further processing is done by the ‘filter’ process. It reads the page extract title, outer text of the page, link text and adds it to the metadata store. Extracts links within the page and add them to the URL frontier. [7]

C. Pseudo Code

The pseudo code for crawler thread is shown below. This gives an insight on operations performed by crawler thread and sequence of those operations.

Description of each procedure is described as:

init: This procedure is called as soon as crawler thread is created. Purpose of this method is to initialize the thread with required data structures.

fetch_url: This is responsible for fetching URL from the URL frontier by using the mutex lock.

navigate_url: This is responsible for sending HTTP request and receiving the HTTP response for a URL.

```
init( )
{ fetch_url( )
}

fetch_url( )
{ while(mutex.closed( ))
     { }
   mutex.lock( )
   new_url=pop(url_frontier)
   mutex.release( )
     If new_url is Nothing then
          {exit}
   navigate_url(new_url)
 }

navigate_url( new_url)
{send_http_request(n_url)
  get_http_response(raw)
  push(raw_data_store,raw)
  fetch_url( )
}
```

D. *Crawler Algorithm*

Assuming that mutex represents the mutual exclusion lock at database level that provide synchronized access to crawler threads.

1. Check the locked status of mutex lock.

LockStatus=CheckMutexLockStatus[mutex]

2. If LockStatus=MUTEX_LOCKED then wait for lock top open by going to step 1. If LockStatus=MUTEX_OPEN then goto step 3.
3. Access the URL Frontier to pick next URL which is to be fetched and crawled to extract metadata.

nextURL=getNextURL()

4. Release the mutex lock so that it can be accessed by other threads

ReleaseMutexLock(mutex)

5. Fetch the raw web page and populate in appropriate data structure:

rawData=fetchRawPage(nextURL)

6. Repeat step 1, 2 to acquire lock. Once the lock is acquired, push the rawData to database:

pushRawPage(rawData)

7. Release the mutex lock :

ReleaseMutexLock(mutex)

8. Repeat steps 1 to 7 until URL frontier is empty.

V. Parser

A. *Structure*

Once the raw page data is pushed into the database the next step is to parse that data and extract meaningful metadata from it. This metadata acts fundamental information for search engine. The kind of elements parsed from raw data to generate metadata may vary as per the search engine requirements. In general the elements which are parsed to extract metadata are hyperlinks, title, Meta tag, headings, etc. For experiment a multithreaded parser was developed that can also generate parser threads at variable rate to extract information of raw pages and push them into database so that it can be readily used by the search engine.[8]

B. *Pseudo Code*

The pseudo code for Filter/Parser is shown below.

```
filter( )
{
   new_raw=pop(raw_data_store)
   new_meta=extract_meta_data(new_raw)
   push(meta_data_store,new_meta)
   extract_links(new_raw )
}
extract_meta_data(new_raw)
{
     //Extracts and returns the Metadata of the
page
}
extract_links(new_raw)
{
   for each url in new_raw
   {
      push(url_frontier,url)
    }
}
```

filter: This procedure is called as soon as filter process is initiated. Purpose of this method is to initialize the thread with required data structures.

extract_meta_data(new_raw): This procedure is responsible for extracting meta data from the page and adding it to raw fetched meta data store.

extract_links(new_raw): This procedure is responsible for extracting all URLs from the page and add them to URL frontier.

C. *Parser Algorithm*

Assuming that mutex represents the mutual exclusion lock at database level that provide synchronized access to parser threads.

1. Check the locked status of mutex lock.

LockStatus=CheckMutexLockStatus[mutex]

2. If LockStatus=MUTEX_LOCKED then wait for lock top open by going to step 1. If LockStatus=MUTEX_OPEN then goto step 3.

3. Access the raw page data from database:

   rawData=GetRawPageData()

4. Release the mutex lock so that it can be accessed by other threads

   ReleaseMutexLock(mutex)

5. Parse the page and extract metadata from it:

   metaData=ExtractMeta(rawData)

6. Repeat step 1, 2 to acquire lock. Once the lock is acquired, push the metaData to database:

   pushMetadata(metaData)

7. Release the mutex lock :

   ReleaseMutexLock(mutex)

8. Repeat steps 1 to 7 until URL frontier is empty

VI. Observations

A. *Time factor*

Consider let T be the combined time to fetch a page from the web, extract metadata and links from it. Now this T is composed of two components: time to fetch the page from web and time to parse the web page to extract links and metadata. Let t_f be the time to fetch the page and t_p is the time to parse the page to extract data from it. Then we can write T as:

$$T = t_{f\,+}\, t_p$$

A set of 2000 URLs is serving as the URL frontier at the beginning of the experiment. We performed our experiment for both crawling using mutex lock and crawling without mutex lock. Crawler threads are only responsible for fetching the web pages not parsing the pages. A ‘Filter’ program is used to parse the fetched web pages, extract links and metadata from them. Since we are using same URL frontier set for both mutex based and non-mutex based crawling, the ‘Filter’ program takes same constant amount of time to parse pages for both the cases. t_f includes the time to fetch the URL from frontier, time to send HTTP request and time to obtain the HTTP response. t_f can be written as:

$$t_f = t_{request} + t_{response}$$

Where $t_{request}$ is the time taken by request to reach the server and $t_{response}$ is the time taken for response to reach the crawler. Above equation holds good for models where the parser can directly get the raw data from the crawler thread for parsing. For models where the parser threads write the raw page data fetched from a URL to the centralized database, the equation can be written as:

$$t_f = t_{request} + t_{response} + t_{pushToStore}$$

$t_{pushToStore}$ is the time to acquire the mutex lock, write the raw data and to release the lock. $t_{request}$ can be further broken down into $t_{pickurl}$ and $t_{httprequest}$. $t_{pickurl}$ is time spent waiting for mutex lock, acquire mutex lock for database, access next URL and release the mutex lock. $t_{httprequest}$ is the time taken to create HTTP request and send it to respective endpoint. $t_{response}$ depends on several factors like speed of the internet connection, load on the web server serving that page and many other factors. The only parameter we can control is $t_{request}$. This is the only factor that can be controlled to minimize the t_f.

B. *Time Minimization*

The minimization of $t_{request}$ was performed in this experiment within the variable rate crawler thread generator. Generator provides provision to set rate at which the crawler thread will be generated. Once the page is crawled, its raw source is pushed into database with other relevant information specific to URL resource. The parser threads are responsible for parsing the raw page and extract useful metadata from it that can be fed to the search engine. These threads too are executed in multithreaded manner where synchronization between thread is done through mutex lock at database level. Based on observations recorded by generating crawler threads at variable rates, a graph is plotted for $t_{pickurl}$ against threading rate and is shown below:

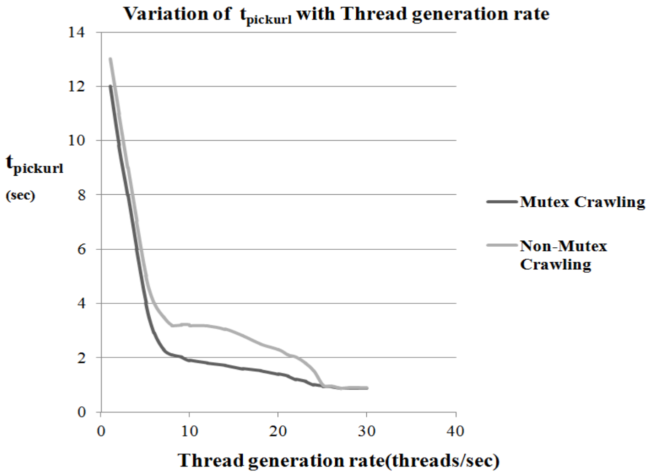


Figure 3. Graph for $t_{pickurl}$ vs. thread generation rate

C. *Utilizing mutex lock waiting time in crawler thread*

Consider the case when a crawler thread holds the mutex lock and other threads are waiting for the lock to read the next URL from the frontier. Here we are considering the mutex model where mutex lock is used for synchronization. Under normal operation conditions the probability of majority of threads waiting for mutex lock is high. This totally depends on the $t_{f,}$ the time to fetch the raw page. It was observed that majority of threads have similar t_f. Thus they end up fetching the page in same time and spend most of time waiting for mutex lock to fetch next URL. The waiting time for crawler thread can be utilized by employing that time for fetching raw data for subsequent URLs. We name this approach as extended crawling. The change required in crawler thread is that rather than picking a single URL from the frontier it picks collection of URLs whose raw data is to be fetched. This collection of URLs is pushed onto a stack STK[URL]. Once raw page data for a URL is fetched crawler checks for availability of mutex lock. If lock is held by any other thread then current thread pushes the raw fetched data onto a stack STK[RAW] and pops the next URL from the STK[URL]. Then crawler fetches the raw page data for this next URL popped. So this way the waiting time is utilized for fetching raw data for collection of URLs. In this model each thread will push the raw data to database in short bursts whenever the mutex lock is acquired by the thread.

The proposed algorithm for utilizing the waiting time for extended crawling can be written as:

```
1. Check the locked status of mutex lock
   LockStatus=CheckMutexLockStatus[mutex]

2. If LockStatus=MUTEX_LOCKED then goto step3. If
   LockStatus=MUTEX_OPEN then goto step 7

3. Pop the URL from STK[URL] to fetch the page while
   the mutex lock is held by other threads:
       nextURL=STL[URL].Pop()

4. Fetch the raw page data using the URL popped in
   previous step:
       rawData=HTTPFetch(nextURL)

5. Push the fetched raw data onto STK[RAW]:
       STK[RAW].Push(rawData)

6. Repeat Step 1 to acquire the lock.

7. Pop the fetched raw page data form top of stack
   STK[RAW] and write it to database:
       rawData==STK[RAW].Pop()
       CommitToDatabase(rawData)

8. Repeat step 7 until stack is empty. Once stack is empty
   repeat steps 1 to 6.
```

Consider the variation of t_f, $t_{request}$ + $t_{response}$ and $t_{pushToStore}$ with thread generation rate for a single thread. The dark shaded region shows the time spent in sending the request and fetching the page in the waiting time for the mutex lock by crawler thread. The dark black line shows the variation of total time to fetch the page (t_f). The light shaded region shows the variation of $t_{pushToStore}$ with thread generation rate. In case the mutex waiting time would not have utilized, the region under dark line (t_f) will be light shaded which mainly consists of time spent waiting on lock after the page is fetched. The following graph shows that large portion of t_f, i.e. waiting time on mutex lock is utilized for fetching raw pages for subsequent URLs.

D. Utilizing mutex lock waiting time in parser thread

Consider t_p, this factor is highly variable based on the amount of elements on crawled page. Higher the number of elements on the crawled page, higher the parsing time. t_p can be broken down into t_{parse} and $t_{pushMetadata}$. t_{parse} is the time taken to parse the raw page and fill the appropriate data structures.

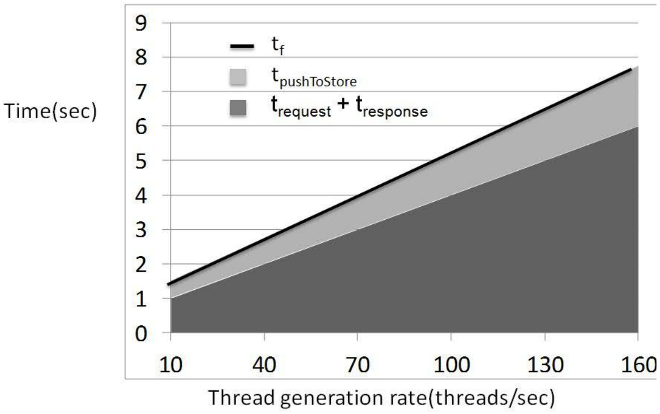


Figure 4. Graph for t_p vs. thread generation rate

The table shows the observations for $t_{request}$ + $t_{response}$ and $t_{pushToStore}$ at different number of crawler threads:

TABLE I. Variation of $t_{request}$ + $t_{response}$ and $t_{pushToStore}$ with thread generation rate

Parser Threads	$t_{request}$ + $t_{response}$ *(sec)*	$t_{pushToStore}$ *(sec)*
10	1	0.5
40	2	0.75
70	3	1
100	4	1.25
130	5.1	1.5
160	6	1.75

$t_{pushMetadata}$ is the time spent waiting for mutex lock, acquire mutex lock, save changes in database, commit the changes and release the mutex lock. Under normal operation conditions the probability of majority of threads waiting for mutex lock is high. The reason is that most of threads might finish parsing operation at same time and they wait for lock if it is acquired by other thread. The waiting time for a parser thread can be utilized by employing that time for parsing subsequent raw pages by picking up another raw page data and parsing it. We name this approach as extended parsing.

The change required in parser thread will be that rather than fetching raw page data for single page the parser will fetch collection of raw page data from the database and push collection onto a stack STK[RAW]. Once the parser finishes parsing raw page and if the mutex is locked then parser can pop raw page data for other pages held in stack and start parsing them. The parsed metadata set can be pushed on the stack STK[META] for pages parsed while waiting for mutex lock. Once the lock is acquired by the thread, it can write all parsed metadata which is held on stack to the database and release the lock. In this model each thread will push parsed metadata to database in short bursts whnever the mutex lock is acquired. This way the waiting time for mutex lock can be utilized for parsing the raw page.

The proposed algorithm for utilizing the waiting time for extended parsing can be written as:

1) *Check the locked status of mutex lock*
2) *LockStatus=CheckMutexLockStatus[mutex]*
3) *If LockStatus=MUTEX_LOCKED then goto step3. If LockStatus=MUTEX_OPEN then goto step 7.*
4) *Pop the raw page data from STK[RAW] to extract metadata from the page while the mutex lock is held by other threads:*
5) *rawData=STL[RAW].Pop()*
6) *Parse the page and extract metadata from it:*
7) *metaData=ExtractMeta(rawData)*
8) *Push the extracted metaData onto STK[META]:*
9) *STK[META].Push(metadata)*
10) *Repeat step 1 to acquire the lock.*
11) *Pop the metadata from top of stack STK[META] and write the metadata to database:*
12) *metadata= STK[META].Pop()*
13) *CommitToDatabase(metadata)*
14) *Repeat step 7 until stack is empty. Once stack is empty repeat steps 1 to 6.*

Consider the variation of t_{parse}, $t_{pushMetadata}$ and t_p with thread generation rate for a single thread. The dark shaded region shows the time spent in parsing the raw pages in the waiting time for the mutex lock by filter thread. The dark black line shows the variation of total time for parsing (t_p). The light shaded region shows the variation of $t_{pushMetadata}$ with thread generation rate. In case the mutex waiting time would not have utilized, the region under dark line (t_p) will be light shaded which mainly consists of time spent waiting on lock after parsing is complete. The following graph shows that large portion of t_p, i.e. waiting time on mutex lock is utilized under the parsing for subsequent set of raw pages.

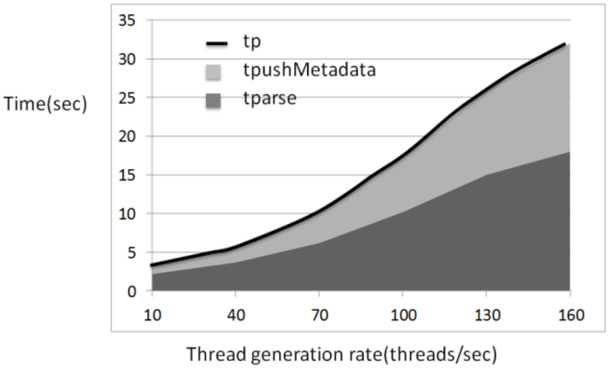


Figure 5. Graph for t_p vs. thread generation rate

The table shows the observations for t_{parse} and $t_{pushMetadata}$ at different number of parser threads:

TABLE II. Variation of t_{parse} and $t_{pushMetadata}$ with thread generation rate

Parser Threads	***t_{parse} (sec)***	***$t_{pushMetadata}$ (sec)***
10	2.2	1
40	3.7	2.2
70	6.2	4
100	10.2	7
130	15	11
160	18	14

VII. RESULTS

The experiment was conducted on Windows XP sp-2 operating system equipped with 512MB RAM, 512 kbps ADSL broadband connection. We are calculating time t_f, the time to fetch the fetch the page. The variation of $t_{pickurl}$ with thread generation rate has been discussed. Also, the experiment results involving utilization of mutex waiting time for parsing raw pages indicates the gravity of the approach. It can be deduced from the graph that multithreaded crawlers works efficiently only with the usage of mutual exclusion lock.

We can observe that for lower rate values, small increase in rate brings down $t_{pickurl}$ by large amounts. For larger rate values, large increase in rate brings small change in $t_{pickurl}$.

Also, it presents new approach to utilize mutex waiting time for parsing operation. This leads to increased performance of the crawler, parser and efficient utilization of resources.

VIII. FUTURE SCOPE

The future work will focus on minimizing the time incurred in acquiring the lock, writing data to database and releasing the lock. This time is represented as grey section in graphs shown in this document.

This may be accomplished by interacting with operating systems at a lower level to speed up the locking and releasing the mutex lock. Also, we will cover the aspects that will enhance the performance by providing an efficient synchronization model across crawler and parser threads.

IX. CONCLUSION

This paper presented a new approach for implementing multithreaded crawlers using mutual exclusion locks, which results in performance improvement as compared to traditional crawlers.

The approach of utilizing mutex waiting time proves efficient if employed for parsing or other useful operations within crawler threads.

REFERENCES

[1] Lawrence Page, Sergey Brin. The Anatomy of a search Engine. Submitted to the Seventh International World Wide Web Conference (WWW98). Brisbane, Australia

[2] Budi Yuwono, Savio L.Lam, Jerry H. Ying, Dik L. Lee. A World Wide Web Resource Discovery System. The Fourth International WWW Conference Boston, USA, December 11-14, 1995.

[3] Gautam Pant, Padmini Srinivasan, Filippo Menczer. Crawling the Web.

[4] Allan Heydon, Marc Najork. Mercator: A Scalable, Extensible Web Crawler

[5] Muhammad Shoaib, Shazia Arshad. Design and Implementation of web information gathering system

[6] Joo Yong Lee, Sang Ho Lee. Scrawler: A Seed by Seed Parallel Web Crawler.

[7] Boldi P., Codenotti B., Santini M., and Vigna S. UbiCrawler: a scalable fully distributed web crawler. Software Pract. Exper., 34(8):711–726, 2004

[8] S.chakraborti, M.van den Crawling: A new approach to topic-specific web resource discovery”. In the 8th International World Wide Web Conference, 1999

AUTHORS PROFILE

Kartik Kumar Perisetla received his Bachelors degree in Computer Science from Lingaya’s Institute of Management and Technology. He is currently working as Software Engineer. His research interest include Grid Computing, Machine Learning and Web Crawling

A Novel Image Encryption Supported by Compression Using Multilevel Wavelet Transform

Ch. Samson[1]
Dept. of Information Technology, SNIST,
Hyderabad, India,

V. U. K. Sastry[2]
Dept. of Computer Science & Engineering., SNIST,
Hyderabad, India,

Abstract— **In this paper we propose a novel approach for image encryption supported by lossy compression using multilevel wavelet transform. We first decompose the input image using multilevel 2-D wavelet transform, and thresholding is applied on the decomposed structure to get compressed image. Then we carry out encryption by decomposing the compressed image by multi-level 2-D Haar Wavelet Transform at the maximum allowed decomposition level. These results in the decomposition vector C and the corresponding bookkeeping matrix S. The decomposition vector C is reshaped into the size of the input image. The reshaped vector is rearranged by performing permutation to produce encrypted image. The vector C and the matrix S serve as key in the process of both encryption and decryption. In this analysis, we have noticed that the reconstructed image is a close replica of the input image.**

Keywords- Image compression; wavelet transform; thresholding; image encryption; compression ratio.

I. Introduction

An image is to be compressed so as to reduce the storage space and increase the speed of transmission. Image compression [1] is of two types: lossy or lossless. In lossless compression, the recovered data is identical to the original, whereas in the case of lossy compression the recovered data is a close replica of the original with minimal loss of data. Lossless compression can be used for text, medical images and legal documents etc. whereas lossy compression is used for natural images, speech signals etc. Images are widely used on several processes, including the Internet, and hence protecting confidential image data from unauthorized access has become an important issue in information security. Cryptography plays a vital role in information security. Cryptography [2] is the art or science that transforms a message (plaintext) into an unintelligible form (ciphertext) and then retransforms that message back to its original form.

Wavelets [3] have gained widespread acceptance in signal processing and image compression applications due to their utility in multi-resolution analysis. A basic wavelet is an oscillatory function that has limited duration. Wavelets are obtained from a single prototype wavelet called mother wavelet by dilations and shifting. Mathematically a wavelet is denoted by the function.

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi(\frac{t-b}{a})$$

where a is the scaling parameter and b is the shifting parameter. The transform based on wavelets is called wavelet transform. Wavelet decomposition of an image is used to analyze the image at different frequencies with different resolutions that gives specific information. This information can be used for processing the image, such as image compression. Wavelet transforms are of two types. One is Continuous Wavelet Transform and the other one is Discrete Wavelet Transform. Several researchers [4-10] have dealt with image compression using wavelet transform.

An alternative representation to wavelet transform is the multiwavelet transform [11-12]. Multiwavelets are very similar to wavelets but have some important differences. In particular, wavelets have an associated scaling function Φ(t) and wavelet function Ψ(t), whereas multiwavelets have two or more scaling and wavelet functions. Multilevel wavelet transforms find a wide variety of applications. They can be used can be used for compression, denoising, egde detection and encryption. In a recent investigation, Debayan et al. [13] have developed an algorithm for text encryption using multilevel 1-D wavelet transform.

In the present paper, our objective is to develop a novel method for image encryption supported by compression using multilevel 2-D Wavelet Transform. Firstly, we compress the input image using multilevel 2- dimensional wavelet transform and the compressed image is then encrypted by using a multilevel 2- dimensional Haar Wavelet Transform.

In what follows we present the plan of the paper. In section 2, we explain the proposed method. Section 3 describes the process of image compression using wavelet packet transform. We present a novel approach for wavelet-based image encryption in section 4. We provide an illustration in section5. Finally in section 6, we deal with computations that are carried out in this analysis and draw conclusions.

II. Proposed Method

When network bandwidth and storage space are limited, image has to be compressed. It is necessary to protect the image data during transmission from unauthorized access. Therefore to reduce the time for encryption, the image is first compressed prior to encryption. Reverse operations are performed at the receiving end to reconstruct the original image. The Schematic diagram of the proposed method is shown in Figure 1.

The proposed method is implemented by the following steps.

1. Decomposition: Choose a multilevel 2-D wavelet transform having the number of decomposition levels as N. Compute the wavelet decomposition of the input image at level N.

2. **Thresholding**: For each level from 1 to N, a threshold is selected and global thresholding is applied to the detail coefficients.

3. Encryption: The compressed image is encrypted by using multilevel 2-D Wavelet Transform (Haar).

4. Decryption: The reverse process of encryption is performed to get the compressed image.

5. Reconstruction: Perform multilevel 2-D wavelet reconstruction of the decrypted image to get a close replica of the original input image.

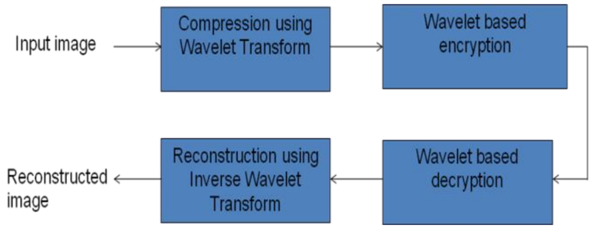

Input image
Compression using Wavelet Transform
Wavelet based encryption
Wavelet based decryption
Reconstruction using Inverse Wavelet Transform
Reconstructed image

Figure 1. Schematic diagram of the proposed method

The algorithm for wavelet based image encryption is given below.

Algorithm for image encryption

1) *Read the input (compressed) image.*

2) *Decompose the input image at the maximum allowed level, using multilevel 2-D Haar Wavelet Transform to get decomposition vector C and the corresponding bookkeeping matrix S.*

3) *Store the vector C and the matrix S.*

4) *Reshape the coefficients of the decomposition vector C to have the size of the input image (N -by-N).*

5) *Rearrange the vector coefficients by performing permutation to produce encrypted image.*

By performing inverse operations for the above steps in the reverse order, we get back the input compressed image which is the decrypted image. It is to be noted here that the decomposition vector C and the corresponding bookkeeping matrix S serve as key for both encryption and decryption.

III. Wavelet Approach For Image Compression

Image compression is one of the most successful applications of wavelet transform. The Wavelet Transform can be implemented using specially designed digital filters. Let us consider an image F(x,y) of size N×N. The samples of the input image are passed through a low pass filter and a high pass filter simultaneously, and the filter outputs are down-sampled by two along rows. Then the filter outputs can be further decomposed using the same filters and down-sampled by two again along columns, giving the approximation coefficients matrix (LL) and the detail coefficients matrices (LH, HL and HH) each of size N/2× N/2 as shown in Figure 2.

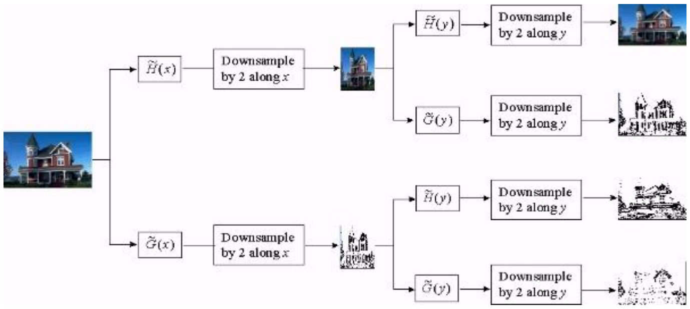


Figure 2. Wavelet Transform implementation.

To have a clear idea, Figure 2 can be seen as shown below.

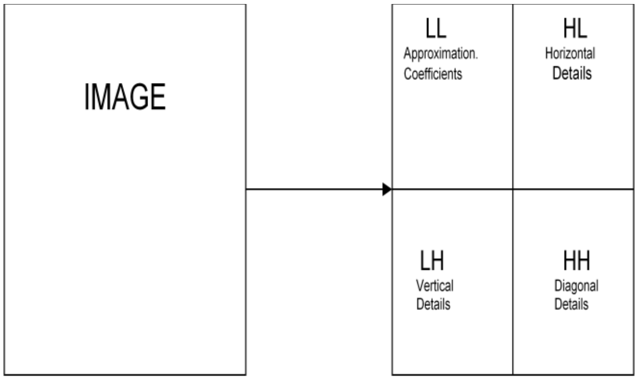


Figure 3. Wavelet Decomposition

The approximation coefficients matrix (LL) is called low resolution sub image. The sub images HL, LH and HH give horizontal, vertical and diagonal details respectively. multiwavelet decompositions produce two low pass subbands and two high pass subbands in each dimension. This kind of decomposition can be repeated to further increase the frequency resolution and the approximation coefficients decomposed with high and low pass filters and then down-sampled. In this analysis, we have conducted experiments using multilevel wavelet transforms based on Haar, Biorthogonal, Coiflet, Discrete Mayer Wavelet, Symlet, and we have taken the number of decomposition levels 3 to 5. However we have included levels 3 and 4 only in our analysis (See table I in section 6) for brevity in representation.

In the process of multilevel wavelet decomposition, many of the wavelet coefficients we have obtained are close to or equal to zero. Most of the information is included among a small number of the transformed coefficients. So, we truncate or quantize the coefficients including little information using thresholding. Thresholding can modify the coefficients to produce more zeros. Three types of thresholding [1] techniques can be used: local thresholding, global thresholding and dynamic thresholding. Local tresholding is one in which a different threshold is applied to each sub image where as a single threshold is applied to all sub images in global thresholding. Dynamic thresholding uses different thresholds for each coefficient separately. In our analysis, level-dependent global thresholds are selected based on Birge-Massart strategy and applied on detail coefficients as approximation coefficients cannot be thresholded. This will produce many consecutive zeros which can be stored in much less space and transmitted more quickly.

It is to be noted here that the low pass filter and the high pass filter are related to each other and they are known as the quadrature mirror filters which will make image reconstruction possible.

IV. Wavelet Based Image Encryprtion

In this section we present a novel method for image encryption using Wavelet Transform. The compressed image which we have obtained in section III is decomposed by multilevel 2-D Haar Wavelet Transform at the maximum allowed decomposition level and get the decomposition vector C and the corresponding bookkeeping matrix S. We reshape the decomposition vector C into a matrix form of size N×N. We rearrange the vector coefficients by performing permutation to obtain the encrypted image.

By performing inverse operations in the reverse order, we get back the input (compressed) image. The advantage of wavelet based image encryption is that the encryption time gets reduced and the decryption time also becomes small.

V. Illustraion of the method involving compression and encryption

Consider the image of Gandhiji of size 256x256 which is shown in Figure 4, given in section VI. Let us focus our attention on a portion P of the image of size 8x8 which lies in between the rows 1 to 8, and the columns 1 to 8. On representing this portion of the image in terms of its pixel values, we get the matrix given below.

$$P = \begin{bmatrix} 204 & 204 & 202 & 201 & 203 & 205 & 203 & 199 \\ 200 & 198 & 197 & 197 & 201 & 204 & 202 & 197 \\ 201 & 199 & 197 & 198 & 204 & 207 & 206 & 201 \\ 206 & 204 & 201 & 201 & 205 & 208 & 208 & 206 \\ 207 & 205 & 202 & 200 & 199 & 200 & 203 & 205 \\ 207 & 204 & 201 & 198 & 195 & 194 & 198 & 203 \\ 208 & 205 & 202 & 200 & 198 & 197 & 200 & 204 \\ 210 & 206 & 203 & 203 & 203 & 203 & 204 & 207 \end{bmatrix}$$

On decomposing P by using multilevel 2-D Wavelet Transform at the decomposition level 3, we get decomposition vector c and the corresponding bookkeeping matrix s in the form

c = (1.0e+003) *[1.6179 -0.0006 -0.0001 -0.0121 -0.0010 -0.0033 -0.0077 -0.0048 0.0055 0.0107 0.0037 -0.0087 -0.0010 0.0003 0.0023 -0.0018 0.0050 -0.0050 0.0005 -0.0015 0.0045 -0.0035 0.0015 -0.0020 0.0015 -0.0010 0.0050 -0.0055 0.0015 -0.0035 0.0035 -0.0035 0.0010 0.0020 0.0025 0.0035 0.0005 -0.0005 0.0025 0.0010 -0.0025 -0.0030 0 0.0005 0.0045 0.0035 -0.0035 -0.0035 -0.0010 0 -0.0005 -0.0005 0.0005 -0.0005 -0.0005 0.0010 0.0005 0 -0.0010 0.0005 -0.0005 0.0015 0.0015 -0.0005],

and

$$s = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 2 & 2 \\ 4 & 4 \\ 8 & 8 \end{bmatrix}$$

Level-dependent thresholds are obtained by using a wavelet detail coefficients selection rule based on Birge-Massart strategy [12]. However, we have to remember that the approximation coefficients cannot be thresholded. On using level-dependent thresholds, the decomposition vector c and the corresponding bookkeeping matrix s, compression is performed, and the resultant compressed image is obtained in the form

$$CP = \begin{bmatrix} 201 & 201 & 201 & 201 & 204 & 204 & 204 & 204 \\ 201 & 201 & 201 & 201 & 204 & 204 & 204 & 204 \\ 201 & 201 & 201 & 201 & 204 & 204 & 204 & 204 \\ 201 & 201 & 201 & 201 & 204 & 204 & 204 & 204 \\ 204 & 204 & 203 & 203 & 201 & 201 & 201 & 201 \\ 204 & 204 & 203 & 203 & 201 & 201 & 201 & 201 \\ 204 & 204 & 203 & 203 & 201 & 201 & 201 & 201 \\ 204 & 204 & 203 & 203 & 201 & 201 & 201 & 201 \end{bmatrix}$$

The compressed image matrix CP is decomposed by the multilevel 2-D Haar Wavelet Transform at the maximum allowed decomposition level to get the decomposition vector C and the corresponding bookkeeping matrix S. The decomposition vector C is reshaped into a matrix form of size N×N, and it is given by

$$rs = \begin{bmatrix} 402 & 0 & 0 & 0 & 0 & 0 & 0 & 408 \\ 402 & 0 & 0 & 0 & 0 & 0 & 0 & 408 \\ 408 & 0 & 0 & 0 & 0 & 0 & 0 & 402 \\ 408 & 0 & 0 & 0 & 0 & 0 & 0 & 402 \\ 402 & 0 & 0 & 0 & 0 & 0 & 0 & 408 \\ 402 & 0 & 0 & 0 & 0 & 0 & 0 & 408 \\ 406 & 0 & 0 & 0 & 0 & 0 & 0 & 402 \\ 406 & 0 & 0 & 0 & 0 & 0 & 0 & 402 \end{bmatrix}$$

The bookkeeping matrix S is given by

$$S = \begin{bmatrix} 4 & 4 \\ 4 & 4 \\ 8 & 8 \end{bmatrix}$$

We obtain the encrypted image matrix by performing permutation. Thus we have

$$E = \begin{bmatrix} 255 & 255 & 0 & 0 & 0 & 0 & 0 & 0 \\ 255 & 255 & 0 & 0 & 0 & 0 & 0 & 0 \\ 255 & 255 & 0 & 0 & 0 & 0 & 0 & 0 \\ 255 & 255 & 0 & 0 & 0 & 0 & 0 & 0 \\ 255 & 255 & 0 & 0 & 0 & 0 & 0 & 0 \\ 255 & 255 & 0 & 0 & 0 & 0 & 0 & 0 \\ 255 & 255 & 0 & 0 & 0 & 0 & 0 & 0 \\ 255 & 255 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

On performing inverse permutation on E, we get the decomposition vector in the form of a matrix of size 8x8. On reshaping it into vector form, we get a column vector given by

rC = [402 402 408 408 402 402 406 406 408 408 402 402 408 408 402 402 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]^T .

Here T denotes transpose of the vector. We have obtained the decrypted matrix D based on the multi-level wavelet decomposition structure [rC,S]. This is given by

$$D = \begin{bmatrix} 201 & 201 & 201 & 201 & 204 & 204 & 204 & 204 \\ 201 & 201 & 201 & 201 & 204 & 204 & 204 & 204 \\ 201 & 201 & 201 & 201 & 204 & 204 & 204 & 204 \\ 201 & 201 & 201 & 201 & 204 & 204 & 204 & 204 \\ 204 & 204 & 203 & 203 & 201 & 201 & 201 & 201 \\ 204 & 204 & 203 & 203 & 201 & 201 & 201 & 201 \\ 204 & 204 & 203 & 203 & 201 & 201 & 201 & 201 \\ 204 & 204 & 203 & 203 & 201 & 201 & 201 & 201 \end{bmatrix}$$

It is to be noted here that the decrypted matrix D and the compressed matrix CP are the same.

Thus, by performing multilevel 2-D wavelet reconstruction based on the decomposition vector c and its corresponding bookkeeping matrix s, we have reconstructed the matrix rP which is a close replica of the original input matrix P. This is given by

$$rP = \begin{bmatrix} 204 & 204 & 202 & 201 & 203 & 205 & 203 & 199 \\ 200 & 198 & 197 & 197 & 201 & 204 & 202 & 197 \\ 201 & 199 & 197 & 198 & 204 & 207 & 206 & 201 \\ 206 & 204 & 201 & 201 & 205 & 208 & 208 & 206 \\ 207 & 205 & 202 & 200 & 199 & 200 & 203 & 205 \\ 207 & 204 & 201 & 198 & 195 & 194 & 198 & 203 \\ 208 & 205 & 202 & 200 & 198 & 197 & 200 & 204 \\ 210 & 206 & 203 & 203 & 203 & 203 & 204 & 207 \end{bmatrix}$$

It may be noted here that the reconstructed matrix rP is an exact replica of the original input matrix P as the elements of the rP are rounded off to the nearest integer.

Here the decomposition vector and the corresponding bookkeeping matrix serve as key in the process of encryption and in the process of decryption.

VI. Computations And Conclusions

In this paper we have implemented a novel approach for image encryption supported by compression using multilevel wavelet transform in MATLAB[14] .

We have considered multilevel 2-D Wavelet Transforms, namely, 'haar' 'bior6.8','coif5','dmey' 'sym8' for image compression and multilevel 2-D Haar wavelet transform for image encryption. We have conducted experiments using the above wavelets for three test images 'Lena', 'Gandhiji' and 'Lady'. The input image of Gandhiji of size 256x256 and its corresponding compressed, encrypted, decrypted and reconstructed images are shown below for the decomposition level 4.

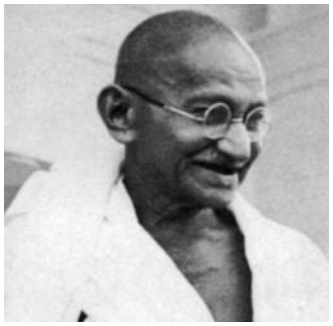


Figure 4. Input image of Gandhiji

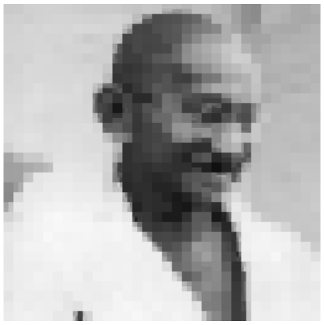


Figure 5. Compressed image

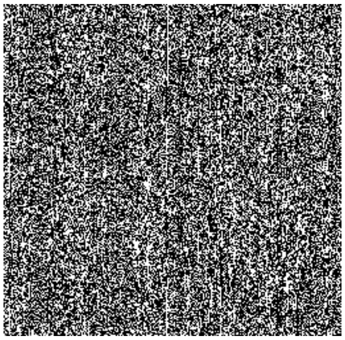


Figure 6. Encrypted image.

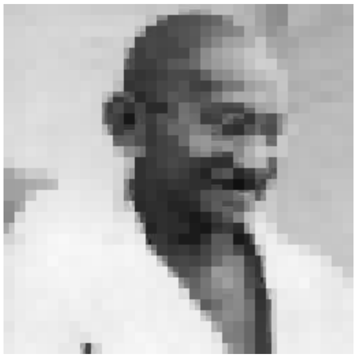


Figure 7. Decrypted image

We have calculated output parameters like compression score, compression ratio that determine the efficiency of the proposed system. Compression score is given by

Compression score in percentage = 100*(number of zeros of the current decomposition)/ number of coefficients)

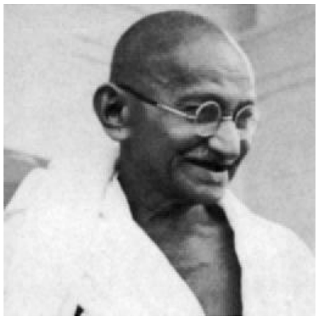


Figure 7. Reconstructed image

Compression ratio (C_R) is defined as

$$C_R = \frac{Uncompressed\ File\ Size}{Compressed\ File\ Size}$$

The performance comparison of five traditional wavelets for three test images is given below in table 1.

TABLE I. Performance comparison

Image	Type of wavelet used	Compression score (%)		Compression ratio	
		N=3	N=4	N=3	N=4
Lena	haar	92.27	97.95	12.94	49.0
	bior6.8	87.10	93.95	7.75	16.5
	coif5	83.09	90.27	5.91	10.2
	dmey	66.14	74.17	2.95	3.87
	sym8	87.4	94.21	7.94	17.2
Gandhiji	haar	92.27	97.9	12.94	49.2
	bior6.8	87.1	93.95	7.75	16.5
	coif5	83.09	90.27	5.91	10.2
	dmey	66.14	74.17	2.95	3.87
	sym8	87.40	94.21	7.94	17.2
Lady	haar	92.27	98.06	12.94	51.68
	bior6.8	87.10	93.95	7.75	16.54
	coif5	83.09	90.27	5.91	10.28
	dmey	66.14	74.17	2.95	3.87
	sym8	87.40	94.21	7.94	17.28

In this analysis, we have found that wavelet transform is very powerful and extremely useful for compressing data such as images. It is quite interesting to see that both compression and encryption are carried out by using wavelet transform.

Wavelet transform 'sym8' demonstrates better performance. It is observed that for a fixed decomposition level, the increase in value of threshold results in greater compression while for a fixed value of threshold, compression score/ratio decreases with increase in decomposition level. Wavelet based image encryption could be useful in a lot of commercial applications whereby large image databases can be rendered illegible to unauthorized users. We conclude that the compression ratio depends on the type of image and type of transforms because there is no filter that performs the best for all images pertaining to different applications.

References

[1] Rafael C. Gonzalez & Richard E. Woods,— Digital Image processing, 2ndEdition Pearson Education 2004.

[2] William Stallings, Cryptography and Network Security, Principles and Practice, Third edition, Pearson, 2003.

[3] K.P. Soman, K.I. Ramachandran, Insight into Wavelets from theory to practice, Second edition, PHI, 2006.

[4] S. Mallat, A Wavelet Tour of Signal Processing, (AcademicPress, 1999).

[5] Bryan Usevitch, "A Tutorial on Modern LossyWavelet Image Compression : Foundations of JPEG 2000," IEEE Signal Processing Magazine, 2001.

[6] Sachin P. Nanavati, Prasanta K. Panigrahi, "Wavelet Transform- A new mathematical microscope", (Resonance, March 2004)

[7] DONOHO D. Compressed sensing [J], IEEE Transactions on Information Theory, 2006, 52(4):1289-1306.

[8] CANDES E. Compressive sampling[A], Proceedings of the International Congress of Mathematicians[C]. Madrid, Spain, 2006, 3: 1433-1452.

[9] Jatan K. Modi, Sachin P. Nanavati, Amit S. Phadke,Prasanta K. Panigrahi, "Wavelet Transforms- Application to Data Analysis – 1",(Resonance, November 2004).

[10] S. Singh, V. Kumar. H. K. Verma ," DWT-DCT hybrid scheme for medical image compression", *Journal* of Medical Engineering & Technology, Vol 31, Issue 2 , pp.109 – 122, March 2007 .

[11] Martin M., "Applications of Multiwavelets to Image Compression," PhD Thesis, Deptartment of Electrical Engineering, Virginia Polytechnic Institute & State University, June 1999.

[12] Sudhakar R. and Jayaraman S., "Image Compression Using Multiwavelets and Wavelet Difference Reduction Algorithm," in Proceedings of the International Conference on Resource Utilization and Intelligent Systems, pp. 1-8, January 2006.

[13] Debayan Goswami, Naushad Rahman, Jayanta Biswas, Anshu Koul, Rigya Lama Tamang,Dr. A. K. Bhattacharjee,' A Discrete Wavelet Transform based Cryptographic algorithm', IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.4, April 2011.

[14] Alasdair Mcandrew, ―Digital Image processing with MatLab, Cengage learning 2004.

Authors Profile

Dr. V. U. K. Sastry is presently working as Professor in the Dept. of Computer Science and Engineering (CSE), Director (SCSI), Dean (R & D), SreeNidhi Institute of Science and Technology (SNIST), Hyderabad, India. He was Formerly Professor in IIT, Kharagpur, India and worked in IIT, Kharagpur during 1963 – 1998. He guided 12 PhDs, and published more than 80 research papers in various international journals. He received the best Engineering College Faculty Award in Computer Science and Engineering for the year 2008 from the Indian Society for Technical Education (AP Chapter) and Cognizant- Sreenidhi Best faculty award for the year 2012. His research interests are Network Security & Cryptography, Image Processing, Data Mining and Genetic Algorithms.

Mr. Ch. Samson obtained his Diploma from Govt. Polytechnic, Hyderabad in 1994, B. E. from Osmania University in 1998 and M. E from SRTM University in 2000. Presently he is pursuing Ph.D. from JNTUH, Hyderabad since 2009. He published 10 research papers in various international journals and two papers in conferences. He is currently working as Associate Professor and Associate Head in the Dept. of Information Technology (IT), SNIST since June 2005. His research interests are Image Processing, Image Cryptography and Network Security.

Ethernet Based Remote Monitoring And Control Of Temperature By Using Rabbit Processor

U. SUNEETHA
Department of Electronics
Sri Krishnadevaraya University
Anantapur, INDIA

K.TANVEER ALAM
Department of Electronics
Sri Krishnadevaraya University
Anantapur, INDIA

N.ANJU LATHA
Department of Instrumentation Sri
Krishnadevaraya University
Anantapur, INDIA

B.V.S.GOUD
Department of electronics and Instrumentation
Acharya Nagarguna University
Guntur, INDIA

B.RAMAMURTHY
Department of Instrumentation
Sri Krishnadevaraya University Anantapur, INDIA

***Abstract*— Networking is a major component of the processes and control instrumentation systems as the network's architecture solves many of the Industrial automation problems. There is a great deal of benefits in the process of industrial parameters to adopt the Ethernet control system. Hence an attempt has been made to develop an Ethernet based remote monitoring and control of temperature. In the present work the experimental result shows that remote monitoring and control system (RMACS) over the Ethernet.**

Keywords- RMACS; Control system; Ethernet.

I. Introduction

The exponential growth of Internet and computer technology enables the development of complex, hybrid systems which offers greater concern in maintenance and has more flexibility in servicing and fault finding [1, 2]. With the advanced technology industries are interested in automation by introducing remote monitoring and control system for the measurement control of industrial process parameters very precisely and accurately for the quality products. The Ethernet provides an inexpensive gateway through which to data transfer for real-time interaction of the remote monitoring and control of the parameter give many advantages.

RMACS is an effective on-line monitoring and control system and to transmit the real time data on to the terminal [3]. The RMACS is an effective also to analyze, manage and feedback the remote information and it combines the most advanced science and technology, in the field of communication technology, Internet technology and other areas.

The main objective of the present work is to develop a system for remote monitoring and control of temperature over the Ethernet. Accurate measurement and control of temperature is essential nearly in all chemical processes which require some times below than +1^{o}C or – 1^{o}C accuracy.

Temperature is measured using different methods currently in use: thermocouples (TC) [4], thermister [5, 6], resistance temperature detectors (RTD) [7] and integrated circuits (IC). Most of the temperature sensors produce an analog at present which is further convert to digital form using A/D converter interfacing with the processing like microprocessor, microcontroller and PC or their combination.

II. Hardware Implementation

The block diagram and schematic diagram of the hardware system RMACS [8] is shown in the fig.1 and fig.2 respectively. The system consists of the following units. They are Temperature sensor (LM 35), signal conditioning unit (MAX 186), Rabbit processor, Ethernet, RTL8019AS, Relay control device (FAN) and PC.

In the system Rabbit 3000 processor is used to measure and control the parameter. Input port is used to sense the temperature and output port is used to control the process. Ethernet provide communication capability to the system. Temperature bath is used to set different temperatures. Sensor is placed within the temperature bath so that its output is being monitored by the system. The signal conditioned by MAX 186, which will be able to provide the signals to be properly detected by the processor. The data logging is achieved continuously by Rabbit processor to the PC via MAX 232 (Level converter). By implement the software program developed on the PC update the database.

A. Temperature Sensor (LM35)

The LM35 series are precision integrated-circuit temperature sensor. It is three-terminal device produces an electric voltage proportional to degree Celsius (10mv/°C). These sensors are capable to measure temperature below 0°C by using a pull down resistor (±1°C from -55°C to +150°C vs. ±3°C from -20°C to +100°C). Thus LM35 has an advantage over linear temperature sensors calibrated in degree Kelvin, as the user is not required to subtract a large constant voltage from its output to obtain convenient centigrade scaling and not require any external calibration.

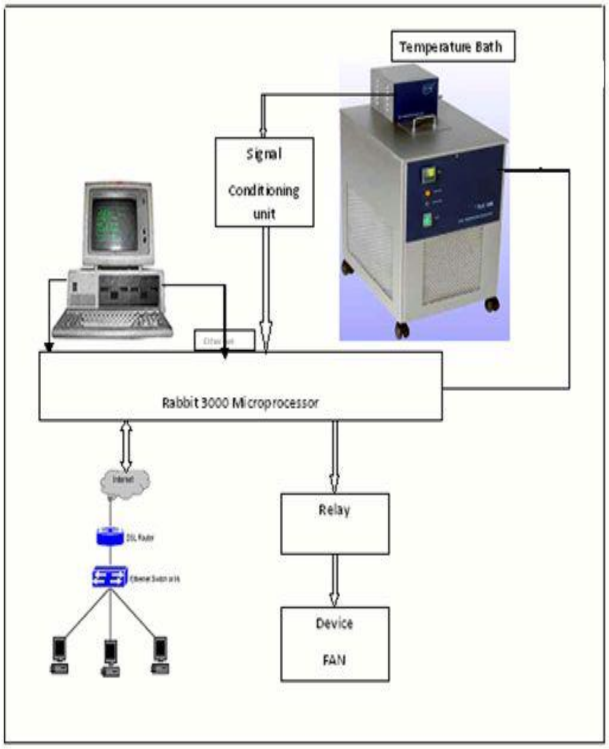


Fig.1 Block diagram of RMACS for Temperature with Ethernet

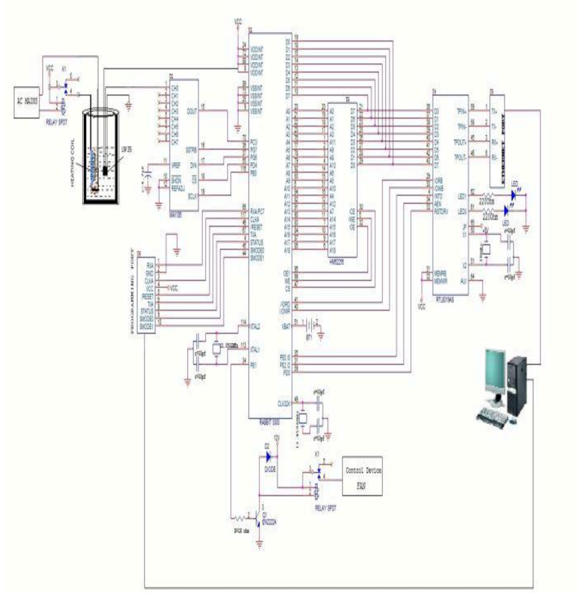


Fig.2 Schematic diagram of RAMACS for temperature with Ethernet.

B. *Control device*

In the present system a FAN is used to control the temperature precisely by switch on/off to reduce temperature of the coil by connecting one of the I/O lines of the processor.

C. *Signal conditioning unit (MAX 186)*

The signal conditioning unit consists of an ADC. The sensor output is typically connected to the ADC(MAX 186) which has 12-bit analog to digital converter combines an 8-channel multiplexer and serial interface together with high conversion speed and ultra-low power consumption. The analog inputs are software configurable for unipolar/bipolar and single-ended/ differential operation.

D. *Rabbit processor*

In the present study a 16-bit Rabbit processor is used as a processing tool for the remote measurement and control of temperature. It has 8-bit external data bus and an 8-bit internal data bus, address lines (A0–A18) and the data lines (D0–D7), the onboard 512K flash memory and 512K SRAM chips, EPROM. 1Mbyte serial flash is also available to store data on Web pages. Rabbit processor is having six serial ports [9] for asynchronous communication.

E. *Ethernet*

The word NETWORK [10] implies a linkage between two or more computing devices together for the purpose of sharing data. The network can be categorized as local area networks (LANs) and wide area networks (WAN)[11,12]. Ethernet is a family of frame based computer network technologies for local area networks (LANs) with the data transfer rate as high as 10 Mbps. The principle to access the Ethernet is carrier sense Multiple access with collision detection (CSMA/CD)[13, 14]. The Ethernet is having two layers one is physical layer which converts the data into electrical signals. Second is data link layer it is further divided into two sub-layers. One is logic link control (LLC) which is responsible for flow and error control. Next is media access control (MAC) which is responsible for the operation of CSMA/CD access method. Now a days, the most popular protocol is TCP/IP[15]. TCP protocol is said to be connection oriented and provides a reliable service.

F. *RTL8019AS*

It is highly integrated Ethernet card (NIC: network interface card) implemented with a plug and play NE 2000 compatible full duplex and power down features having built in 16 byte SRAM in a single chip Consisting PPT (point to point protocol) protocol for logic link layer.

This will carry the signals to data link layer played to the RJ-45 Ethernet jack.

G. *Personal computer*

Personal computers have high-speed Internet allowing access to the World Wide Web. In the present work the data logging is achieved continuously from the Rabbit processor to personal computer via MAX 232.

Data is received by implementing the software program developed for the present work which updates the data by using html. On the client side this html file is transferred by using hypertext transfer protocol (HTTP).

III. SOFTWARE DEVELOPMENT

The code is developed for RMACS using the Dynamic C software with a certain socket libraries and http libraries. By initializing this library functions we can develop server software. Library files with Dynamic C provide a full range of serial communication support. The temperature monitored on html document. On client side html file can be browsed with Internet browser which can display the temperature on http service. The html file exists within the http services. If the temperature variation takes place the html file is updated and http server refresh the data. In this way Remote Monitoring and Control of temperature is done by using rabbit processor. The flowcharts depicting the monitoring and the control temperature is shown in fig.3.

IV. RESULTS AND DISCUSSIONS

Main object of the present work is to develop an RMACS for the industrial process parameter like temperature on real time basis with Ethernet which is tiny, rugged, low cost and low power consumption ideally suited for industrial control applications. The graphical representation of real-time vs temperature for different set points is shown in figure 4 and photograph of the remote monitoring system with Ethernet Figure 5.

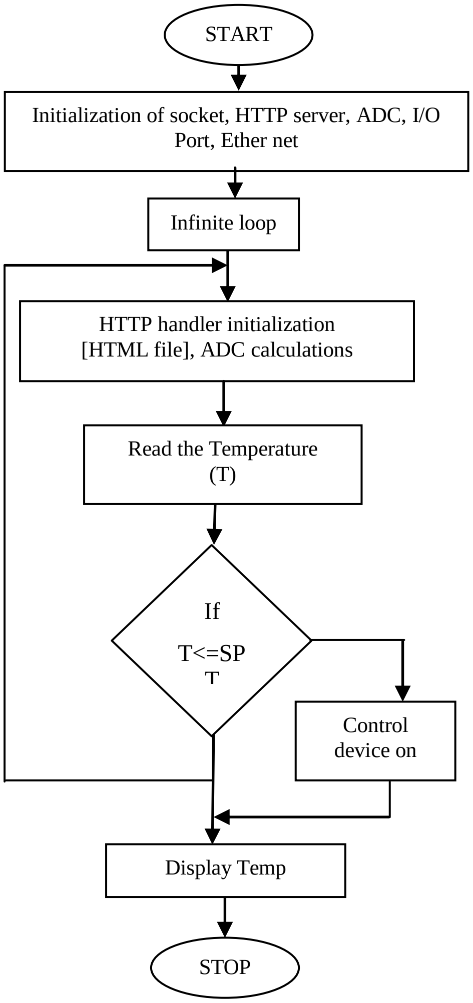

START
Initialization of socket, HTTP server, ADC, I/O Port, Ether net
Infinite loop
HTTP handler initialization [HTML file], ADC calculations
Read the Temperature (T)
If T<=SP T
Control device on
Display Temp
STOP

Fig 3**:** Flowchart for RMACS of temperature with Ethernet**.**

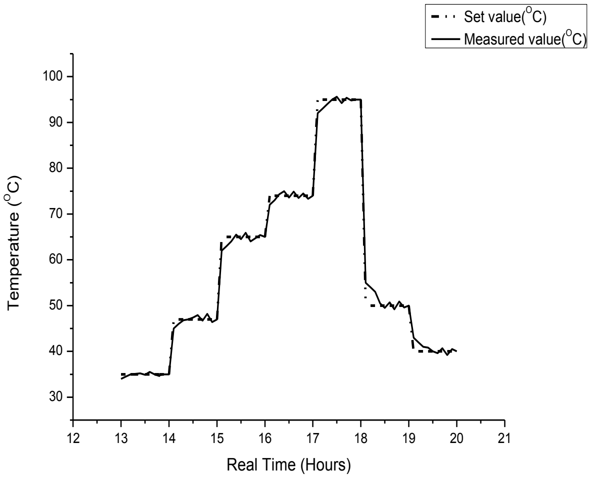


Fig4***:*** Graphical representation of RMACS for temperature with Ethernet.

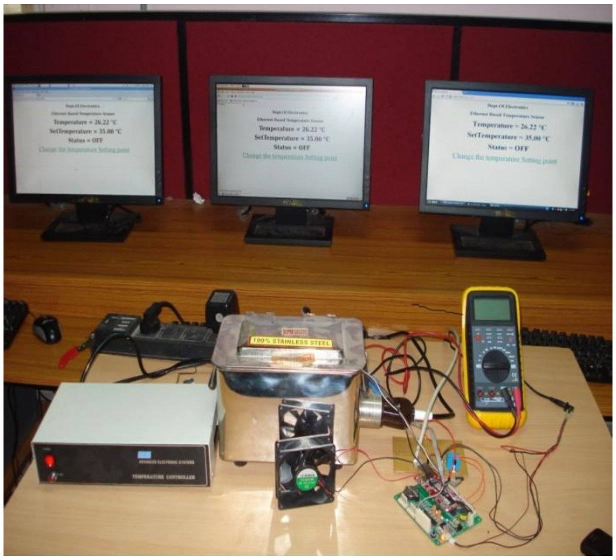


Fig.5**.** photograph of RMACS for temperature with Ethernet.

V. CONCLUSION

A low cost approach of the present developed work is novel and has achieved the target to control process parameter like temperature remotely using the Ethernet. The system is low cost as compared to the previously existing systems like GSM, Wifi with an accuracy of +or -1^{o}C.

REFERENCES

[1] P. line, “Design and implementation of an internet-based virtual lab system for learning support”, Proc.5thIEE International conf on Advanced learning Technologies, Kaohsiung, Taiwan, 005, pp.295-296.

[2] Bagnasco & A.Scapolla, ”Agrid of remote laboratory for teaching electronics”, 2^{nd} International LeGE-WG workshop on e learning and grid technologies: a fundamental challenge for Europe, Paries, 2003.

[3] N.C.Corbett, “Remote Monitoring and Control of advanced gas turbines” Computing and control Engineering Journal, April 2001

[4] Application Note AN107, “Practical Thermocouple Temperature Measurements”, Dataforth Corporation.

[5] RTD Measurement and Theory www.omega.com / temperature /z/thertd .html.

[6] N .Mondal, On Certain Topics in Temperature Measurement Using Thermistors, MEE Thesis, Jadavpur University, 1996.

[7] X.LIN, G.HUBBARD, Sensor and Electronic Biases/Errors in Air Temperature Measurements in Common Weather Station Networks, Journal of atmospheric and oceanic technology, Volume 21, 2004.

[8] Dr. B. Ramamurthy, S. Bhargavi and Dr. R. Shashi Kumar, "Design and implementation Of GSM based Remote Monitoring and Control System for Industrial Process Parameters", IJCSIS, vol.8, no.5, August 2010.

[9] Technical Note TN227**,** Interfacing External I/O with Rabbit 3000 Designs

[10] H. Ohsaki, M. Murata, and H. Miyahara,"Modeling end-to-end packet delay Dynamics Of the internet using systems identification", proceedings of the International teletraffic congress 17, Salvador da bahia, Brazil, pp.1027-1038, December 2001.

[11] Wright, Gary R and W Richard Stevens TCP/IP Illustrated, vol.2, the implementation. Reading, Wesley, 1998.

[12] BehrouzA.Forouzan TCP/IP protocol suite, vol.3, Basics of protocols.

[13] S.H.Yang, and J.L.Alty,"Development of a distributed simulator for control experiments through the internet", Future generation computer systems. Vol.18, No.4, pp.57-59, April 1999.

[14] S.H.Yang, X.Chen, and J.L.Alty, "design issues and implementation of internet based process control systems", Control Engineering practice, vol.11, No.6.pp.709 720, June 2003.

[15] Hong-Yanli, "Web-based remote monitoring and control for Process plants", International conference on mechine learning and cybernetics, 18- 21 august 2005.

AUTHORS PROFILE

U.Suneetha is doing P.hd in the Department of Electronics and communications, Sri Krishna Devaraya university, Anantapur. She also working as a Teaching assistant in the Department of Electronics and communications, Sri Krishna Devaraya university, anantapur, Andhrapradesh, India. She having nine years of teaching experience. Her area of interest are Wireless communications and embedded systems.

Prof.B.V.S.Goud is currently Head of the department to the department of electronics and instrumentation technology and co-ordinator to the PG Exames, Acharya nagarjuna university, Nagarjuna nagar, guntur. Pior to be he was worked as Deputy manager, Elico plvt ltd. He is the chairman to the board of studies in electronics and instrumentation technology. He is a lifemember of instrumentation socity of india. He has 20 years of teaching experience over 4 years of industrial experience. He has produced several reasearch and review papers in national and international journals. He has complete two major reasearch projects funded by UGC and DAE. He conducted several simphosias, seminors and workshops.

A study on Security within public transit vehicles

A.N.Seshukumar
M. Tech II Year
Department of CSE
VR Siddhartha Engineering College
Vijayawada

Dr.S.Vasavi
Professor
Department of CSE
VR Siddhartha Engineering College
Vijayawada

Dr.V.Srinivasa Rao
Professor & HOD
Department of CSE
VR Siddhartha Engineering College
Vijayawada

***Abstract*— In public transit vehicles, security is the major concern for the passengers. Surveillance Systems provide the security by providing surveillance cameras in the vehicles and a storage that maintains the data. The applications that allow monitoring the data in surveillance systems of public transit vehicles will provide different features to access the video and allow to perform number of operations like exporting video, generating snapshots at a particular time, viewing the live as well as playback videos. This paper studies automation process of video surveillance system that can also be applied in the surveillance system of public transit vehicles. A new feature that enhances the security to the passengers such as tracking of vehicle through the GPS (Global positioning system) tracking system and also capability of providing the vehicle information like acceleration, speed, on the user interface of application to the user.**

Keywords-Surveillance Cameras; Global Positioning System; Automation; public transit vehicles.

I. Introduction

Surveillance is the monitoring of the behavior, activities, or other changing information, usually of people. It usually refers to observation of individuals or groups by government organizations. The word surveillance may be applied to observation from a distance by means of electronic equipment (such as CCTV cameras), or interception of electronically transmitted information (such as Internet traffic or phone calls). It may also refer to simple, relatively no- or low-technology methods such as human intelligence agents and interception. The present surveillance system in public transit vehicles consists of surveillance Cameras mounted within the vehicle and a digital video recorder for the storage of the video data from the cameras. The surveillance system in public transit vehicles will prevent theft by providing onboard security cameras to monitor bus activity and act as preventative measure against acts of theft between riders. The unpredictable

nature of bus passengers throughout the day can many times lead to violent incidents. Such an incident could stem from an argument between riders or a passenger under the influence of alcohol or drugs losing composure. Surveillance cameras can monitor for such unsavory activity, enabling operators to alert authorities. Users of the bus system want to be confident that their mode of transportation is a safe one. Onboard video surveillance cameras give riders the assurance that authorities are doing everything in their power to provide a high level of security. Onboard security cameras can prove valuable in criminal investigations of incidents taking place on buses as well as outside crimes involving specific suspects whose images may be uncovered. Along with all the above specified benefits we propose a new feature that enhances the security is vehicle tracking during the movement of the vehicle in remote location and also providing the acceleration, speed, direction of the vehicle during the motion of it. So, that at any point when ever any unusual events happens the authorities can take immediate decision and also it reminds the driver to drive in controlled manner without crossing the limited field. This tracked information should be displayed on the user interface of the application while playing the live/playback video. We also studied the automation process of video surveillance systems in a static location like shopping malls, airports etc., The main aim of video surveillance is to develop intelligent video surveillance to replace the traditional passive video surveillance that is proving ineffective as the number of cameras exceed the capability of human operators to monitor them. The aim of visual surveillance is not only to put cameras in the place of human eyes, but also to accomplish the entire surveillance task as automatically as possible.

This paper presents related work on this automation process which can also be applicable to surveillance of public transit vehicles. Section 2 provides summary of existing approaches. Our Proposed approach is given in section 3. Conclusions and future work are given in section 4 and 5.

II. Related Work

This section presents information on Global positioning system and a study on the automation process of video surveillance systems.

The surveillance applications provide features like live/playback video monitoring, exporting the video, generating snapshots. To enhance the security of passengers a new feature namely tracking vehicle while monitoring the video can be added to the existing system. Here we present information on global positioning system.

Global Positioning System is a system that specifies the time and position of an object on the earth. Even though we have different kinds of positioning systems they have their own drawbacks like limited area, some of the positioning systems are LANDMARKS, LORAN and CELESTIAL etc. Since it is satellite based navigation system, it is made up of 27 Earth orbiting satellites among 27 only 24 will be in operation and the remaining 3 are useful when any one among 24 satellites fails. A gps receiver will take the help of satellites to find the position of an object on the earth.Here position in the sense location and time of an object. The location is represented by latitude and longitude values based on which location of a vehicle can be identified on the earth.

The process of identifying the suspicious or abnormal behavior in the video from the surveillance cameras requires an operator to identify. If there are hundreds of areas to be monitored, then it needs more number of operators to perform the analysis. Since it requires more number of operators the automation process came into picture. The same can be applicable in case of surveillance in public transit vehicles. We studied the techniques involved in the automation process.

The automation of video surveillance involves the following steps as shown in figure 1.

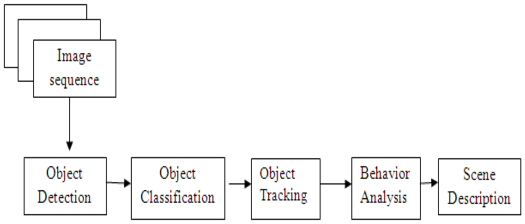


Figure 1: Automation Process of Video Surveillance

The automation process can be done with or without data mining techniques. We studied number of techniques involved in the automation process and the related work on this is presented below.

[1] Presents centralized and decentralized architecture for video surveillance systems. It also presents a typical sequence of video analysis operations in an automatic video surveillance system. It provides about each and every step in automation of video surveillance like preprocessing, Object detection/motion detection, object tracking and object analysis. It explains the algorithms used in motion detection like background subtraction and also that trained object detectors are used to detect objects of a particular category against a complex, possibly moving, background.

[2] Proposed a framework for analysis of surveillance videos by summarizing and mining of the information in the video for learning usual patterns and discovering unusual ones. This framework is useful because it is not possible for a human operator to continuously watch hours of video, either online through a webcam or offline and analyze the video from multiple perspectives. This framework forms the video data in to clusters using an incremental clustering algorithm which can be used with any data type (numerical or symbolic) and is independent of predefining the number of clusters and cluster radii. The incremental clustering algorithm helps in dealing with the large volume of data in case of offline analysis of stored videos. The two techniques component based clustering and cluster algebra for summarization as well as automatic selection of component clusters are used to discover unusual patterns in a surveillance video.

[3] Proposed a review on video surveillance systems. It presents the need of video surveillance and the entire process of video surveillance automation beginning from motion/object detection to behavior analysis. Different techniques are used for the motion segmentation such as background subtraction, temporal differencing and optical flow. Object classification distinguishes between the different objects present in the image into predefined classes such as human, vehicle, animal, clutter, etc. Two approaches namely shape based and motion based classification were used for classification. The final step of an automated surveillance system is to recognize the behavior of the objects and create a high level semantic description of their activities.

[4] Proposed an algorithm for background model initialization. Motion detection and tracking algorithms rely on the process of background subtraction, a technique which detects changes from a model of the background scene. It presents a new algorithm for the purpose of background model initialization. The algorithm takes as input a video sequence in which moving objects are present, and outputs a statistical background model describing the static parts of the scene.

[5] Proposed an approach for automated analysis of passenger’s behaviors with a set of visual low-level features, which can be extracted robustly. The approach was performed on a set of global motion features computed in different parts of the image. The complete image, the face and skin color regions, a classification with Support Vector Machines was performed.

[6] Provided a survey on behavior analysis in video surveillance applications. The different methods of behavior analysis were mentioned in the survey. The automation of video surveillance was also provided in detail manner with all the methods in each step of the process.

[7] Proposed Dynamic Oriented Graph method that is used to detect and predict abnormal behaviors, using real-time unsupervised learning. The Dynamic Oriented Graph method processes sequential data from tracked objects, signalizes unusual events and sends alarm warnings for possible abnormal behaviors. This method also constructs a structure to learn and maintain a set of observed patterns of activities, using real-time learning and without the requirement to perform any kind of training. The Dynamic Oriented Graph classifier demonstrated to be extremely fast, learning, classifying and predicting activities on-line and in a dynamical form. The classifier detect the behavior of a very large number of objects in real-time simultaneously.

[8] Proposed an approach for the automatic human behavior recognition and its explanation for video surveillance. This system could automatically report on human activity in video would be extremely useful to surveillance officers who can be overwhelmed with increasingly large volumes of data.

[9] Proposed a framework for moving objects recognition system using its appearance information. Moving objects are extracted with adaptive Gaussian mixture model first. Its silhouette image is unified to a certain mode. Subspace feature of different moving object classes is obtained through training with large numbers of these silhouette images. A more suitable dimension reduction method called marginal Fisher analysis is used to obtain projection eigenvector.

[10] Proposed a framework for Interactive Motion Analysis for Video Surveillance and Long Term Scene Monitoring**.** It consists of two feedback mechanisms which allow interactions between tracking and background subtraction. This improves tracking accuracy, particularly in the cases of slow-moving and stopped objects which is completely complement to the existing process.

[11] Proposed a method to detect passengers on-board public transport vehicles with the aim of monitoring their behaviors under suspicious circumstances. It comprises an elliptical head detection algorithm using the curvature profile of the human head as a cue.

[12] Presents real-time implementation of Moving Object Detection Video Surveillance Systems Using FPGA. FPGA is a device that captures the video stream, performs pre-processing, image analysis and reduces the data transfer between FPGA and CPU by transferring the processed results to CPU.

III. Proposed Approach

The present surveillance applications provide the monitoring of live/playback video from the surveillance cameras and allow performing some operations on the video data like exporting video, snapshot generation. But to enhance the security of the passengers we propose a new feature in the surveillance application. The new feature is tracking of vehicle while monitoring the data and also providing the details of the speed, acceleration, direction on the user interface of the application to the user. The tracking of device can be done in two ways.

1. Connecting GPS device separately along with the digital video recorder within the vehicle and receiving the latitude, longitude values from the device and mapping the location on the maps.

2. Enabling a digital video recorder with GPS Connectivity so that a separate device for GPS tracking can be eliminated.

From the two approaches, the second approach would be less expensive than the first approach. The tracking can be done by using the open source tools like Google maps to track the vehicle.

The tracking system allows the user to find the location of the vehicle at any instant of time and also allow the authorities to react to any unusual events like accidents, bus failure situation immediately. This feature should be enabled on the application side so that whenever the operator monitors the video, he can also track the location of the vehicle. This also makes the driver of the bus to drive in a limited speed.

The architecutre for the proposed approach is shown below in figure 2.The architecture explains that the digital video recorder integrated with the GPS connectivity will allow the operator to track the vehicle on the user interface while monitoring the video simultaneously.

The following figure 2 is the architecture for the proposed framework.

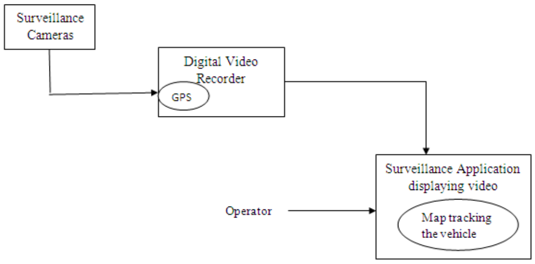


Figure 2: Architecture for enhanced security using GPS tracking system in surveillance applications of public transit vehicles.

The proposed feature can be achieved by using open source Google maps to perform tracking of vehicle. The digital video recorder that contains the gps connectivity will store the latitude and longitude values of the location based on the current position of the vehicle. The application can show the tracking of vehicle by accessing the latitude, longitude values from the device, use the Google maps and shows the exact location of the vehicle. Along with the location, if the device has been provided with the details of the acceleration, speed then the developer can show the details of the acceleration and speed on the application itself.

We also studied the automation process of surveillance systems. We found that the video content from the surveillance cameras contain unstructured enormous data that is not useful for real time processing

[13] Proposed a framework that mines the raw video content from surveillance cameras in surveillance systems of public transit vehicles.

The automation process involves object detection, object classification and object tracking and Behavior analysis. We observed that motion /object detection is being done by enabling it as a feature on digital video recorder itself instead of separate implementation.

Object Tracking gives structure to the observations and enables the object's behavior to be analyzed, for instance detecting when a particular object crosses a line. The aim is to automatically detect passengers which might be a threat to others or themselves.

Object classification is the process of identifying what kind of object is present in the environment and is useful when distinctly different types of objects are present in the environment.

Finally behavior analysis involves analysis and recognition of motion patterns, and the production of high-level description of actions and interactions between or among objects. Human face and gait are now regarded as the main biometric features that can be used for personal identification in visual surveillance systems.

The same process can be applicable to the existing surveillance system of public transit vehicles that would reduce the man power required to monitor the video.

IV. Conclusion

In this paper, the existing system of surveillance system in public transit vehicles is specified and also proposed a new feature that enhances the security of the passengers. GPS tracking system ensures more security to the passengers. The digital video recorder that is embedded with GPS connectivity can provide the proposed feature. Also the device that is provided with the acceleration and speed details then by accessing those details they can be displayed on the application to the user.

The process of surveillance system needs a human operator to monitor the video data. But a lot of research was done on the automation of surveillance process that reduces the man power required. At present this automation is being done only on applications of static locations. This paper provided the detailed study on the automation process that could be applicable in the surveillance system of public transit vehicles.

V. Future Work

The surveillance applications at present are able to provide the user to monitor the video of a single vehicle at a time in user interface. In future, it can be extended to monitor multiple vehicles simultaneously on a single user interface and also reduce the man power required to monitor the video by implementing the automation process.

The surveillance applications can be provided with the automated process of detecting the violent incidents that may take place in the vehicle.

References

[1] “An Introduction to Automatic Video Surveillance”, Andrew W. Senior, Privacy Protection in Video Surveillance, 2009.

[2] Ayesha Choudhary, Santanu Chaudhury and Subhashis Banerjee : A Framework for Analysis of Surveillance Videos in Computer Vision,Graphics& Image Processing ,2008. Sixrh Indian Conference in December 2008.

[3] Garima Sharma: Video Surveillance System: A Review in IJREAS, Volume 2 Issue2 February 2012.

[4] D. Gutchess†, M. Trajkovi´c‡, E. Cohen-Solal‡, D. Lyons‡, A. K. Jain†: A Background Model Initialization Algorithm for Video Surveillance in IEEE transactions 2001.

[5] Dejan Arsi´c, Bj¨orn Schuller and Gerhard Rigoll:Suspicious Behavior Detection in Public Transportation by Fusion of Low-Level video Descriptors in IEEE transactions,2007.

[6] TeddyKo:A Survey on Behavior Analyis in Video Surveillance Applications paper published in 2011.

[7] Duarte Duque, Henrique Santos and Paulo Cortez : Prediction of Abnormal Behaviors for Intelligent Video Surveillance Systems in IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007).

[8] Neil Robertson , Ian Reid and Michael Brady: Automatic Human Behavior Recognition and Explanation for CCTV Video Surveillance in security journal,2008.

[9] Zhang Yi, Wang Hancheng: A General Framework of Moving Objects Recognition System.

[10] A.W. Senior, Y. Tian, and M. Lu, "Interactive Motion Analysis for Video Surveillance and Long Term Scene Monitoring", in Proc. ACCV Workshops (1), 2010.

[11] Boon Chong Chee, Mihai Lazarescu and Tele Tan: Detection and Monitoring of Passengers on a Bus by Video Surveillance in IEEE transactions,2007.

[12] Kryjak, Tomasz and Gorgon, Marek (2011) Real-Time Implementation of Moving Object Detection in Video Surveillance Systems Using FPGA.

[13] JungHwan Oh, Babitha Bandi : Multimedia DataMining Framework for Raw Video Sequences in third International Work on Multimedia DataMining 2002.

Authors Profile

A.N. Seshu Kumar is pursuing Master of Technology with specialization in Computer Science and Engineering in V.R.Siddhartha Engineering College, Vijayawada. His current research interests include video surveillance applications, automation of new applications, quality control where his publications are focused.

Dr.S.Vasavi working as professor in the department of Computer Science and Engineering in V.R.Siddhartha Engineering College,Vijayawada has 16 years of teaching experience.Her areas of research are Semantic interoperability , Data mining and Image processing.

Dr.V.Srinivasa Rao working as professor & HOD in the department of Computer Science and Engineering in V.R.Siddhartha Engineering College,Vijayawada has 22 years of teaching experience.Her areas of research are Bioinformatics, Image processing.

Image Denoising using Adaptive Thresholding in Framelet Transform Domain

S.Sulochana
Research Scholar, Institute of Remote Sensing (IRS)
Anna University
Chennai, India

R.Vidhya
Associate Professor, Institute of Remote Sensing (IRS)
Anna University
Chennai, India

Abstract— **Noise will be unavoidable during image acquisition process and denosing is an essential step to improve the image quality. Image denoising involves the manipulation of the image data to produce a visually high quality image. Finding efficient image denoising methods is still valid challenge in image processing. Wavelet denoising attempts to remove the noise present in the imagery while preserving the image characteristics, regardless of its frequency content. Many of the wavelet based denoising algorithms use DWT (Discrete Wavelet Transform) in the decomposition stage which is suffering from shift variance. To overcome this, in this paper we proposed the denoising method which uses Framelet transform to decompose the image and performed shrinkage operation to eliminate the noise .The framework describes a comparative study of different thresholding techniques for image denoising in Framelet transform domain. The idea is to transform the data into the Framelet basis, example shrinkage followed by the inverse transform. In this work different shrinkage rules such as universal shrink(US),Visu shrink (VS), Minmax shrink(MS), Sure shrink(SS) , Bayes shrink(BS) and Normal shrink(NS) were incorporated . Results based on different noise such as Gausssian noise, Poission noise , Salt and pepper noise and Speckle noise at (σ=10,20) performed in this paper and peak signal to noise ratio (PSNR) and Structural similarity index measure(SSIM) as a measure of the quality of denoising was performed.**

***Keywords-** Discrete Wavelet Transform(DWT); Framelet Transform(FT); Peak signal to noise ratio(PSNR); Structural similarity index measure(SSIM).*

I. INTRODUCTION

The quality of image is degraded by various noises in its acquisition and transmission. Image Denoising has remained a fundamental problem in the field of image processing [1] . There are various noise reduction techniques used for removing noise. Most of the standard algorithms use to denoise the noisy image and perform the individual filtering process which reduces the noise level. But the image is either blurred or over smoothed due to the lose of edges. Noise reduction is used to remove the noise without losing detail contained in the images. Wavelet transform[2] has proved to be effective in noise removal and also reduce computational complexity, better noise reduction performance.

Wavelet transform may not require overlapped windows due to the localization property and wavelet filter does not correspond to time domain convolution [3][4]. Apply discrete wavelet transform (DWT) which transforms the discrete data from time domain into frequency domain. The values of the transformed data in time frequency domain [5]-[10] are called the coefficients where large coefficients correspond to the signal and small ones represent mostly noise. The denoised data is obtained by inverse transforming the suitably threshold coefficients. DWT does not provide shift invariance. Shift variance results from the use of critical sub sampling in the DWT. For this reason every second wavelet coefficient at each decomposition level is discarded. This can lead to small shifts in the input waveform causing large changes in the wavelet coefficients. Large variations in the distribution of energy at different scales introduce many visual artifacts in the denoised output.

To overcome the problem of DWT, Framelet transform which is similar to wavelets but has some differences. Framelets has two or more high frequency filter banks, which produces more subbands in Decomposition. This can achieve better time frequency localization [11] ability in image processing. There is redundancy between the Framelet subbands, which means change in coefficients of one band can be compensated by other subbands coefficients. After framelet decomposition, the coefficient in one subband has correlation with coefficients in the other subband. This means that changes on one coefficient can be compensated by its related coefficient in reconstruction stage which produces less noise in the original image.

In this paper, we combine the Framelet transform and apply it to image denoising.A tight frame filter bank[12][13] provides symmetry and has a redundancy that allows for approximate shift invariance. This leads to clear edges with effective denoising which is lacked in critically sampled discrete wavelet transform. Experimental results show that using Framelet transform, result in high peak signal to noise ratio for all denoised images. The organization of this paper is as follows. In section 2 Mathematical Representation of Framelet transform is presented. Section 3 and 4 Denoising Algorithm and different thresholding techniques explained. Section 5and 6 Evaluation criteria and experimental results were explained.

II. FRAMELET TRANSFORM

In contrast to wavelets, Framelets have one scaling function $\varphi(t)$ and two wavelet functions $\psi_1(t)$ and$\psi_2(t)$.

A set of functions $\{\psi_1, \psi_2, \dots \dots \dots . . \psi_{N-1}\}$ in a square integrable space L^2 is called a frame if there exist *A*>*0*, *B*<∞ so that, for any function $f \in L^2$

$$A\|f\|^2 \leq \sum_{i=1}^{N-1} \sum_j \sum_k \left|\langle f, \psi^i(2^j - k)\rangle\right|^2 \leq B\|f\|^2 \qquad (1)$$

Where *A* and *B* are known as frame bounds. The special case of A = *B* is known as tight frame. In a tight frame we have, for all $f \in L^2$ .In order to derive fast wavelet frame, multiresolution analysis is generally used to derive tight wavelet frames from scaling functions

Now we obtain the following spaces,

$$V_j = span_k \{\varphi(2^j t - k\} \qquad (2)$$

$$W_j = span_k \{\psi^i(2^j t - k\} \quad i = 1.2, \dots \dots N-1 \qquad (3)$$

With $V_j = V_{j-1} \cup W_{1,j-1} \cup W_{2,j-1} \cup \dots \dots . \cup W_{N-1,j-1}$ (4)

The scaling function $\varphi(t)$ and the wavelets $\psi_1(t)$ and $\psi_2(t)$ are defined through these equations by the low pass filter $h_0(n)$ and the two high pass filters $h_1(n)$ and $h_2(n)$

Let $\varphi(t) = \sqrt{2} \sum_n h_0(n) \varphi(2t - n)$ (5)

$\psi_i(t) = \sqrt{2} \sum_n h_i(n) \varphi(2t - n) \quad i = 1,2 \dots .$ (6)

Perfect Reconstruction conditions and Symmetry Conditions

The Perfect Reconstruction (PR) conditions for the three band filter bank can be obtained by the following two equations

$$\sum_{i=0}^{2} H_i(z) H_i(z^{-1}) = 2 \qquad (7)$$

$$\sum_{i=0}^{2} H_i(-z) H_i(z^{-1}) = 0 \qquad (8)$$

A wavelet tight frame with only two symmetric or anti symmetric wavelets is generally impossible to obtain with a compactly supported symmetric scaling function (t) . Therefore if $h_0(n)$ is symmetric compactly supported. Then antisymmetric solution $h_1(n)$ and $h_2(n)$ exists if and only if all the roots of

$2 - H_0(z) H_0(z^{-1}) + H_0(-z) H_0(-z^{-1})$ has even multiplicity.

case$H_2(z) = H_2(-z)$: The goal is to design a set of three filters that satisfy the PR conditions in which the low pass filter $h_0(n)$ is symmetric and the filters $h_1(n)$ and $h_2(n)$ are either symmetric or anti symmetric. There are two cases. Case I denotes the case where $h_1(n)$ is symmetric and $h_2(n)$ is anti symmetric. Case II denotes the case where $h_1(n)$ and $h_2(n)$ are both anti symmetric. The symmetric condition for $h_0(n)$ is

$$h_0(n) = h_0(N - 1 - n) \qquad (9)$$

Where N is the length of the filter $h_0(n)$ .We dealt with case I of even length filters. Solutions for Case I can be obtained from solutions where $h_2(n)$ time reversed version of is $h_1(n)$ and where neither filter is anti symmetric. To show this suppose that $h_0(n)$, $h_1(n)$ and $h_2(n)$ satisfy the PR conditions and that

$$h_2(n) = h_1(N - 1 - n) \qquad (10)$$

Then by defining

$$h_1^{new} = \frac{1}{\sqrt{2}}(h_1(n) + h_2(n - 2d)) \qquad (11)$$

$$h_2^{new} = \frac{1}{\sqrt{2}}\left(h_1(n) - h_2(n - 2d)\right) \qquad with\ d\epsilon z \qquad (12)$$

The filters h_0 h_1^{new}, h_2^{new} also satisfy the PR conditions, and h_1^{new} and h_2^{new} are symmetric and symmetric as follows

$$h_1^{new}(n) = h_1^{new}(N_2 - 1 - n) \qquad (13)$$

$$h_2^{new}(n) = -h_2^{new}(N_2 - 1 - n) \qquad (14)$$

Where $N_2 = N + 2d$

The polyphase components of the filters $h_0(n)$, $h_1(n)$ and$h_2(n)$ are given in [13] with symmetries in Equ(9) And Equ (10) satisfies the PR conditions . The 2D extension of filter bank is illustrated on “Fig 1”.

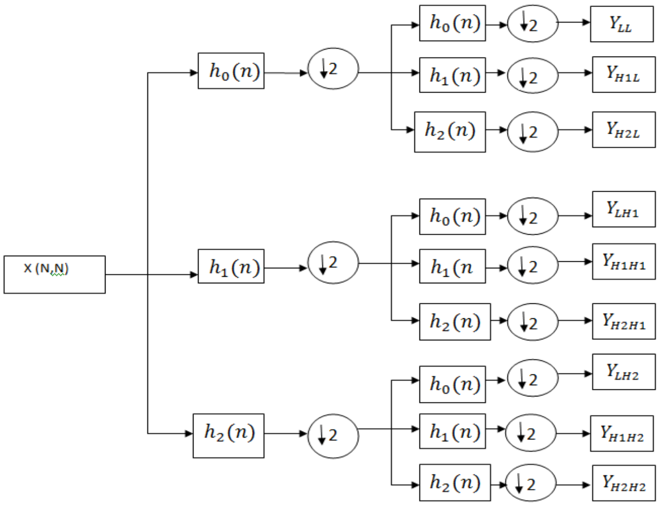


Fig.1. An over sampled filter bank for 2D image

III. Image Denoising Algorithm

Noise is present in an image either additive or multiplicative form. Gaussian noise is most commonly known as additive white Gaussian noise which is evenly distributed over the signal. Each pixel in the noisy image is the sum of the true pixel value and random Gaussian distributed noise vale. Salt and pepper noise is represented as black and white dots in the images. This is caused due to errors in data transmission. Speckle noise is a multiplicative noise which occurs all coherent imaging systems like laser and Synthetic Aperture Radar imagery.

Additive noise satisfies

$w(x, y) = s(x, y) + n(x, y)$

Multiplicative noise follows the rule

$w(x, y) = s(x, y) \times n(x, y)$

Where $w(x, y)$ is the original image $n(x, y)$ denoted noise and $n(x, y)$ reprsents pixel location in the image. The image is corrupted when noise is introduced in the images. Depending on the specific sensor there are different types of noises.

The goal is to estimate the image $s(x, y)$ from noisy observations (x, y) . The image denoising algorithm has the following steps.

1. Perform Decomposition using discrete wavelet transform (DWT) and Framelet transform (FTT).

2. Calculate threshold value of detailed parts using shrinkage rules.

3. Apply soft thresholding to the noisy coefficients.

4. Invert the decompositions to reconstruct the denoised image.

IV. Thresholding

Thresholding is a simple non-linear technique, which operates on one wavelet coefficient at a time. Each coefficient is threshloded by comparing against threshold which is calculated using shrinkage techniques. If the coefficient is smaller than the threshold it is set to zero otherwise it is modified. Replacing the small noisy coefficients by zero and inverse transform on the result provide reconstruction with the essential image characteristics without noise with mean squared error (MSE) is minimum.

There are two primary thresholding methods: hard thresholding and soft thresholding [15]. Hard thresholding and soft thresholding are denoted as following

The hard thresholding operator is defined as

$$D(U,\lambda) \quad = U \quad if|U| > \lambda$$
$$D(U,\lambda) = \quad 0 \quad otherwise$$

The soft thresholding operator is defined as

$$D(U,\lambda) = (sgn(U) * \max(0, |U| \geq \lambda))$$

The soft-thresholding rule is chosen over hard thresholding. Hard thresholding is found to introduce artifacts in the recovered images. But soft thresholding [14] is most efficient and it is also found to yield visually more pleasing images. Different shrinkage rules used in this framework are described below.

V. Shrinkage Rules

The choice of a threshold is an important point of interest. It plays a major role in the removal of noise in images because denoising most frequently produces smoothed images, reducing the sharpness of the image. Care should be taken so as to preserve the edges of the denoised image. There exist various methods for wavelet thresholding, which rely on the choice of a threshold value. Some typically used methods for image noise removal as follows.

A. Universal Shrink (US)

The universal threshold can be defined

$$\lambda = \sigma\sqrt{2log(N)}$$

N being the signal length, σ is noise variance. Universal threshold give a better estimate for the soft threshold if the number of samples is large. It tends to over smooth the signal, thereby losing some details of the original signal, which result in an increased estimation error.

B. Visushrink (VS)

Visu shrink is thresholding by applying Universal threshold proposed by Donoho and Johnston [1994].

The threshold is given by

$$\lambda = \sigma\sqrt{2log(M)}$$

Where M is the number of pixels in the image. VisuShrink does not deal with minimizing the mean squared error. Another disadvantage is that it cannot remove speckle noise. It can only deal with an additive noise. For the de-noising purpose this method is found to give up a smoothed estimate.

C. Minimax Shrink (MS)

The threshold value is calculated using minmax principle. The minimax estimator is the one that realizes the minimum of the maximum MSE obtained for the cost function. The minimax threshold is computed by

$$\lambda = 0.394 + 0.264\log(M)$$

It has the advantage of giving good predictive performance.

D. Sure Shrink (SS)

Sure Shrink is a thresholding by applying sub band adaptive threshold, a separate threshold is computed for each detail sub band based upon SURE (Stein's unbiased estimator for risk), It is a combination of the universal threshold and the SURE threshold. The sure threshold is define as

$$\lambda = \min(t, \sigma\sqrt{2log(M)}$$

Where M is number of wavelet coefficients in the particular subbands. *t* denotes the value that minimizes Stein's Unbiased Risk Estimator.σ is noise variance. n is the size of the image. SURE shrink has yielded good image denoising performance and comes close to the true minimum MSE of the optimal soft-threshold estimator.

E. Bayes Shrink

BayesShrink is an adaptive data-driven threshold for image de-noising via wavelet soft-thresholding. The threshold is driven in a Bayesian framework, and we assume generalized Gaussian distribution for the wavelet coefficients in each detail sub band and try to find the threshold which minimizes the Bayesian Risk.

Bayes shrink is calculated as follows

$$\sigma_y^2 = \sigma_x^2 + \sigma^2$$

σ_y^2 Variance of noisy image
σ_x^2 Variance of original image
σ^2 Noise variance

$$\sigma_y^2 = \frac{1}{M}\sum_{m=1}^{M} B_m^2$$

Where B_m are the coefficients of wavelet in every scale,

M is the total number of subband coefficients.

$$\sigma_x^2 = \sqrt{\max(\sigma_y^2 - \sigma^2, 0)}$$

Where $\sigma = \frac{median(|d_{ij}|)}{0.6745}$

$$\lambda = \frac{\sigma^2}{\sigma_x}$$

The reconstruction using Bayes Shrink is Smoother and more visually appealing than one obtained using Sure Shrink.

F. Normal Shrink (NS)

Normal shrink method is computationally more efficient and adaptive because the parameters required for estimating the threshold depends on subband data. The threshold value is calculated as

$$\lambda = \beta \frac{\sigma^2}{\sigma_y}$$

Where β is scale parameter.

$$\beta = \sqrt{log\left[\frac{L_k}{J}\right]}$$

L_k is the length of the subband at kth scale and J is total number of decompositions. Performance of Normal shrink is similar to Bayes shrink. But normal shrink preserves edges better than Bayes shrink.

VI. EVALUATION CRITERIA

Image Quality [19][20] is a characteristic of an image that measures the perceived image degradation Peak signal to noise ratio (PSNR) and structural similarity index measure (SSIM) were used to measure the efficiency of the proposed method.

A - Perfect image, B - Denoised image, *i* – pixel row index, *j* – pixel column index

Mean squared error

This parameter carries the most significance as far as noise suppression is concerned

$$MSE = \frac{1}{mn}\sum_{1=1}^{m}\sum_{j=1}^{n}(A(i,j) - B(i,j))^2$$

Peak Signal to Noise Ratio

$$PSNR = 10 \times log_{10}\left(\frac{Peak^2}{MSE}\right)$$

PSNR is the peak signal to noise ratio in decibles(DB).The PSNR is measured in terms of bits per sample or bits per pixel.The image with 8 bits per pixel contains from 0 to 255. The greater PSNR value is, the better the image quality and noise suppression.

The **structural similarit index measure** (SSIM)

The structural similarity index is a method for measuring the similarity between two images .The SSIM index is a full reference metric, measuring of image quality based on an initial noise free image as reference. SSIM is designed to improve on traditional methods like peak signal to noise ratio.

$$SSIM(A,B) = \frac{(2\mu_A\mu_{B+}C_1)(2\sigma_{AB} + C_2)}{(\mu_A^2 + \mu_B^2 + C_1)(\sigma_A^2 + \sigma_B^2 + C_2)}$$

Where μ_A and $\mu_B$are the estimated mean intensity along A,B directions and σ_A and σ_B are the standard deviation respectively. σ_{AB} Can be estimated

$$\sigma_{AB} = \left(\frac{1}{N-1}\sum_{i=1}^{N}(A_i - \mu_A)(B_i - \mu_B)\right)$$

C_1 and C_2 are constants and the values are given as

$$C_1 = (K_1L)^2$$

$$C_2 = (K_2L)^2$$

Where K_1 , K_2 <<1 is a small constant and L is the dynamic range of the pixel values (255).The resultant SSIM index is a decimal value between -1 and 1, and value 1 is only reachable in the case of two identical sets of images.

VII. EXPERIMENTAL RESULTS

The experiments were conducted on gray scale test image LENA of size 512x512 at different noise levels (**σ**=10,20) combined with various thresholding such as Universal shrink(US),Visu shrink(VS),Minmax shrink(MS),Sure shrink(SS) ,Bayes shrink(BS),and Normal shrink(NS).The proposed method is compared with Discrete wavlet transform(DWT)[16][17][18] based image denoising.In this experiment ,we choose PSNR and SSIM as evaluated standard.The greater PSNR and SSIM value shows that our proposed method gives better noise suppression without artifacts. PSNR and SSIM values of test image LENA with DWT and FT shown in Table 1&2.

VIII. CONCLUSION

In this work image denoising scheme based on Framelet transform was implemented using MATLAB platform.Various shrinkage rules combined with soft thresholding function were applied to the test image at noise levels (**σ**=10,20) with Gaussian noise,possion noise ,Salt &pepper Noise and speckle noise. Experimental results shows that the Framelet transform offers superior performance then discrete wavelet transform (DWT) based denoising techniques both visually and in terms of PSNR&SSIM.

REFFERENCES

[1] Rafeal.C.Gonzalez and Richard E.Woods, "Digital image processing", 2'"t,ed. Addison Wesley Longman, 1999.

[2] K.P.Soman and K.I.Ramachandran "Insight into wavelets from theory to practice", Prentice Hall, 2004

[3] I.Daubechies "The Wavelet transform, time frequency localization and signal analysis, IEEE Trans. Inform. Theory Vo1.36, pp.961-100S, Sep. 1990.

[4] M.Vetterii and C.Harley "Wavelets and filter banks: theory and design", IEEE Trans. Signal Processing Vo1.40, No.9, pp.2207-2232,Dec.l993.

[5] Lakhwinder Kaur and Savita Gupta and R.C.Chauhan, “Image denoising using wavelet thresholding”, ICVGIP, Proceeding of the Third Indian Conference On Computer Vision, Graphics & Image Processing, Ahmdabad, India Dec. 16-18, 2002.

[6] S.Kother Mohideen, Dr. S. Arumuga Perumal, Dr. M.Mohamed Sathik. “Image De-noising using Discrete Wavelet transform”, IJCSNS International Journal of Computer Science and Network Security, vol .8, no.1, January 2008.

[7] S. Grace Chang, Bin Yu and M. Vattereli, “Adaptive Wavelet Thresholding for Image Denoising and Compression”, IEEE Trans.Image Processing, vol. 9, pp. 1532-1546, Sept. 2000.

[8] G. Y. Chen and T. D. Bui, “Multi-wavelet De-noising using Neighboring Coefficients,” IEEE Signal Processing Letters, vol.10,no.7, pp.211-214, 2003.

[9] Rohit Sihag, Rohit Sihag, Varun Setia, “Wavelet Thresholding for Image De-noising”. International Conference on VLSI, Communication & Instrumentation (ICVCI) 2011,Proceedings published by International Journal of Computer Applications (IJCA).

[10] Mr. Sachin Ruikar, Dr. DDDoye, “Image Denoising Using Wavelet Transform”. 201O International Conference on Mechanical and Electrical Technology (ICMET 2010).

[11] Hadeel N.Al-Taai, “A Novel Fast Computing Method for Framelet Coefficients,” American Journal of Applied Sciences 1522-1527, 2008.

[12] Ivan W. Selesnick,”A Higher Density Discrete Wavelet Transform”. IEEE Transactions on signal procesing, Vol. 54, NO. 8, August 2006.

[13] A. Farras Abdelnour, Ivan W. Selesnick ,”Symmetric Nearly Shift-InvariantTight Frame Wavelets”. IEEE Transactions on signal processing, Vol. 53, No. 1, January 2005

[14] I.W.Selesnick and A.F.Abdelnour, “Symmetric Wavelet Tight Frame with two generators”, Applied and Computational Harmonic Analysis, Vol.17, pp.211-225, 2004.

[15] David L.Donoho, "Denoising by soft thresholding",IEEE Trans. Information Theory Vo1.41, No.3, pp. 613-627, May 1995.

[16] Tongzhou Zhao, Yanli Wang, Ying Ren, Yalan Liao, “Approach of Image Denoising Based on Discrete Multi-wavelet Transform”. 2009 IEEE.

[17] Harnani Hassan, Harnani Hassan, **“**Still Image Denoising Based on DiscreteWavelet Transform”. 2011 IEEE International Conference on System Engineering and Technology (ICSET)

[18] ngyu Yang, Yao Wang, Wenli Xu, Qionghai Dai, “Image and Video Denoising Using AdaptiveDual-Tree Discrete Wavelet Packets” IEEE Transaction on circuits and systems for video technology , V0l. 19, No. 5, May 2009

[19] Venkata Rao D, Sudhakar N , Ravindra Babu B, Pratap Reddy L ,” An Image Quality Assessment Technique Based on Visual Regions of Interest Weighted Structural Similarity”. GVIP Journal, Volume 6, Issue 2, September, 2006

[20] Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactionson Image Processing, vol. 13, no. 4, pp. 600-612, Apr. 2004.

About The Authors

Dr.R.Vidhya is currently working as associate professor in Institute of Remote Sensing (IRS) under Civil Department, Anna University Chennai. She has more than 20 years of Experience in teaching. She is an Expert in Remote Sensing, Image Processing and Water Resource. She has produced 2 MS and Guiding 12 PhD scholars. She has published more than 25 technical papers in national and international journals and conference

S.Sulochana is currently doing PhD in Institute of Remote Sensing (IRS), Anna University, and Chennai.

TABLE.1 PSNR&SSIM of LENA IMAGE USING DWT

Noise Type	***Noise Level***	***Universal Shrink (US)***	***Visu Shrink (VS)***	***Min-Max Shrink (MS)***	***Sure Shrink (SS)***	***Bayes Shrink (BS)***	***Normal Shrink (NS)***
Gaussian noise	**σ**=10	29.685	30.632	28.678	29.856	27.632	30.123
		0.6523	**0.7212**	**0.6932**	**0.5963**	**0.7550**	**0.7189**
	σ=20	27.632	28.952	28.152	27.960	28.123	27.650
		0.7453	**0.8296**	**0.7903**	**0.8523**	**0.7012**	**0.6977**
Poission Noise	**σ** =10	26.532	25.960	26.352	26.506	26.921	25.262
		0.7620	**0.8632**	**0.7291**	**0.8310**	**0.7320**	**0.6352**
	σ =20	24.345	24.320	25.312	24.202	25.156	24.260
		0.7960	*0.8110*	*0.8001*	*0.7350*	*0.7513*	*0.7125*
Salt& Pepper noise	**σ** =10	28.350	27.220	26.650	24.620	23.250	23.960
		0.5450	***0.6325***	***0.7315***	***0.6320***	***0.7255***	***0.6150***
	σ =20	27.601	26.220	25.998	24.890	22.141	23.927
		0.6460	***0.7125***	***0.8123***	***0.7013***	***0.7556***	***0.7540***
Speckle noise	**σ** =10	27.620	26.652	25.567	26.789	25.127	24.996
		0.5675	***0.5963***	***0.7321***	***0.7502***	***0.8123***	***0.8423***
	σ =20	25.239	24.976	25.134	24.936	24.456	25.789
		0.7962	***0.8532***	***0.8532***	***0.8334***	***0.8632***	***0.8706***

TABLE.2 PSNR&SSIM of LENA IMAGE USING FT

Noise Type	**Noise Level**	**Universal Shrink (US)**	**Visu Shrink (VS)**	**Min-Max Shrink (MS)**	**Sure Shrink (SS)**	**Bayes Shrink (BS)**	**Normal Shrink (NS)**
Gaussian noise	**σ**=10	33.654	32.926	33.786	32.110	31.356	31.330
		0.8130	***0.8650***	***0.8023***	***0.8561***	***0.8534***	***0.8943***
	σ=20	32.650	30.332	31.800	30.550	32.113	31.556
		0.8250	***0.8761***	***0.8134***	***0.8672***	***0.8725***	***0.8790***
Poission Noise	**σ** =10	30.325	28.663	26.336	27.333	29.332	29.300
		0.8741	***0.8650***	***0.9123***	***0.8932***	***0.8790***	***0.8870***
	σ =20	28.452	26.665	28.320	26.500	26.110	28.215
		0.9250	***0.8760***	***0.9250***	***0.9320***	***0.8875***	***0.8960***
Salt& Pepper noise	**σ** =10	32.745	30.512	32.630	33.118	30.110	32.127
		0.9614	***0.8960***	***0.9320***	***0.9415***	***0.8964***	***0.9153***
	σ =20	31.342	30.132	31.651	32.160	29.115	30.632
		0.8969	***0.8967***	***0.9324***	***0.9143***	***0.8976***	***0.9220***
Speckle noise	**σ** =10	31.896	30.360	30.112	31.660	32.632	31.650
		0.9065	***0.9156***	***0.9215***	***0.9618***	***0.9715***	***0.9867***
	σ =20	30.620	29.750	31.632	30.655	30.830	32.632
		0.9120	***0.9240***	***0.9354***	***0.9645***	***0.8964***	***0.9743***

Improved Accuracy of PSO and DE using Normalization: an Application to Stock Price Prediction

Savinderjit Kaur
Department of Information Technology, UIET, PU,
Chandigarh, India

Veenu Mangat
Department of Information Technology, UIET, PU,
Chandigarh, India

Abstract— **Data Mining is being actively applied to stock market since 1980s. It has been used to predict stock prices, stock indexes, for portfolio management, trend detection and for developing recommender systems. The various algorithms which have been used for the same include ANN, SVM, ARIMA, GARCH etc. Different hybrid models have been developed by combining these algorithms with other algorithms like roughest, fuzzy logic, GA, PSO, DE, ACO etc. to improve the efficiency. This paper proposes DE-SVM model (Differential Evolution-Support vector Machine) for stock price prediction. DE has been used to select best free parameters combination for SVM to improve results. The paper also compares the results of prediction with the outputs of SVM alone and PSO-SVM model (Particle Swarm Optimization). The effect of normalization of data on the accuracy of prediction has also been studied.**

Keywords- Differential evolution; Parameter optimization; Stock price prediction; Support vector Machines; Normalization.

I. Introduction

Stock Market prediction is an attractive field for research due to its commercial applications and the attractive benefits it offers. It follows stochastic, non-parametric and nonlinear behavior. An important hypothesis related to stock market which has been debated and researched time and again is EMH (Efficient Market Hypothesis). According to EMH, the stock market immediately reflects all of the information available publicly. But in reality, the stock market is not that efficient, so the prediction of stock market is possible.

This paper proposes a hybrid of DE-SVM (Differential Evolution-Support Vector Machines). The performance of SVM is based on the selection of free parameters C (cost penalty), ϵ (insensitive-loss function) and γ (kernel parameter). DE will be used to find the best parameter combination for SVM. DE-SVM has already been used by Zhonghai Chen et al. [6] for air conditioning load prediction, Yong Sun et al. [7] for gas load prediction, Jośe Garćıa-Nieto et al. [8] for feature selection, Shu Jun et al. [9] for rainstorm forecasting and for studying the lithology identification method from well logs by Jiang An-nan et al. [10]. The paper also compares the results of DE-SVM with PSO-SVM and SVM. The effect of normalization on datasets has also been studied.

II. Literature Review

Yohanes et al. [1] showed that ARIMA (Autoregressive Integrated Moving Average) can be outperformed by ANN. ESS (Each sum square) result with ARIMA is 284.95 and with ANN is 170.40 [1]. Qiang Ye et al. [2] proved that stock price prediction results using amnestic NN are better than common ANN. The ratio of right classified stocks is 58.25% when forgetting coefficient is 0.10 as compared to 56.25% for forgetting coefficient of 0.00 (for common ANN) [2]. Ling-Feng Hsieh et al. [3] integrated DOE (Design of Experiment) with BPNN to show that experimental validation of the optimal parameter settings can effectively improve the forecasting rate to 84%. Mustafa E. Abdual-Salam et al. [4] proved that DE converges to global minimum faster and gives better accuracy than PSO when used as training algorithms for ANN. Zhang Da-yong et al. [5] proposed a hybrid model ARMA-SVM (Autoregressive Moving average-SVM) which has MSE of 1.1433 against 1.1494 for BPNN.

A. Support Vector Machines (SVM):

SVM was developed by Vapnik and Cortes in 1995. SVM is a promising method for the classification of both linear and nonlinear data [11]. SVM can be used both for classification and regression. SVMs can be trained with lesser input samples and are less prone to overfitting. The training time of even the fastest SVMs can be extremely slow, but they are highly accurate, owing to their ability to model complex nonlinear decision boundaries [11]. SVM follows supervised learning. For classification purposes, when data is linearly separable a straight line can be drawn to separate the tuples of one class from the other. For nonlinear data, the data is mapped into higher dimensional space where the different classes can be separated using a hyperplane. A number of hyperplanes are possible but SVM searches for the maximum marginal hyperplane (MMH). The vectors in the training set that have minimal distance to the maximum margin hyperplane are called support vectors [12].

SVM selects the minority of observations (support vectors) to represent the majority of the rest of the observations [13]. The soft margins were introduced to penalize but not prohibit classification errors while finding the maximum margin hyperplane [11].

If the margin can be significantly increased, the better generalization can outweigh the penalty for a classification error on the training set [11]. To maximize the prediction ability of a model, both underfitting and overfitting need to be depressed at the same time in data processing [25]. The error of training is called Empirical Risk denoted by $R_{emp.}$ SVM uses SRM (Structural Risk Minimization) instead of ERM (Empirical Risk Minimization) which aims at minimizing (1) :

$$\min \left[R_{emp+} \sqrt{\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln\left(\frac{\eta}{4}\right)}{4}} \right] \quad (1)$$

Here, l is number of samples in training set, 1-η is the probability of the equation ((1) ≥ R_{pred} , R_{pred} is the total risk of prediction) to be true and h is VC dimension to depress overfitting in data processing [25].

SVM parameters: The performance of SVM is based on three basic parameters C (cost penalty), ϵ (insensitive loss function parameter) and γ (kernel parameter).

Cost penalty: C determines the trade-off cost between minimizing the training error and minimizing the model's complexity [26]. The parameter C determines the trade-off between model complexity and the tolerance degree of deviations larger than ε [20].

ϵ loss-insensitive function: Parameter ϵ controls the width of the ϵ-insensitive zone, used to fit the training data [27]. Larger ϵ-value result in fewer SVs selected, and result in more 'flat'(less complex) regression estimates [20]. If the value of ϵ is too big, the separating error is high, the number of support vectors is small, and vice versa [26].

Kernel parameter: γ ($2\sigma^2$) of the kernel function implicitly defines the nonlinear mapping from input space to some high-dimensional feature space [28]. The main kernels used are:

1) Linear kernel: x.y

2) Polynomial kernel: $K(x_i,x_j)= (x_i.x_j+1)^d$

3) Radial Basis kernel: $K(x_i,x)=\exp(-\frac{\|x_i-x\|^2}{2\sigma^2})$

4) Sigmoid kernel function: $K(x_i,x_j)=\tanh(x_i.x_j+p)$

RBF kernel is mostly used for stock price prediction because only one parameter needs to be confirmed, there are less SVR training parameters constructed by it and it is easy to confirm SVR training parameters [18]. The kernel width parameter σ in RBF is appropriately selected to reflect the input range of the training/test data. For univariate problems, RBF width parameter is set to σ ~[0.1–0.5]* range(x) [20].

B. Differential Evolution (DE):

Differential evolution (DE) was introduced by Kenneth Price and Rainer Storn in 1995 for global continuous optimization problem. It has won the third place at the 1st International Contest on Evolutionary Computation [14]. DE belongs to the family of Evolutionary Algorithms (EA). DE algorithm is similar to genetic algorithms having similar operations of crossover, mutation and selection. DE can find the true global minimum regardless of the initial parameter values. DE provides fast convergence and uses fewer control parameters. DE constructs better solutions than genetic algorithms because GA relies on crossover while DE relies on mutation operation. It is a stochastic population-based search method that employs repeated cycles of recombination and selection to guide the population towards the vicinity of global optimum. DE uses a differential mutation operation based on the distribution of parent solutions in the current population, coupled with recombination with a predetermined parent to generate a trial vector (offspring) followed by a one-to-one greedy selection scheme between the trial vector and the parent [15]. Depending on the way trial vector is generated, there exist many trial vector generation strategies and consequently many DE variants. High convergence characteristics and robustness of DE have made it one of the popular techniques for real-valued parameter optimization. DE uses three parameters conventionally, they are: the population size NP, the scale factor F and the crossover probability CR/ Cr. Some conditions for these variables include: NP>4, F>0 and is a real valued constant and is often set to 0.5, CR Є (0, 1) and is often set to 0.9 [16]. Different stages in DE are:

1. Population structure : The current population, symbolized by P_c, is composed of those D-dimensional vectors $X^g_i = \{x^g_{i,1}, x^g_{i,2}, \ldots, x^g_{i,D}\}$, the index g indicates the generation to which a vector belongs [17]. In addition, each vector is assigned a population index, i, which varies from 1 to N_p, knowing that N_p is the population size [17]. Once initialized, DE mutates randomly chosen vectors to produce an intermediary population P_v of Np mutant vectors V_i^g[22]. Each vector in the current population is then recombined with a mutant to produce a trial population P_u of N_p trial vectors U_i^g[22].

2. Initialization : This stage consists in forming the initial population. For example, if our objective is the optimization of the membership functions, the initialization step consists in arbitrarily choosing the interval of this function [17].

3. Mutation [17, 22]: For each vector (for example, a vector which represents the interval of the membership functions) $V_i^g=\{v^g_{i,1}, v^g_{i,2},\ldots., v^g_{i,D}\}$ a mutant vector is produced according to the following formulation [22]:

$$v^g_{i,j}=x^g_{r0,j} + F(x^g_{r1,j} - x^g_{r2,j}) \quad (2)$$

The scale factor F is a positive real number that controls the rate at which the population evolves. While there is no upper limit on F, effective values seldom are greater than 1.

4. Crossing [17,22,4]:The relative vector is mixed with the transferred vector to produce a test vector $T^g_{i,j}$:

$$T^g_{i,j}=\begin{cases} v^g_{j,i} \text{ if } (r^g_{j,i} \le \text{CR or } j=j_r) & (3) \\ x^g_{j,i} \text{ otherwise} \end{cases}$$

The crossover probability CR ϵ [0,1] is a user-defined value that controls the fraction of parameter values that are copied from the mutant. To determine which source contributes, a given uniform crossover parameter compares CR to the output of a uniform random number generator $r_{j,i}^{g}$. If the random number is less than or equal to CR, the trial parameter is inherited from the mutant $v_{j,i}^{g}$; otherwise, the parameter is copied from the vector $x_{j,i}^{g}$. In addition, the trial parameter, with randomly chosen index j_r is taken from the mutant to ensure that the trial vector does not duplicate x_{i}^{g}. Because of this additional demand, CR only approximates the true probability.

5. Selection [17]: All the solutions in the population have the same chance that the parents of being selected, regardless of their fitness function value. The child produced (new vector) after the crossing operations is evaluated. Then, the performances of the child vector and its relative are compared and the best one is selected. If the relative is still better, it is maintained within the population.

Once the new population is installed, the process of mutation, recombination and selection is repeated until the optimum is located, or a prespecified termination criterion is satisfied, e.g., the number of generations reaches a preset maximum, gmax [4].

C) Particle Swarm Optimization (PSO): PSO (Particle Swarm Optimization) was proposed by James Kennedy and Russell Eberhart in 1995. It is motivated by social behavior of organisms such as bird flocking and fish schooling [29]. It can be used for nonlinear and mixed integer optimization. PSO is different from evolutionary computing, as in it flying potential solutions through hyperspace are accelerating towards "better" solutions, while in evolutionary computation schemes operate directly on potential solutions which are represented as locations in hyperspace [4]. The position of a particle is influenced by the best position visited by itself (i.e. its own experience) and the position of the best particle in its neighborhood (i.e. the experience of neighboring particles) [30]. Particle position, x_i, are adjusted using:

$$x_i(t+1)=x_i(t)+v_i(t+1) \qquad (4)$$

where the velocity component, vi, represents the step size. For the basic PSO,

$$v_{i,j}(t+1)=wv_{i,j}(t)+c_1r_{1,j}(t)(y_{i,j}(t)-x_{i,j}(t))+c_2r_{2,j}(t)(\hat{y}_j(t)-x_{i,j}(t)) \qquad (5)$$

where w is the inertia weight [31], c1 and c2 are the acceleration coefficients, $r_{1,j}$, $r_{2,j}$ ~ U(0, 1), y_i is the personal best position of particle i, and $\hat{y}_i$ is the neighborhood best position of particle i [30]. The neighborhood best position $\hat{y}_i$, of particle i depends on the neighborhood topology used [32,33].

The main steps involved in PSO are [34]:

1) Initialize a population array of particles with random positions and velocities on D dimensions in the search space.

2) For each particle, evaluate the desired optimization fitness function in D variables.

3) Compare particle's fitness evaluation with its previous best. If current value is better than previous best, then set previous best equal to the current value, and previous best position equal to the current location in D-dimensional space.

4) Identify the particle in the neighborhood with the best success so far.

5) Change the velocity and position of the particle according to (4) and (5)

6) If a criterion is met (usually a sufficiently good fitness or a maximum number of iterations) then optimal result is given out otherwise optimization continues.

DE and PSO have been used to optimize the parameters of SVM during training and then those parameters have been used to create the best possible model for prediction purposes.

III. Implementation Details

1) Dataset: The daily datasets of Honeywell International Inc. (listed on NYSE) and Apple Inc. (listed on NASDAQ), have been used for implementation purposes. The data sets are available at (http://wikiposit.org/Finance/Stocks/) and are available in csv, html, tab delimited, xml and raw formats. The reference site for this data is www.finance.google.com. Opening price, high, low, adjusted closing price and volume have been used as inputs and the closing price the following day is the output for SVM model. The training datasets have 500 records each from 6 April, 2009 to 29 March, 2011 for Honeywell and from 17 July, 2009 to 12 July, 2011 for Apple. The testing datasets have 200 records each from 30 March, 2011 to 12 Jan, 2012 for Honeywell and 13 July, 2011 to 27 April, 2012 for Apple.

The paper compares prediction results of both normalized and non-normalized datasets.

The data has been normalized as inspired by [18] to:

1) Avoid the data with large range "submerge" those with small range and balance their functions in the training to make data comparable [18].
2) To enhance training efficiency and to avoid the problem of inner product calculation when calculating kernel function [18].

The formula used for normalization is [18]:

$$x' = x_{low} + \frac{(x_{up}-x_{low})(\mathrm{x}-x_{min})}{x_{max}-x_{min}} \qquad (6)$$

Here, x is the original data, x' is the data after normalization, x_{min} is the minimum of original data, x_{max} is the maximum of original data, x_{low} is the lower bound of the data after normalization, x_{up} is the upper bound of the data after normalization. Here, we use $x_{low} = -1$ and $x_{up} = +1$.

2) Performance indicator: The performance measure used is MSE (Mean Square Error):

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{i=n}(a_i-p_i)^2 \qquad (7)$$

Here, a is the actual value, p is the predicted value, i represents the term index which ranges from 1 to n, where n represents the last term index. MSE helps to avoid NAs and negative terms in result which can arise because of normalization of data.

3) Methodology: The basic methodology for the both normalized and non- normalized approaches is same.

To find the optimal range of all three parameters C,ϵ,γ first two parameters are fixed and the other one is varied to see its effect on Training MSE, Testing MSE and number of support vectors. And then the second parameter is fixed and so on. All these collected values are considered to find the optimal range to be used for the purpose of stock price prediction. The general range of these parameters can vary over a large solution space but the optimal range differs for different applications and is also dataset dependent. Training MSE, testing MSE and number of support vectors of all three parameters are checked for overfitting and underfitting to select the optimal range.

The following points have been considered while selecting values of C and γ:

i) Selecting C: A ‘good’ value for C can be chosen equal to the range of output (response) values of training data [19]. However, such a selection of C is quite sensitive to possible outliers (in the training data) [20] so, C has been fixed using the formula suggested in [20]:

$$C=max(|y'+3\sigma_y|, |y'-3\sigma_y|) \qquad (8)$$

Here, y’ and σ_y are the mean and standard deviation of the y values of the training data. This C value coincides with prescription suggested by Mattera and Haykin (1999) when the data has no outliers, but yields better C-values when the data contains outliers [20]. Based on above formula C is calculated as 69.167.

ii) Selecting γ: RBF kernel has been used for implementation of SVM. This use is inspired from [18]. Radial basis kernel expression is as follows:

$$K(x_i,x)=\exp\left(-\frac{\|x_i-x\|^2}{2\sigma^2}\right) \qquad (9)$$

According to [20] for multivariate d-dimensional problems the RBF width parameter should be such that $\sigma^d \sim (0.1\text{-}0.5)$ so γ or $2\sigma^2$ has been selected as 0.0625.

iii) Mattera and Haykin (1999) propose to choose ϵ-value so that the percentage of SVs in the SVM regression model is around 50% of the number of samples [19]. [20] suggests that optimal generalization performances can be achieved with the number of SVs more or less than 50%. The range of values where number of SVs is from 200 to 300 has been chosen for optimization purpose in the implementation.

Dataset for Apple:

Finding range for ϵ:

i) Selecting C: The value of C has been fixed at 450.8346 using (8).

ii) Selecting γ: Value of γ is fixed at 0.0625 according to [20] as explained above.

i) Normalized dataset parameters decision making:

Finding range for ϵ: After fixing values of C and γ at 450.8346 and 0.0625 respectively, the values of different aspects for ϵ have been calculated over the range [0.01,0.30]. The results for no. of support vectors, training MSE and testing MSE are shown in Figure 1(a), 1(b) and 1(c) respectively. The favorable range for ϵ has been found as [0.033,0.052] based on required number of support vectors, decrease in training and testing MSEs.

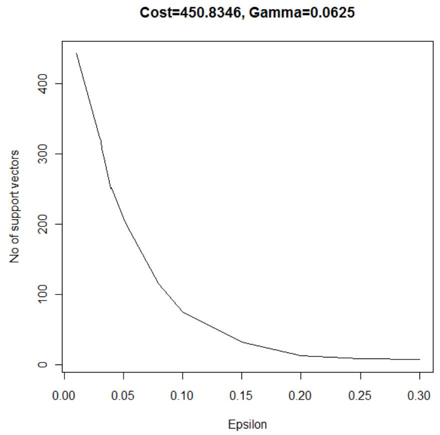


Figure 1(a)

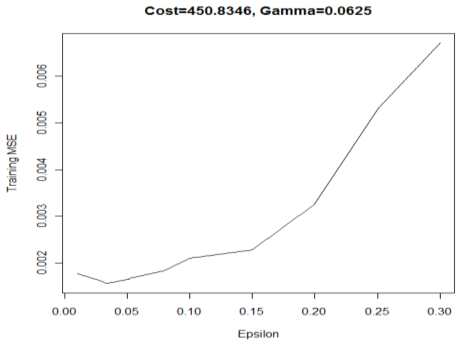


Figure 1(b)

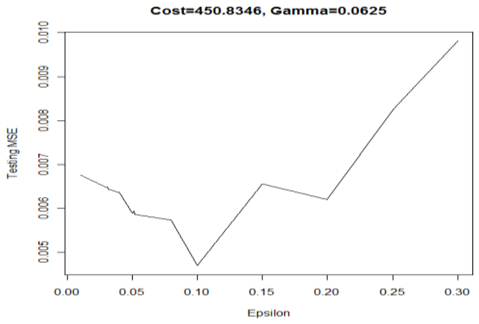


Figure 1(c)

Finding range for C: i) ϵ has been selected from above found range of [0.033,0.052]. It has been set as 0.039.
ii) γ is set as 0.0625.

The values of C are examined over [0.1,6000] while fixing ϵ and γ. The results of no of support vectors, training and testing errors are shown in Figure 2(a), 2(b) and 2(c). The range of C has been selected as [1,550]. Figure 2(a) shows that number of support vectors never fall below 200. So, C has been selected such that training MSE decreases and there is no significant increase in testing MSE.

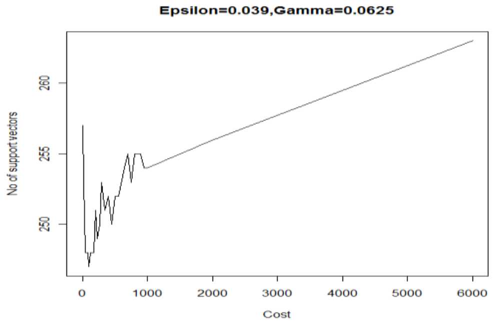


Figure 2(a)

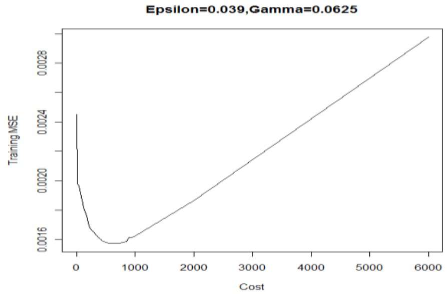


Figure 2(b)

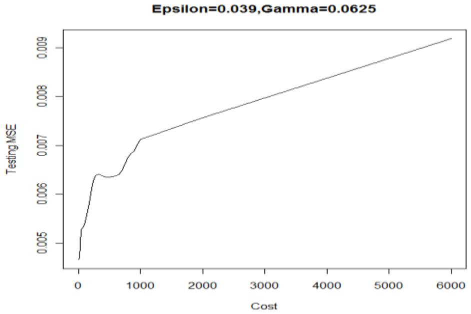


Figure 2(c)

Finding range for γ: i)) ϵ has been selected from above found range of [0.033,0.052]. It has been set as 0.039.

ii) C has been selected from above found range of [1,550]. It has been set as 500.

The values of γ are examined over [0.0,0.4] after fixing ϵ and C. The results for no of support vectors, training and testing MSEs are shown in Figure 3(a), 3(b) and 3(c). The range of γ has been selected as [0.01,0.11]. Figure 3(a) shows that number of support vectors never fall below 200. The range has been selected so that there is no significant increase in training and testing MSEs.

ii) Non-normalized dataset parameters decision making:

Finding range for ϵ: After fixing the values of C and γ at 450.8346 and 0.0625, the SVM model is created, training and testing MSEs along with no. of support vectors are recorded for ϵ over the range of [0.01,0.20].

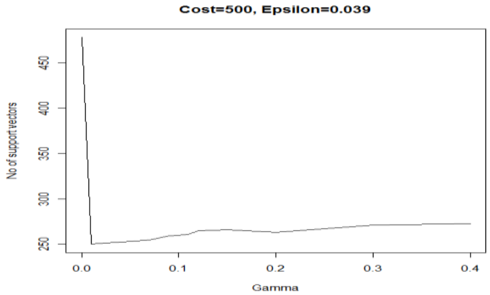


Figure 3(a)

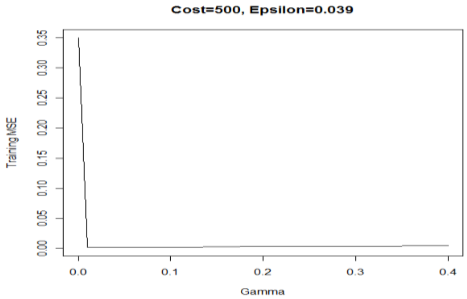


Figure 3(b)

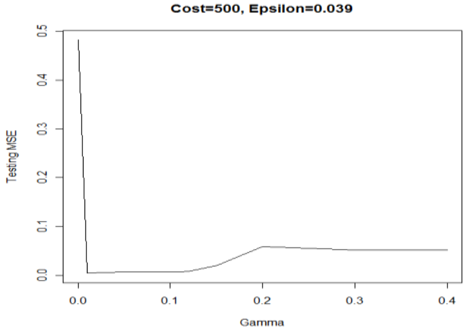


Figure 3(c)

The results are shown in Figures 4(a), 4(b) and 4(c). The range for ϵ has been selected as [0.033,0.052] after considering appropriate number of support vectors and after examining that training and testing errors don't increase significantly in this range.

iii) Finding range for C:

i) ϵ has been selected from above found range of [0.033,0.052]. It has been set as 0.035.

ii) γ is set as 0.0625.

The results for no of support vectors, training error and testing error over range of C~[1,3000] are shown in Figures 5(a), 5(b) and 5(c).

The range for C has been selected as [1,300].

Figure 5(a) shows that number of support vectors never fall below 200. The range has been selected such that training error decreases.

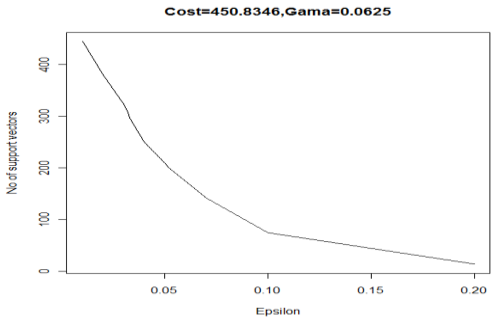


Figure 4(a)

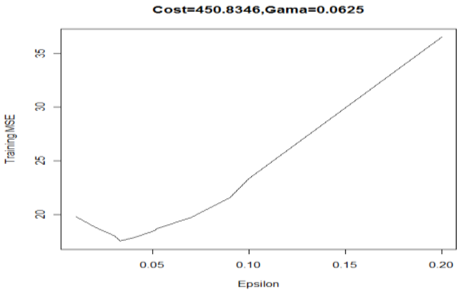


Figure 4(b)

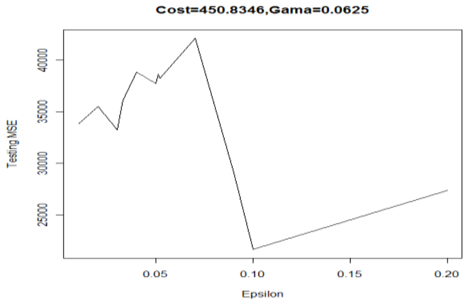


Figure 4(c)

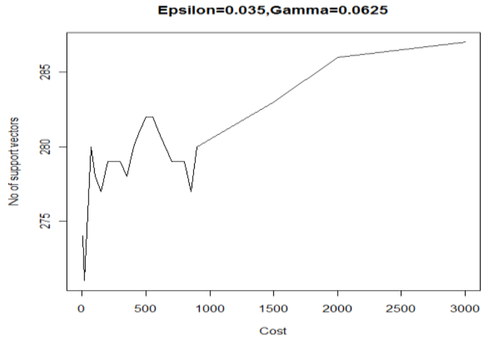


Figure 5(a)

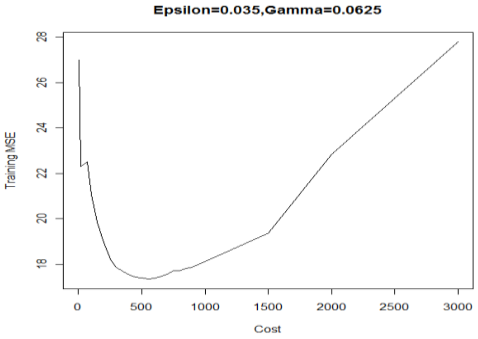


Figure 5(b)

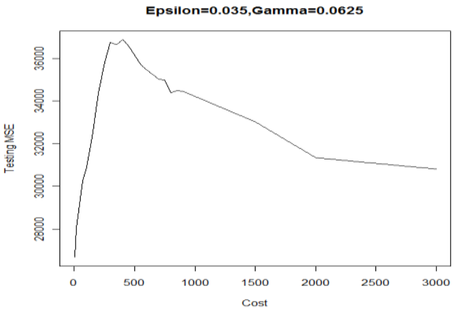


Figure 5(c)

Finding range for γ:

i)) ϵ has been selected from above found range of [0.033,0.052]. It has been set as 0.035.

ii) C has been selected from above found range of [1,300]. It has been set as 200.

γ has been examined over [0.0,0.4] and the results are shown in Figure 6(a), 6(b) and 6(c). The range of γ has been selected as [0.01,0.1]. At γ > 0.1 testing MSE decreases but training error increases.

Dataset for Honeywell: The above approach was also used with Honeywell dataset, both normalized and non normalized.

i) Normalized dataset: For ϵ, C and γ were fixed at 69.167 and 0.0625 using (8) and according to [20] which gave range as [0.08, 0.15]. Setting ϵ at 0.1 from the selected range and γ at 0.0625, C has favorable range in [1,440]. Now, ϵ at 0.1 and C at 210 from selected favorable range γ had favorable range in [0.02,0.08].

ii) Non normalized dataset: For ϵ, C and γ were fixed at 69.167 and 0.0625 using (8) and according to [20] which gave range as [0.05,0.07]. Setting ϵ at 0.05 from the selected range and γ at 0.0625, C has favorable range in [1,60]. Now, ϵ at 0.05 and C at 30 from selected favorable range γ had favorable range in [0.01,0.1].

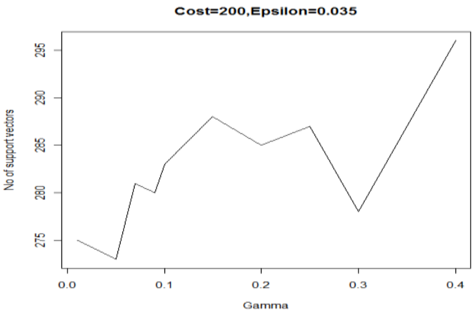


Figure 6(a)

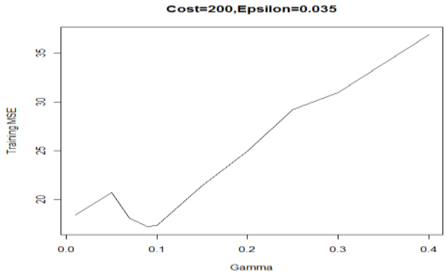


Figure 6(b)

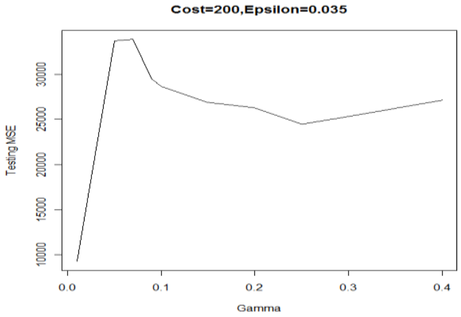


Figure 6(c)

4) DE-SVM model:

All implementation has been done in R on a system with AMD Turion-X2 2GHz Dual Core processor having 2GB RAM and Windows 7 Ultimate (32 bit) OS. The model used is shown in Figure 7.

IV. Results

Apple:

Normalized dataset: The range of parameters C,ϵ,γ are [1,550], [0.033,0.052] and [0.01,0.11] respectively. The time taken for both PSO and DE to converge is 13hrs approx. The results for both DE and PSO are shown in Table 1. Table 2 shows prediction results for SVM (with default parameters of C=1, ϵ=0.1,γ=0.2), DE-SVM and PSO-SVM together.

Non normalized dataset: The range of parameters C,ϵ,γ are [1,300], [0.033,0.052] and [0.01,0.1] respectively. The results for DE-SVM and PSO-SVM are shown in Table 3. Table 4 shows prediction results for SVM (with default parameters of C=1, ϵ=0.1,γ=0.2), DE-SVM and PSO-SVM together. The large values of MSEs for testing are because of the highly inaccurate predicted values produced because of the wide range of the output values.

Table 1

	Optimized C,ϵ,γ	Training MSE	Testing MSE	No of support vectors
DE	286.37295110, 0.03567755, 0.08290609	0.001520	0.006520839	277
PSO	312.57590986, 0.03556743, 0.08097982	0.001519326	0.006537451	277

Table 2

	Training MSE	Testing MSE	No. of support vectors
SVM	0.003224442	0.008572274	94
DE-SVM	0.001520	0.006520839	277
PSO-SVM	0.001519326	0.006537451	277

For both DE and PSO, normalized and non normalized cases, the population size has been fixed at 30 and iterations at 200. The prediction results of DE-SVM and PSO-SVM are better than SVM alone in both cases.

Table 5 shows predicted stock price values for both normalized and unnormalized datasets for SVM, DE-SVM and PSO-SVM models.

Table 3

	Optimized C,ϵ,γ	Training MSE	Testing MSE	No. of support vectors
DE	298.574181, 0.035675, 0.082388	17.029155	30013.67	276
PSO	296.05980293, 0.03569455, 0.08201110	17.01084	30175.51	276

Table 4

	Training MSE	Testing MSE	No. of support vectors
SVM	36.10602	30383.61	94
DE-SVM	17.029155	30013.67	276
PSO-SVM	17.01084	30175.51	276

Table 5

Original	Normalized data			Unnormalized data		
price	SVM	DE-SVM	PSO-SVM	SVM	DE-SVM	PSO-SVM
603	623.7251	615.6471	615.6017	264.1839	248.7355	248.1471
545.17	552.267	553.2582	553.4743	264.2948	263.0166	262.1655
493.42	497.0696	496.938	496.959	267.2491	264.4146	263.1924
369.8	359.912	367.1551	367.1906	354.1847	359.2065	359.2020

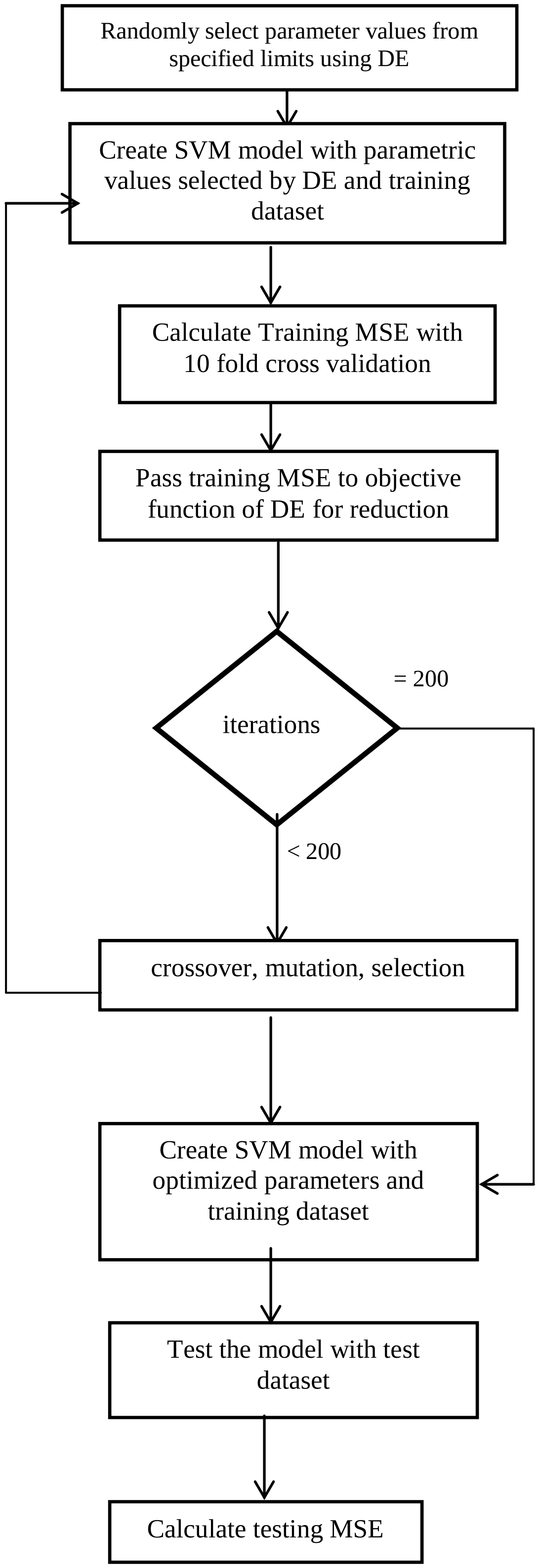

Randomly select parameter values from specified limits using DE
Create SVM model with parametric values selected by DE and training dataset
Calculate Training MSE with 10 fold cross validation
Pass training MSE to objective function of DE for reduction
iterations
= 200
< 200
crossover, mutation, selection
Create SVM model with optimized parameters and training dataset
Test the model with test dataset
Calculate testing MSE

Figure 7

Honeywell:

Normalized dataset: The range of parameters C,ϵ,γ are [1,440],[0.08,0.15] and [0.02,0.08] respectively. Table 6 shows prediction results.

Table 6

SVM	Optimized C, ϵ,γ	Training MSE	Testing MSE	No. of support vectors	Time taken
DE (CR= 0.7, F= 0.9)	439.864990, 0.080024, 0.079993	0.003211084	0.03108 918	222	4 hrs 10 min
PSO	440, 0.08 0.07999401	0.003210875	0.03117 105	222	4 hrs 30 min

Non normalized dataset results: The range of parameters C,ϵ,γ are [1,60],[0.05,0.07] and [0.01,0.1] respectively. Table 7 shows the results. Predicted values for both normalized and non normalized datasets are shown in Table 8.

For DE CR=0.7 and F=0.9. DE / local-to-best / 1 / bin strategy has been used for DE for all the implementations in this paper.

Table 7

	C,ϵ,γ	Training MSE	Testing MSE	No. of support vectors
DE-SVM	40.543474, 0.056122, 0.010113	0.4726898	1.603333	256
PSO-SVM	40.7422794, 0.06411372, 0.01000007	0.4727329	1.620285	225
SVM	1,0.1,0.2	0.8239745	3.8681	148

Table 8

Original	Normalized			Unnormalized		
Closing price	SVM	DE-SVM	PSO-SVM	SVM	DE-SVM	PSO-SVM
41.94	42.503	42.340	42.34	43.019	42.95	42.95808
62	61.555	61.63	61.64	56.079	61.56	61.53316
43.22	44.15	44.002	44.00	45.36	44.65	44.66156
53.56	53.84	54.15	54.15	54.185	54.44	54.47695
56.43	56.373	56.64	56.65	56.546	56.654	56.70812

V. Conclusion

The performance of SVM can be significantly affected by choice of its free parameters of cost (C), insensitive loss function (ϵ) and kernel parameter (γ). The results show that DE-SVM model's performance is comparable to that of PSO-SVM. Performance of these models can be improved by normalization of datasets. Normalization helps to significantly improve the accuracy of the output when the range of values is vast. Normalization gives equal weightage to all the input variables by converting the values of all the variables within a pre-specified range. This helps to avoid dominance of one variable over others in the created model. So, it helps to improve the efficiency of the created model. SVM alone performs better when data is normalized because in hybrid models optimization techniques help to tune the model according to requirement of datasets. With normalization of data, the range for optimization of C,ϵ,γ improves.

VI. Future Scope

DE responds to the population progress after a time lag. The whole population in DE remains unchanged until it is replaced by a new population [15]. Hence, it results in slower convergence. To alleviate this problem, a dynamic version of DE called Dynamic Differential Evolution (DDE) has been proposed by Anyong Qing [23]. DEPSO algorithm, which represents more stability by dual evolution, proposed by Ying-Chih Wu [24] can be used for optimization of SVM. The above mentioned methods will help to further improve the efficiency of SVM and hence improve results.

References

[1] Yohanes Budiman Wijaya, S.Kom,Togar Alam Napitupulu, “Stock price prediction: comparison of Arima and artificial neural network Methods”, in Second International Conference on Advances in Computing, Control, and Telecommunication Technologies, IEEE, 2010.

[2] Qiang Ye, Bing Liang, Yijun Li, “Amnestic Neural Network for Classification: Application on Stock Trend Prediction”, in Proceedings of ICSSSM, IEEE, 2005.

[3] Ling-Feng Hsieh, Su-Chen Hsieh b, Pei-Hao Tai, “Enhanced stock price variation prediction via DOE and BPNN-based optimization”, in Expert Systems with Applications, 2011.

[4] Mustafa E. Abdual-Salam, Hatem M. Abdul-Kader, and Waiel F. Abdel-Wahed, “Comparative Study between Differential Evolution and Particle Swarm Optimization Algorithms in Training of Feed-Forward Neural Network for Stock Price prediction”, in 7th International Conference on Informatics and Systems, IEEE, 2010.

[5] Zhang Da, Yong Song, Hong wei, Chen Pu, “Stock Market Forecasting Model Based on A Hybrid ARMA and Support Vector Machines”, in International Conference on Management Science & Engineering (15th), IEEE, 2008.

[6] Zhonghai Chen, Yong Sun, Guoli Yang, Tengfei WU, Guizhu Li, Longbiao XIN, “Air Conditioning Load Prediction Based on DE-SVM Algorithm”, in Third International Symposium on Intelligent Information Technology and Security Informatics, IEEE, 2010

[7] Yong Sun, Guoli Yang, Limin Wang, Yongjiang Shi, Yongqiang Wu, “Gas Load Prediction Based on DE-SVM Algorithm”, in 2nd International Conference on Future Computer and Communication, IEEE, 2010.

[8] Jośe Garćıa-Nieto, Enrique Alba, Javier Apolloni, “Hybrid DE-SVM Approach for Feature Selection: Application to Gene Expression Datasets”, in 2nd International Logistics and Industrial Informatics, IEEE, 2009.

[9] Shu Jun, Li Jian, “A Combination of Differential Evolution and Support Vector Machine for Rainstorm Forecast”, in 2009 Third International Symposium on Intelligent Information Technology Application, IEEE, 2009.

[10] Jiang An-nan, Jin Lu, “Studying the Lithology Identification Method from Well logs Based on DE-SVM”, in Control and Decision Conference, IEEE, 2009.

[11] Jiawei Han, Micheline Kamber, Data Mining:Concepts and Techniques, Morgan Kaufmann Publishers: San Francisco, 2nd ed., 2006.

[12] David Taniar, Data Mining and Knowledge Discovery Technologies, Idea Group Publishing: USA, 2008.

[13] David Taniar, Research and Trends in Data Mining Technologies and Applications, Idea Group Publishing: USA, 2007.

[14] Lei Peng, Yuanzhen Wang, “UDE: Differential Evolution with Uniform Design”, in 3rd International Symposium on Parallel Architectures, Algorithms and Programming, IEEE, 2010.

[15] G. Jeyakumar, C. Shunmuga Velayutham, “A Comparative Performance Analysis of Differential Evolution and Dynamic Differential Evolution Variants”, in World Congress on Nature & Biologically Inspired Computing, IEEE, 2009.

[16] Youyun Ao, Hongqin Chi, “Experimental Study on Differential Evolution Strategies”, in Global Congress on Intelligent Systems, IEEE, 2009.

[17] Nizar Hachicha, Bassem Jarboui, Patrick Siarry,” A fuzzy logic control using a differential evolution algorithm aimed at modelling the financial market dynamics”, in Information Sciences, pp. 79-91, 2011.

[18] Xie Guo-qiang, “The Optimization of Share Price Prediction Model Based on Support Vector Machine”, in International Conference on Control, Automation and Systems Engineering , IEEE, 2011.

[19] Mattera, D., & Haykin, S., “Support vector machines for dynamic reconstruction of a chaotic system”. In B. Scho¨lkopf, J. Burges, & A. Smola (Eds.), Advances in kernel methods: Support vector machine. Cambridge, MA: MIT Press, 1999.

[20] Vladimir Cherkassky, Yunqian Ma, “Practical selection of SVM parameters and noise estimation for SVM regression”, in Neural Networks, pp. 113–126, 2004.

[21] Scott Hemby, Sabine Bahn, Functional Genomics And Proteomics in the Clinical Neurosciences, Elsevier: UK, 1st ed., pp. 102, 2006.

[22] Kenneth V. Price, Rainer M. Storn, Jouni, A. Lampinen, Differential Evolution A Practical Approach to Global Optimization, Springer-Verlag: Germany, 2005.

[23] Anyong Qing, “Dynamic Differential Evolution Strategy and applications in electromagnetic inverse scattering problems,” in IEEE Transactions on Geoscience and Remote Sensing, Vol. 44, No. 1, January 2006.

[24] Ying-Chih Wu, Wei-Ping Lee, Ching-Wei Chien, “Modified the Performance of Differential Evolution Algorithm with Dual Evolution Strategy”, in 2009 International Conference on Machine Learning and Computing, IPCSIT vol.3, IACSIT Press, Singapore, pp. 57-63, 2011.

[25] Nianyi Chen, Wencong Lu, Jie Yang, Support vector machine in chemistry, World Scientific Publishing Co. Pte. Ltd: USA, 2004.

[26] Sheng Liu, Na Jiang, “SVM Parameters Optimization Algorithm and Its Application”, in Proceedings of 2008 IEEE International Conference on Mechatronics and Automation, pp. 509-513, IEEE, 2008.

[27] Cherkassky, V., &Mulier, F., Learning from data: Concepts, theory, and methods, Wiley:New York, 1998.

[28] Kaibo Duan, S. Sathiya Keerthi, Aun Neow Poo, “Evaluation of simple performance measures for tuning SVM hyperparameters”, in Neurocomputing 51 pp. 41 – 59, 2003.

[29] Wilke, D. N., Analysis of the particle swarm optimization algorithm, Master's Dissertation, University of Pretoria, 2005.

[30] Mahamed G.H. Omran, Andries P. Engelbrecht, Ayed Salman, “Discrete Optimization Bare bones differential evolution”, in European Journal of Operational Research, pp. 128-139, 2009.

[31] Kennedy, J., Eberhart, R.C., Shi, Y., Swarm Intelligence, Morgan Kaufmann: San Francisco, 2001.

[32] Kennedy, J., “Small worlds and mega-minds: Effects of neighborhood topology on particle swarm performance”, in IEEE Congress on Evolutionary Computation, vol. 3, pp. 1931–1938, 1999.

[33] Kennedy, J., Mendes, R., “Population structure and particle performance”, in IEEE Congress on Evolutionary Computation. IEEE Press, pp. 1671–1676, 2002.

[34] Riccardo Poli, James Kennedy, Tim Blackwell “Particle swarm optimization An overview”, Swarm Intell, DOI 10.1007/s11721-007-0002-0, Springer Science, 2007.

The Impact on Effectiveness and User Satisfaction of Menu Positioning on Web Pages

Dr Pietro Murano
University of Salford, School of Computing, Science and Engineering,
Salford, M5 4WT, UK

Kennedy K. Oenga
University of Salford, School of Computing, Science and Engineering,
Salford, M5 4WT, UK

Abstract— **The authors of this paper are conducting research into the usability of menu positioning on web pages. Other researchers have also done work in this area, but the results are not conclusive and therefore more work still needs to be done in this area. The design and results of an empirical experiment, investigating the usability of menu positioning on a supermarket web site, are presented in this paper. As a comparison, the authors tested a left vertical menu and a fisheye menu placed horizontally at the top of a page in a prototype supermarket web site against a real supermarket web site using a horizontal menu placed at the top of a page. Few significant results were observed, which gave rise to the conclusion that overall there were not many differences between the tested menu types. Furthermore, an explanation for the results observed is discussed in terms of cognitive, physical, functional and sensory affordances. It is suggested that observation of the affordances may be a more crucial aspect to menu design than the actual menu positioning.**

Keywords- Usability; menu design; menu positioning; affordances.

I. Introduction

Most web sites make use of some sort of structure for organising content. The content is then usually accessed by means of various different types of navigation elements. The most common of these tend to be menu-based, where the menu(s) is potentially placed at different locations on the user interface - depending on the designer. Some examples of commonly used navigation schemes are to have menus at the top, bottom, left or right sides of a web page and in some cases combinations of these will also be used. One of the common combinations one can see on certain web sites is to have a horizontal menu at the top of a page and a vertical menu at the left side of the page (inverted-L configuration).

Various dedicated Human Computer Interaction texts also devote some effort in discussing the appropriate design of menus, e.g. Dix, Finlay, Abowd and Beale [4], Benyon [1] and Rogers, Sharp, and Preece [13]. However, despite being able to access guidelines and advice for menu design, the real life situation is that empirical evidence regarding the effectives and user satisfaction of different menu designs is inconclusive overall when various studies are considered as a whole.

While the authors of this paper acknowledge that each type of menu has stylistic aspects to it, we are more concerned with effectiveness and user satisfaction of menu types and their positioning used on web pages. Various studies have been carried out to try and assess the effectiveness and user satisfaction of different menu types. However to our knowledge the results are not completely conclusive and therefore this is still a worthy area of research. We are seeking to contribute to the body of knowledge concerning effectiveness and user satisfaction of menu types and positioning in web pages. Furthermore it is our aim to explain our results in terms of the theory of affordances as expressed by Hartson [7] in the context of user interfaces. Overall this is a very important area of research because the success of a web site involves several different aspects. One of these is the usability of the menus and their positioning. A web site that lacks usability in some form, e.g. having bad menu design and positioning, could lead to loss of business for the owners or simply to not receiving any visitors.

Therefore this paper will firstly discuss some related works in similar areas to our research. This will then be followed by a description of an experiment carried out aiming to determine effectiveness and user satisfaction of different menu positioning on web pages. This will be followed by a presentation of the main results and linked to the theory of affordances [7]. Lastly conclusions and future work will be discussed.

II. Related Work

As suggested above, some researchers have been investigating similar issues to the work presented in this paper. However, to our knowledge the overall conclusions regarding effectiveness and user satisfaction are inconclusive and therefore worthy of being studied further.

In a study by Burrell and Sodan [3], menu positioning was investigated using six different menu positions. These were: 1. A tabbed menu placed horizontally and at the top of the page 2. A horizontal menu placed at the top of the page (not tabbed) 3. A menu placed horizontally at the top of the page and a menu placed vertically at the left side of the page 4. A vertical menu placed at the left of the page 5. A menu placed horizontally at the top of the page and also a menu placed horizontally at the bottom of the page 6. A menu placed horizontally at the top of the page and a menu placed vertically at the right side of the page.

The authors then conducted a study with prototype web sites and the above described menu positions. The actual tasks carried out by participants are not clearly indicated in the paper. However it may be that free form exploration was used by the participants so as to experience the different menu positions. The authors concluded from their data that the tabbed menu was preferred by participants. While the idea for the investigation of the six menu positions was interesting, this work is lacking in various ways. Some examples include that ideally such a study should also investigate effectiveness by using some carefully designed tasks and measures. This did not seem to be a part of the study. Also, as mentioned above, the details of the tasks are not revealed in the paper and it is therefore difficult to judge if there were any biasing factors in the task design.

In another study by Dos Santos, De Lara, Watanabe, Filho and Fortes [5] eight different types of horizontal menus were tested with participants. The menu types were, with actual names/descriptions:

1) *A superfish dropdown menu*
2) *A mega-dropdown menu*
3) *A Vimeo style dropdown menu*
4) *A Simple jQuery Dropdown menu*
5) *A Sexy dropdown menu*
6) *'A different top navigation' menu*
7) *A horizontal menu 'that creates columns for grouping information in sub-menus' [4]*
8) *A jQuery (mb) Menu 2.7.*

The study was a within users study (for the menu types) with a between users element (age grouping), i.e. this was a mixed design. The study involved participants carrying out two tasks involving some menu usage. Although the tasks are not clearly described in the paper, the paper indicates that there was an element of 'looking for/finding information' as part of the tasks. Based on these activities, the researchers recorded the average time to complete the tasks, the errors committed by participants and the task completion rate. This data was collected remotely by means of software logging.

The overall results of the work are not entirely clear. The authors' argumentation is mostly centred around averages. However it is not clear in several cases if there are any statistically significant differences observed. A visual inspection of some of the bar charts presented in the paper may indicate some significance, but without the actual data being available it is not possible to categorically confirm this aspect. The authors do argue that their results suggest that menus 1 and 3 (see above) were better. However they do not directly state if this is in terms of fewer errors or faster task completion times, or both. Furthermore the authors' comparisons were restricted to horizontal menus placed at the top of a page, which makes the study limited in nature. Menus placed in other positions on a page should ideally have been investigated.

Another study worthy of consideration is by Leuthold, Schmutz, Bargas-Avila, Tuch, and Opwis [8]. This was a study where the authors compared three types of vertical menu positioned at the left side of the page. These were:

1) *A simple menu consisting of clickable links*
2) *A menu like the simple menu in 1, but with more links which were grouped under various headings ('service navigation items') and*
3) *A dynamic menu where various headings could be expanded by a user by clicking on a heading, which would reveal further clickable options. This was essentially a compacted version of menu 2.*

The authors used eye-tracking equipment to gather data. The context of the web site was a storefront which purported to sell books, DVDs and music. The study involved a series of participants taking part in a simple task and then a more complex task. The tasks basically involved navigating through some of the links to find some information and potential items for purchase. The authors measured user performance, navigational approaches and user preferences.

The authors' results suggested that there was a greater success rate with a first click whilst participants used menu 2 (described above) for the simple and more complex tasks. Also fewer eye fixations were required for the simple and more complex tasks whilst using menu 2. However with menu 1, participants were faster whilst undertaking a simple task and participants were faster with menu 2 during a more complex task. Regarding subjective opinions, overall the authors found that menu 2 was the preferred option.

The research presented by the authors is interesting. However it does have some weaknesses. The first of these is that the menus tested were of the vertical type placed at the left side of the page. It would have been interesting to have tested menus in different positions on the page – although it is accepted that for the purposes of their hypothesis other positions were not necessarily required. Also some of the results are rather obvious in nature, e.g. since menu 3 was a compacted version of menu 2, it is rather obvious that menu 3 would require more clicks to use and therefore more time. Regarding the user satisfaction aspects, we would argue that the measures reported in the paper were rather coarse grained in nature and more detail should have been elicited from the participants in order to reveal detailed perceptions of satisfaction.

A further aspect to consider about the results and the study as a whole is that it gives us some insight into the kinds of links that could be suitable to include in a menu. However this work does not deal with the issue of the actual positioning of menu items and menus on a page.

The last study we will consider in this paper is by Bernard, Hamblin and Chaparro [2]. This was a study where three menu types were evaluated. These were: 1 An index menu, where the menu options appearing as links were all displayed in the centre of the page, 2. A horizontal menu at the top of the page and 3. A vertical menu positioned at the left of the page.

The authors aimed to design realistic tasks based on browsing for information on a web site. Some of the tasks involved finding specific products, while some of the tasks were more vague in nature, because participants were presented with a scenario type context. This context did not specifically give a specific product to find, but some product would be implied in the context given.

The experiment used a between users design and the authors measured task completion time, search efficiency and participants' subjective opinions regarding their interaction experience.

Their results showed with statistical significance that the index menu incurred faster times for task completion compared to the other two menus included in the study. Several aspects of participants' perceptions were elicited and most of these were not statistically significant. However participants did indicate that their first choice would be the index menu design. Although it is unclear from [2] if this finding was statistically significant.

Overall the work in [2], in our examination, seems to be one of the more rigorous studies published in this area and is in our opinion the most rigorous we have summarised here. However there were details missing regarding the actual procedure followed during their experiment and it would have been safer to have had a slightly larger participant sample. Also as acknowledged by the authors of [2], the menus used in their study did not descend to very deep levels.

Another aspect of previous work that we wish to briefly summarise concerns the theory of affordances which we will use to explain our observations. The theory of affordances was initially suggested by Gibson [6]. However, over time, some researchers began to apply and extend the theory to user interfaces, e.g. Hartson [7] and Norman [11, 12].

Hartson suggested the existence of cognitive, physical, functional and sensory affordances. He reasoned that when users are doing some computer related task, the users are using cognitive, physical and sensory actions.

Cognitive affordances involve ‘a design feature that helps, supports, facilitates, or enables thinking and/or knowing about something’ [7]. A simple example of this, concerns presenting feedback to a user that is clear and precise. If a designer labels a button, the label should easily indicate to the user what will happen if the button is clicked.

Physical affordances are ‘a design feature that helps, aids, supports, facilitates, or enables physically doing something’ [7]. Hartson suggests that a button that can be clicked by a user is a physical object acted on by a human and the button size should be big enough to allow easy clicking. This would therefore be a physical affordance characteristic. Functional affordances concern having some purpose in relation to a physical affordance. A simple example is that clicking on a button should have some purpose with a goal in mind. The opposite is that just clicking anywhere on the screen is not purposeful and has no goal.

Finally, sensory affordances concern ‘a design feature that helps, aids, supports, facilitates or enables the user in sensing (e.g. seeing, feeling, hearing) something’ [7]. Sensory affordances are linked to cognitive and physical affordances as they complement one another. Therefore the users need to be able to ‘sense’ the cognitive and physical affordances so that these affordances can help the user.

This brief consideration of some of the key work in this research area (including the theory of affordances), shows that there is still much more work to be done in order to discover more conclusively which menu design may be more effective and satisfying for users. Although other researchers have done some work in this area, often limits in rigour and limitations in the types of menu design tested, show that more evidence needs to be gathered for the benefit of the research community and user interface designers. In the next section we present the details and results of a study where different menu designs were evaluated.

III. Menu Comparison Experiment

In order to add to the body of knowledge regarding the usability of different menu types and layouts, a small prototype was developed to simulate a supermarket web site. The prototype used the same colour scheme and products available on the real supermarket web site. This was then compared with the real supermarket web site, with the main varying components being the menu design. The prototype web site used a left vertical menu and a fisheye menu placed horizontally at the top of the page. The real supermarket web site used a horizontal menu placed at the top of the page. Overall the aim was to discover if these differences in menus and their placement on the web page affected user performance and satisfaction.

A. Hypotheses

We devised several hypotheses around the area of efficiency of use and user satisfaction for the purposes of this experiment. In all cases we were looking for statistically significant differences in the data to be collected.

1) a) H_0: There will be no difference in the number of navigation errors made in using either of the two web sites.

b) H_1: Participants using the prototype supermarket web site will make fewer navigation errors than those using the real supermarket web site.

2) a) H_0: There will be no difference in the ease of use (efficiency) of the two web sites' menu navigation systems.

b) H_1: Participants will find the prototype supermarket web site navigation easier to use (efficient) than the real supermarket web site menu navigation systems.

3) a) H_0: There will be no difference in the participants’ satisfaction level between the two web sites.

b) H_1: Participants’ satisfaction level for the prototype supermarket web site will be higher than that of the real supermarket web site.

4) a) H_0: There will be no difference between the two web sites for task time.

b) H_1: Participants using the prototype supermarket web site will incur shorter task times.

B. Users

Since the experiment involved testing aspects of menu design and positioning on a web page, it was deemed important to have participants with a certain amount of experience in using web sites and computers in general. This is because if there happened to be a number of beginners to such activities, these could potentially bias times and outcomes. Therefore:

- 56 undergraduate students took part in the experiment.
- All participants had not visited the real supermarket web site in the past and had not seen the prototype web site prior to the experiment.
- All participants had basic IT skills.
- All participants had experience with the Internet and online shopping experience.
- All participants were in the 20-40 age range.

These aspects were elicited by means of a carefully designed pre-experiment questionnaire.

C. Experimental Design

Since the tasks and 'products' being used within the tasks were the same for both web sites, a between users design was deployed. This would help to avoid the possibility of some 'learning' taking place, regarding the specific products used. Each participant was randomly assigned to one of the two conditions of the experiment. The two conditions were the prototype supermarket web site and the real supermarket web site.

D. Variables

The independent variables were (1) the types of menu being investigated (horizontal menu placed at the top of the page, left vertical menu and the horizontal fisheye menu) and (2) the type of task involving using the menus described, in finding a series of typical products sold in supermarkets.

The dependent variables were the participants' performance in carrying out the tasks and their subjective opinions.

The dependent measures were that the performance was measured by examining the time to complete the tasks, the number of errors made and the success level. The success level was determined by observation of the participants' 'behaviour' and interaction with the web sites. This involved making a decision regarding whether a task was completed with ease, completed with difficulty or not completed at all. An error was recorded if a participant deviated from the optimum path to achieve a task by clicking on an incorrect link. This was a good indicator of aspects of the interface that misled the user.

The subjective opinions were measured by means of a post-experiment questionnaire. The time and errors were recorded by using the Morae [10] software suite. The timing was started by clicking the 'record' button when the participant felt ready to begin and the time was stopped when the participant clicked on the home page link of each respective web site. For the context of this study, clicking on incorrect links that did not lead to the expected information, were categorised as errors. Lastly the post-experiment questionnaire that was used for eliciting subjective opinions had a series of sections where the participant responses were made using Likert [9] type scales. The main areas covered by the post-experiment questionnaire were opinions about the navigation styles being tested, ease of learning of the navigation types, ease of remembering one's position on the web pages, efficiency and feelings of satisfaction in using the web sites and their navigation types.

E. Apparatus and Materials

The experiment took place in a well lit room containing a desk, and three computer chairs.

Two laptops were used in this experiment and for each the screen display was set to a resolution of 1280 by 800 pixels with the colour set to highest (32 bit). Laptop 1 was a Sony Vaio with a 64 bit processor Intel (R) Core TM2 Duo, CPU T6600 2.20GHz and 4.00 GB RAM. This was used by the researcher and was running Morae Observer [10]. The Morae Observer in the Sony laptop connected with Morae Recorder [10] on laptop 2 on a wireless network using an IP address. The two web sites were also run on laptop 2. This was a 32 bit HP Compaq 6735s, with an AMD Sempron, 2.10 GHZ and 2GB RAM. The operating system for both laptops was Windows Vista Home Basic and Internet Explorer 8 was used for the web browser.

Morae Recorder was used for digitally capturing the participants' interaction and Morae Observer was used by the experimenters to observe the participants' interactions in real time without interrupting the participants' interaction in any way.

Five tasks were designed for this experiment. Each centred around typical shopping type activities on a supermarket web site. Further, the information/products participants had to find, involved using the menus to various sub-levels in the hierarchical structure of the web sites. To introduce the tasks and make them more realistic a small scenario was presented to the participants, as follows:

You intend to buy a few Christmas presents for your younger brother. These items are to be purchased online as the shops are overcrowded around this time when people are busy with their Christmas shopping. The Items you require are available at the supermarket's online shop. You are interested in buying a packet of milk, an iPod and a Scooter. You are not required to purchase the items and for this experiment only, you do not use the search button to get to the products.

Therefore the five tasks were as follows:

Task 1: Use the navigation links to locate *Whole Milk 1.13L (2 pint)* and add it to the shopping cart. Then click on the home button to end this task.

Task 2: Use the navigation links to locate the *New iPod Shuffle 2GB – Pink* and add it to the shopping cart. Then click on the home button to end this task.

Task 3: Start the task from the Horizontal link buttons, and then the Icons that appear in the subsequent pages (do not use the left menu buttons) to locate *Lightning Strike Scooter - Pink* and add it to the shopping cart. Then click on the home button to end this task.

Task 4: This task is a continuation from tasks 1, 2 and 3 above. Each time you purchase an item from the supermarket, you can collect loyalty points. How can you get **double** loyalty points from your purchase? The researcher will provide you with paper to write your answer.

Task 5: What is the difference in Giga Bytes (GB) between the new iPod Shuffle and New iPod Nano Silver (pictures provided). Write your answer on the paper provided by the researcher.

F. Procedure

The procedure followed and described in this section was initially pilot tested with three independent individuals. The pilot testing showed that the designed procedure was accomplishing the objectives of the study without any obvious problems.

Therefore, each participant was asked to present themselves to a specific room in the institution set aside for the experiment. During the experiment each participant was seated at the desk in the room with the laptop facing them and the researcher sat opposite the participant with the second monitoring laptop facing the researcher.

Each participant was briefed about the web applications and it was stated that the study was evaluating the web applications rather than the participants. Participants were told that the tasks started from the home page and that after adding the items to the shopping basket, each task ended when the participant pressed the home button on the website. Then the participants were given instructions on how to perform the tasks.

They were also asked to complete an informed consent form and fill out a pre-experiment questionnaire that included questions about demographics and computer skills.

A piece of paper was provided for the participants to write answers for tasks four and five. The researcher ensured that the participants started from the home page and ended their tasks by pressing the home button.

The researcher explained that the amount of time taken to complete each task would be measured and that they should not engage themselves in any exploratory behaviour outside the task flow until after the experiment had been completed. Participants were also given the opportunity to ask questions.

Then the participants were given the tasks to do in the order described in the previous section. Time was allowed for them to read the tasks and understand them fully before they started. There were five tasks for each web site and the participants were free to follow any navigation route they wished, except for task three which required them to use the horizontal fisheye menu first (see Variables section above for a description of the dependent measures and how they were recorded).

After completing the tasks, the participants were prompted to fill out an electronic post-experiment questionnaire concerning user satisfaction (see Variables section above for a summary of the areas covered by the post-experiment questionnaire).

G. Results

In this section, for brevity, only the significant results are presented. For the data collected, the distributions were examined which included the respective means (M) and standard deviations (SD). Then the data was subjected to Multifactorial Analysis of Variance (MANOVA) testing and where significance was found, the significance was confirmed by means of post-hoc testing using either t-tests or Tukey HSD tests (post-hoc tests not included in this paper for brevity).

For task 4, which involved aspects of finding double loyalty points, there was a significant difference for the number of errors committed, where the prototype supermarket web site (M = 0.93 errors, SD = 0.77) incurred significantly (F (5, 50) = 3.26, P<0.05) more errors than the real supermarket web site (M = 0.39 errors, SD = 0.63).

The subjective question concerning overall ease of navigation was scored by participants using a Likert [9] type scale of 1-5, where 5 was the most positive response possible and 1 was the most negative response possible. This question incurred a significant difference in opinions between the various age groups which took part in the experiment. The slightly older groups in both experimental conditions (M = 3.75 for 36-45 age group, M = 3.36 for 31-35 age group) rated the ease of navigation significantly (F (5, 50) = 2.60, P<0.05) less easy than the younger groups in both experimental conditions (M = 4.08 for 25-30 age group, M = 4.61 for 19-24 age group).

The subjective question concerning the web sites being easy to learn to use by anyone was scored by participants using a Likert [9] type scale of 1-5, where 5 was the most positive response possible and 1 was the most negative response possible. This question incurred a significant difference in opinions across the two web sites. The prototype supermarket web site (M = 3.39 opinion score, SD = 1.31) incurred significantly (F (5, 50) = 2.71, P<0.05) lower/more negative opinion scores than the real supermarket web site (M = 4.32 opinion score, SD = 1.12).

The subjective question concerning the web sites' navigation being suitable for all levels of users was scored by participants using a Likert [9] type scale of 1-5, where 5 was the most positive response possible and 1 was the most negative response possible. This question incurred a significant difference in opinions across the two web sites. The prototype supermarket web site (M = 2.89 opinion score, SD = 1.59) incurred significantly (F (5, 50) = 3.89, P<0.01) lower/more negative opinion scores than the real supermarket web site (M = 3.68 opinion score, SD = 1.28).

The subjective question concerning the text size used for menu labelling and the ease of reading such labels, was scored by participants using a Likert [9] type scale of 1-5, where 5 was the most positive response possible and 1 was the most negative response possible. This question incurred a significant difference (F (5, 50) = 2.76, P<0.05) in opinions between the various age groups which took part in the experiment - across the two experimental conditions (M = 1.25 for 36-45 age group, M = 3.21 for 31-35 age group, M = 2.75 for 25-30 age group and M = 2.44 for 19-24 age group).

The subjective question concerning a lack of willingness to use the web sites in the future was scored by participants using a Likert [9] type scale of 1-5, where 5 indicated full agreement in not wanting to use the web sites again and 1 indicated complete disagreement in that participants would want to use the web sites again. This question incurred a significant difference in opinions across the two web sites. The prototype supermarket web site (M = 3.32 opinion score, SD = 1.19) incurred significantly (F (5, 50) = 3.32, P<0.05) more positive opinions towards being willing to use the web site in the future than the real supermarket web site (M = 4.25 opinion score, SD = 1.08).

The next section will discuss the results presented above in relation to the hypotheses already presented above and the theory of affordances [7].

H. Experiment Results Discussion

Overall the significant results presented above, if taken in isolation from the other analysed 'variables' in the experiment, suggest the original real supermarket web site was preferred over the prototype supermarket web site. However the authors feel that the context of the many other subjective questions (see the Variables section above for a summary of the areas covered by the post-experiment questionnaire) asked should not be ignored. These other questions gave insignificant results across the two conditions being tested. This clearly suggests that overall opinion across the two conditions was rather uniform in nature with only a minimal amount of factors showing some statistical significance.

Furthermore, one of the significant results suggested a preference for the prototype supermarket web site in the context of being willing to use the web site in the future. It seems that the prototype supermarket web site in some way fostered some positive emotion in users as they indicated a stronger feeling of wanting to come back to the site.

Therefore due to the results not being so clear cut in terms of all the 'variables' under analysis, we cautiously accept Hypothesis 3(a) - $H_{0,}$ which stated that there would be no difference in the participants’ satisfaction level between the two web sites.

In addition, as can be seen in the previous section, significantly more errors were incurred with the prototype supermarket web site. This was specifically for Task 4. However, there were 5 tasks overall and clearly the other 4 tasks did not show any significant differences for errors. Also across the 5 tasks, the times taken to complete tasks did not show any significant differences across the different types of menus/the two web sites. Lastly, for each task, the success levels were also recorded and these did not show any significant differences across the two experimental conditions being tested.

Therefore, due to the results also not being so clear cut regarding the performance in the tasks (with the exception of the errors in Task 4), we cautiously accept Hypotheses 1(a) - H_0 and 4(a) - H_0. These stated that there would be no difference in the number of navigation errors made in using either of the two web sites and that there would be no difference between the two web sites for task time.

We also accept Hypothesis 2(a) - H_0 which stated that there would be no difference in the ease of use (efficiency) of the two web sites' menu navigation systems. The authors feel that overall there were not enough significant results in terms of the participant subjective responses and aspects of task time and success level in the tasks.

Lastly, as described in the previous section, we did observe some significant differences within the age groups of the participant groups. While this is not ‘age related’ research, we could not find any particular explanation for these findings, as the recruitment process did attempt to recruit participants with similar IT skills. However this could be a worthy avenue of further research for ‘age related’ studies.

Having linked back to the initial hypotheses (see Hypotheses section above), we are also interested in understanding the reasons for few significant results and therefore mostly no large differences between the different menu types. While there could be issues in the experimental design and execution, retrospective examination of the experimental design and execution reveal no obvious confounding variables. However, an examination of the menu types in relation to the theory of affordances as rendered by Hartson [7] could reveal some light on the matter.

Task 4 was the only task out of five tasks to have significant differences in terms of errors, even though for both web sites the information was only two clicks away from the starting point of the task. We suggest that the menu options needing to be chosen to reach the 'answer', perhaps violated the cognitive affordances, by not being labelled with a term that could indicate that double points information was available by clicking a certain menu option. Also the other tasks involved selection of menu options, for both web sites, that had labels which more clearly indicated the path to be followed and therefore observed the cognitive affordances more appropriately.

Regarding the participants' perceptions of the user interfaces used in the experiment, as discussed above, few significant results were observed. We would suggest that this is because the cognitive affordances were mostly equivalent for both web sites, e.g. in most cases the labelling of menu options was relatively clear for supermarket web sites of this kind. This therefore resulted in the 'side effect' of very few significant differences for user satisfaction. Also, we would suggest that the physical affordances were approximately equivalent for both web sites, because the sizing of the menus and each menu option, were of a size that made it easy to select an option. Finally we would suggest that the functional and sensory affordances were also largely equivalent for both web sites. The menu options had clear 'purpose' and worked as intended. Also as discussed above, the designs of both web sites had good visual clarity.

IV. Conclusions And Future Work

The results suggest overall that particularly in a supermarket shopping web site context, whether the menu is placed horizontally at the top of the page or is replaced with a horizontal fisheye menu and a left vertical menu does not seriously affect interaction time, accuracy and subjective perceptions. Furthermore, we would suggest that irrespective of menu positioning, designers should ensure as far is possible that the cognitive, physical, functional and sensory affordances are not violated in any way, as any violation of these could be more crucial than the actual positioning of the menu itself. However the authors of this paper suggest that more work still needs to be done to obtain empirical results for other types of menus and also to investigate in more depth issues of sub-menus and nesting. We suggest this because clearly other researchers have had different results which may indicate that there are other issues at play still to be discovered.

Also future experiments with more difficult tasks could lead to more understanding of the issues. This approach may show clearer results favouring a particular type of menu design. Furthermore more investigation needs to be done in relation to the theory of affordances [7] to gain more evidence that violation of the affordances creates more problems than the actual menu positioning. Lastly we suggest that perhaps other psychological aspects and/or user experience could have an effect that we have not identified yet.

REFERENCES

[1] D. Benyon, Designing Interactive Systems A Comprehensive Guide to HCI and Interaction Design, 2nd Edition, Addison Wesley, 2010.

[2] M. L. Bernard, C.J. Hamblin and B.S. Chaparro, Comparing Cascading and Indexed Menu Designs for Differences in Performance and Preference, Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting – 2003.

[3] A. Burrell and A.C. Sodan, Web Interface Navigation Design: Which Style of Navigation-Link Menus Do Users Prefer?, Proceedings of the 22nd International Conference on Data Engineering Workshops, ICDEW 2006, 3-7 April 2006, Atlanta, GA, USA, IEEE Computer Society.

[4] Dix, A, Finlay, J, Abowd, G.D and Beale, R, Human Computer Interaction, Pearson/Prentice Hall, 2004.

[5] E.P.B. Dos Santos, S.M.A. De Lara, W.M. Watanabe, M.C.A. Filho and R.P.M. Fortes, Usability Evaluation of Horizontal Navigation Bar With Drop-Down Menus by Middle Aged Adults, 29th ACM International Conference on Design of Communication, SIGDOC 2011, October 3-5, Pisa, Italy.

[6] J.J. Gibson, The Ecological Approach to Visual Perception, Houghton Mifflin Co. 1979.

[7] H.R. Hartson, Cognitive, Physical, Sensory and Functional Affordances in Interaction Design, Behaviour and Information Technology, Sept-Oct 2003, 22 (5), p.315-338.

[8] S. Leuthold, P. Schmutz, J.A. Bargas-Avila, A.N. Tuch, and K. Opwis, Vertical Versus Dynamic Menus on the World Wide Web: Eye Tracking Study Measuring the Influence of Menu Design and Task Complexity on User Performance and Subjective Preference, Computers in Human Behaviour, 27(2011) 459-472.

[9] R.A. Likert, Technique for the Measurement of Attitudes, Columbia University Press, NY, 1932.

[10] Morae, http://www.techsmith.com/morae.html, Accessed Feb 2012.

[11] D.A. Norman, Affordance, Conventions, and Design, Interactions, May-June, 1999, p.39-42.

[12] D.A. Norman, The design of Everyday Things, Basic Books, 2002.

[13] Y. Rogers, H. Sharp, and J. Preece, Interaction Design Beyond Human Computer Interaction, 3rd Edition, Wiley, 2011.

AUTHORS PROFILE

Dr Pietro Murano is a Computer Scientist at the University of Salford, UK. Amongst other academic and professional qualifications he holds a PhD in Computer Science. His specific research areas are in Human Computer Interaction and Usability of software systems.

Kennedy K. Oenga is a Computer Scientist. He has obtained an MSc in Databases and Web-Based Systems from the University of Salford, UK. One of his research interests is in Human Computer Interaction.

A New Automatic Method to Adjust Parameters for Object Recognition

Issam Qaffou
Département Informatique, FSSM
Université Cadi Ayyad,
Marrakech, Morocco

Mohamed Sadgal
Département Informatique, FSSM
Université Cadi Ayyad
Marrakech, Morocco

Aziz Elfazziki
Département Informatique, FSSM
Université Cadi Ayyad
Marrakech, Morocco

Abstract— **To recognize an object in an image, the user must apply a combination of operators, where each operator has a set of parameters. These parameters must be "well" adjusted in order to reach good results. Usually, this adjustment is made manually by the user. In this paper we propose a new method to automate the process of parameter adjustment for an object recognition task. Our method is based on reinforcement learning, we use two types of agents: User Agent that gives the necessary information and Parameter Agent that adjusts the parameters of each operator. Due to the nature of reinforcement learning the results do not depend only on the system characteristics but also the user's favorite choices.**

Keywords- component; Parameters adjustment; image segmentation; Q-learning; reinforcement learning.

I. Introduction

New tools and new algorithms for vision applications cause new system parameters that must be properly adjusted. This adjustment requires a specific knowledge, takes a long time, and sometimes even has to be done in an experimental process. To accomplish a segmentation task, the user must apply some operators, where each one has a set of parameter to adjust. The lack of a general rule that guides the user in his choices, the fixation of parameter values is usually made intuitively. The user proceeds by trying manually all possible cases until finding the desired result. Usually, in the majority of vision tasks we need to apply a combination of several operators where each one has a multitude of parameters to adjust. So, the manual adjustment becomes very tedious and not trustworthy. Therefore, an automatic method to adjust the values of each parameter is needed. The quality of results depends essentially on the operator chosen and the values assigned to its parameters.

Some GUI, for example Ariane [1], help users to accomplish a vision task by proposing them an interactive interface, but the values assigned to the parameters are selected manually by the user. Very few systems had succeeded to automate the process of parameter adjustment.

In [2], B.NICKOLAY et al. proposed a method to automatically optimize the parameters of a machine vision system for surface inspection by using specific Evolutionary Algorithms (EA). A few years later, Taylor proposed a reinforcement learning framework which uses connectionist systems as function approximators to handle the problem of determining the optimal parameters for a computer vision application even in the case of a highly dimensional, continuous parameter space [3]. More recently, Farhang et al. [9] introduced a new method for the segmentation of the prostate in transrectal ultrasound images, using a reinforcement learning (RL) scheme. He divided the initial image into sub-images and works on each one in order to reach a good result.

In this paper we propose a new method to adjust automatically the parameters of vision operators. Our method is based on reinforcement learning. We use two agents: User Agent (UA) and Parameter Agent (PA). The UA gives the necessary information to the system. It gives the combination of applicable operators, the set of adjustable parameters for each operator, values' ranges for each parameter. The PA uses reinforcement learning to assign the optimal values for each parameter in order to extract the object of interest from an image.

Due to the nature of RL, in terms of the interaction between state, action and reward, our approach takes in account not only the system opportunities but also the user preferences, and through the learning mechanism it will suggest trustworthy solutions.

An overview of reinforcement learning is given in section 2. Section 3 outlines the proposed approach and introduces a general framework for parameter adjustment. Section 4 presents the experimental results, and section 5 concludes the paper.

II. Reinforcement Learning

Reinforcement learning (RL) is learning what to do, how to map situations to actions, so as to maximize a numerical reward signal. The learner is not told which actions to take, as in most forms of machine learning, but instead must discover which actions yield the best reward by trying them. One of the challenges that arise in reinforcement learning and not in other kinds of learning is the tradeoff between exploration and exploitation. To obtain a lot of reward, a reinforcement learning agent must prefer actions that it has tried in the past and found to be effective in producing reward. But to discover such actions it has to try those that it has not selected before.

Reinforcement learning uses a formal framework defining the interaction between agent and its environment in terms of states, actions, and rewards, Fig 1.

Reward or punishment is determined from the environment, depending on the action taken. The agent must find a trade-off between immediate and long-term returns. It must explore the unseen states, as well as the states which maximize the return by choosing what the agent already knows. Therefore, a balance between the exploration of unseen states and the exploitation of familiar (rewarding) states is crucial. Watkins has developed Q-learning, a well-established on-line learning algorithm, as a practical RL method [6]. In this algorithm, the agent maintains a numerical value for each state-action, representing a prediction of the worthiness of taking an action in a state.

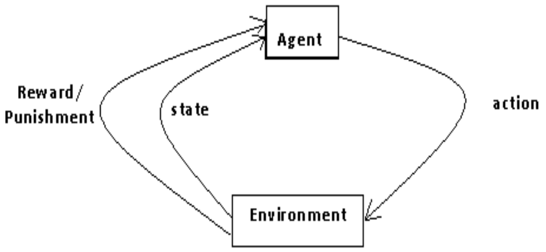


Figure 1: A general model for Reinforcement learning agent

Table 1 represents an iterative policy evaluation for updating the state-action values where r is the reward value received for taking action in states, s' is the next state, α is the learning rate, and γ is the discount factor [7]. There are some policies for taking action a given states. One of them is the Boltzman policy which estimates the probability of taking each action in each state. There are other policies for Q-learning such as ε- greedy and greedy. In the greedy policy, all actions may not be explored, whereas the ε-greedy selects the action with the highest Q-value in the given state with a probability of 1 – ε, and other ones with a probability of ε [7,8]. In this work an ε-greedy policy is used to make a balance between exploration and exploitation. The reward r(s, a) is defined according to each state-action pair (s, a). The goal is to find a policy to maximize the discounted sum of rewards received over time. The principal concerns in RL are the cases where the optimal solutions cannot be found, but can be approximated. The online nature of RL distinguishes it from other techniques that approximately solve Markov decision processes (MDP) [5,7].

TABLE 1. Q-LEARNING ALGORITHM

```
Initialize $Q(s,a)$ arbitrary
  Repeat (for each episode)
    Initialize state s
    Repeat (for each step of episode)
    Choose action $a$ from state s using policy derived from $Q$ (e.g., $\varepsilon - greedy$)
      Take action $a$, observe reward r, next state s'
         $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$
            $s \leftarrow s'$;
      Until s is terminal
```

In this paper, we attempt to introduce the RL concept for parameters adjustment.

III. The Proposed Approach

Generally, to accomplish an object recognition task, the user must apply sequentially some operators, and for each operator there is some parameter to adjust. Because there is no general rule that guides the user in his choices, he is based usually on his intuition to select values for each parameter. In the majority of vision tasks, we have to apply a multitude of operators that have several parameters to adjust. So adjusting manually these parameters basing only on the experience and on the intuition is not evident. It's a tedious work with a huge wasted time. In this paper we propose a new automatic method to find the best values for each parameter in a recognition task.

In our method we use two types of agents: User Agent (UA) and Parameter Agent (PA). Fig 2 shows the general framework of our method.

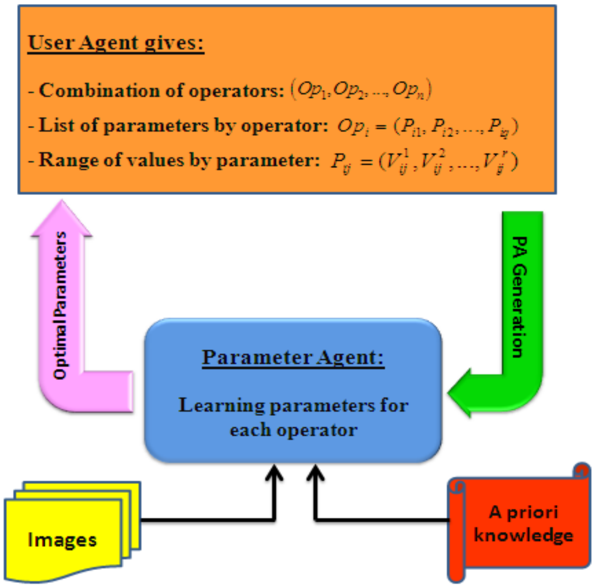

User Agent gives:
- Combination of operators: $(Op_1, Op_2, \ldots, Op_n)$
- List of parameters by operator: $Op_i = (P_{i1}, P_{i2}, \ldots, P_{ig})$
- Range of values by parameter: $P_{ij} = (V_{ij}^1, V_{ij}^2, \ldots, V_{ij}^r)$
Optimal Parameters
PA Generation
Parameter Agent:
Learning parameters for each operator
Images
A priori knowledge

Figure 2: General framework for the proposed approach.

The UA gives to the PA the needed information: the combination of operators to apply, the set of parameters for each operator and the values' ranges for each parameter.

The PA receives this information and proceeds automatically to find the best values for each parameter. Fig. 3 shows the general functioning of the PA.

The agent PA interacts with its environment by actions, states. A set of images containing the object of interest is given to PA. Each image has its ground-truth, the object extracted by an expert.

An image with its ground truth is introduced to the system. A combination of operators to extract an object of interest is proposed. Each operator has some parameters that have to be well adjusted. Each value given to a parameter gives a different result. The agent PA must find the optimal values that give the best result. It proceeds then by trial and error until finding the best parameter values. For that it uses reinforcement learning. Actions, states and a reward function must then be defined.

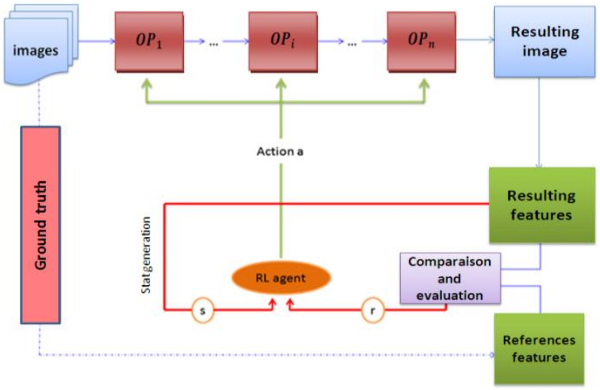


Figure 3: the general process of PA using reinforcement learning.

A. A. Defining actions

Generally, all possible combination of parameters values is defined as an action for the RL agent. The set of the actions is then the set of all possible values combination, see fig. 3.

Each operator OP_k has a series of parameters:

$$(P_1^k, P_2^k, ..., P_n^k)$$

Each parameter P_j^k has a range of values:

$$V_j^k = \{V_{j1}^k, V_{j2}^k, ..., V_{jm}^k\}$$

An elementary action of the operator OP_k is:

$$a_k = (u_{j1}^k, ..., u_{jr}^k) \text{ where } u_{j1}^k \in V_j^k$$

An action of the agent PA is defined by the combinations of the elementary actions of operators as it is defined above:

$$a = (a_1, a_2, ..., a_n)$$

Actions for object recognition task are given in the experience.

B. Defining states

A state is defined by a set of features extracted from the resulting image:

$$s = [\chi_1, \chi_2, ..., \chi_n]$$

χ_i is a feature reflecting the state of the image after the processing. The type of the extracted features depends on the task at hand. Here we give a general definition, and in the experience we define them explicitly for a recognition task.

C. Defining the reward

The return is a reward if the agent chooses the right action, else it is a punishment. The reward is defined according to the quality of the processing result. This quality is assessed by using ground-truth models. To define the return we calculate the similarity between the resulting image and its ground truth. The similarity is calculated according to some features extracted from the two images. The type of these features depends on the task at hand. For example, if we want to detect an object in an image we extract the number of the objects, their areas, their sizes, etc. We express the difference between these scalars by:

$$D = \sum_i w_i D_i$$

The weights w_i are chosen according to the importance of each feature.

A general form of the reward definition in the proposed approach is presented by:

```
Reward: r= -10, 0 or 10;
    if (D < $\varepsilon$) r = +10; f=true;
            elseif ( (D > $\varepsilon$) && (D < $\varepsilon + \delta$) )
                r = 0;
                else r = -10;
        end

        end
```

The values 10 and -10 represent respectively the reward and the punishment depending to a predefined threshold.

IV. Experience

We use a dataset of 30 textured images containing the same object to extract. The object is a textured disc injected in all the 30 images. The used images are textured so the UA proposes a combination of two operators: GLCM (Gray Level Cooccurrence Matrices) to segment textures and k-means to classify them. Each one of these operators has some parameters to be adjusted in order to be executable. In GLCM, texture is always defined in relation to some local window. The size n x n of this window affects the result of the segmentation, so we propose the size of the window as the parameter to adjust for GLCM. UA proposes a range of values for n, it may have seven values, the odd values between 9 and 21 {9, 11,…, 21}. GLCM extracts fourteen texture features [8]. In this paper we limit our self to four of the most popular features: Angular Second Moment (energy), Contrast, Correlation and entropy. After extracting these textures, we classify them using the algorithm k-means. The parameter to adjust for this operator is k, the number of the possible clusters. It can take five values {1,…,5}. How are actions, states and reward are defined according to our experience is given below.

A. Actions Definition

The UA proposes two operators: GLCM and k-means. GLCM has n the size of the sliding window as the parameter to adjust. n can take seven odd values: 9, 11, 13, …, 21, so an elementary action for GLCM is one of these values. k-means has k the number of possible cluster, its possible values are {1,2,3,4,5}. An elementary action for k-means is one of these values.

Then an action for the agent PA is constituted by a couple of a value of “n” and a value of “k”. All actions are all possible combinations of the values of “n” and “k”.

B. States Definition

States are defined, according to the features which represent the status of the resulting image. For object recognition we extract four features to define the state space:

$$S = [x_1, x_2, x_3, x_4] \qquad (1)$$

Where x_i is the selected feature.

x_1 is the Number of Objects in the resulting image after segmentation.

x_2 is the ratio between the area of the extracted object and the area of the whole image.

x_3 is the ratio between the area of the resulting object and the object reference.

x_4 is the mean of the used textural features: Angular Second Moment (energy), Contrast, Correlation and entropy.

C. *Reward Definition*

The rewards and punishments are defined according to the quality criterion that represents how well the image is segmented. A straightforward method is comparing the resulting image with its ground truth. This comparison is made between the scalar features of the obtained regions and those of the desired one. In this paper we define the reward according to a difference between the components of the image. We define this difference as: D = weighted sum of the four following differences in the two images (the resulting image and its ground truth):

D1= difference of the number of the objects;

D2= difference of the sizes of the objects;

D3= difference of the surfaces of the objects;

D4= difference of the feature textures.

$$D = \sum_i w_i D_i$$

Fig. 4 shows the three images taken randomly from the process of recognition. The three images contain the same object of interest, the textured disc.

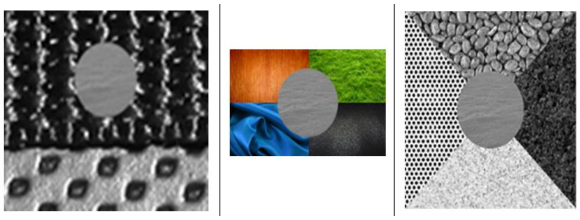


Figure 4: the three images containing the disc to extract.

The agent PA proceeds by reinforcement learning and finds that the optimal action that gives the best result is $(t_w = 13, Nb_c = 3)$. So the best value for the size of the silding window is 13 with the best number of possible clusters is 3. Fig. 5 shows the reference disc and the resulting one by our approach.

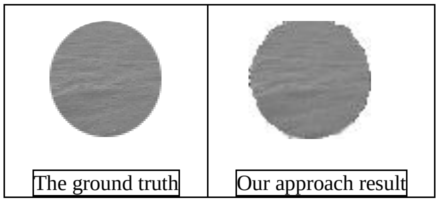


Figure 5: the resulting image and its reference.

Fig. 6 shows the curve of learning of the agent PA. At first it has not much knowledge and experience to behave, so it uses several steps per episode.

Over time the curve learning becomes almost constant, which proves that really there is a learning while the processing is done, the number of steps decreases with episodes. It means that our agent RL accumulates an experience that will help him to take decision in the future.

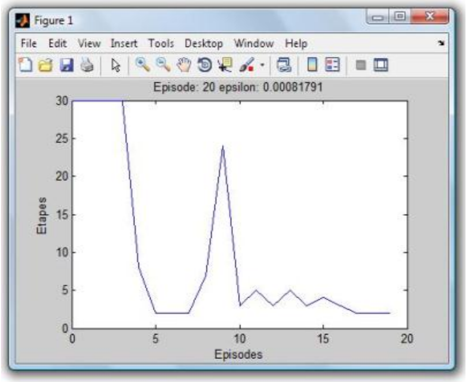


Figure 6: Learning that makes our RL agent during its processing

V. CONCLUSION

Determining the values of parameters of the vision operators is a challenging task. In this paper, we have proposed a reinforcement learning approach to handle this problem even in the case of a vision task needing many operators to sequence. A texture segmentation application is presented to test our approach.

Our goal isn't comparing our method to others, but our goal is to present another manner of thinking that uses learning concepts and show that really it gives good results. Our method can be applied to any decision process using parametric methods.

Due to the nature of reinforcement learning, the proposed approach takes in account not only the system opportunities but also the user preferences, and through the learning mechanism it suggests trustworthy solutions. As perspectives, our approach will be used on a large set of different images and its results will be compared to other methods.

REFERENCES

[1] R. Clouard, A. Elmoataz & F.Angot, "PANDORE : une bibliothèque et un environnement de programmation d'opérateurs de traitement d'images", Rapport interne du GREYC, Caen, France, Mars 1997.

[2] http://www.greyc.ensicaen.fr/~regis/ariane/

[3] B. Nickolay, B. Schneider, S.Jacob, "Parameter Optimization of an Image Processing System using Evolutionary Algorithms" CAIP 1997: 637-644.

[4] G. W.Taylor, "A Reinforcement Learning Framework for Parameter Control in Computer Vision Applications" Proceedings of the First Canadian Conference on Computer and Robot Vision (CRV'04), IEEE 2004.

[5] Sutton RS, Barto AG: "Reinforcement Learning" Cambridge, MA: MIT Press; 1998.

[6] Watkins CJCH, Dayan P: "Q-Learning". Machine Learning 1992, 8:279-292.

[7] Russell SJ, Norvig P: “Artificial intelligence: a modern approach” Englewood Cliffs, N J: Prentice Hall; 1995.

[8] Singh S, Norving P, Cohn D: “Introduction to Reinforcement Learning” Harlequin Inc; 1996.

[9] R. M. Haralick, K. Shanmugan, I. Dinstein. “Textural Features for Image Classification”. IEEE Transactions on Systems, Man and Cybernetics, Vol. 3, 1973, No 6, 610-621.

[10] F. Sahba, H. R. Tizhoosh, M. Salama. “Application of reinforcement learning for segmentation of transrectal ultrasound images” BMC Medical Imaging 2008.

An Emergency System for Succoring Children using Mobile GIS

Ayad Ghany Ismaeel
Department of Information Systems Engineering-
Erbil Technical College-Technical Education Foundation
Erbil-Iraq

***Abstract*— The large numbers of sick children in different diseases are very dreaded, and when there isn't succor at the proper time and in the type the sick child need it that makes us lose child. This paper suggested an emergency system for succoring sick child locally when he required that, and there isn't someone knows his disease. The proposed system is the first tracking system works online (24 hour in the day) but only when the sick children requiring the help using mobile GIS. In, this emergency system the child will send SMS (for easy he click one button) contains his ID and coordinates (Longitude and Latitude) via GPRS network to the web server (the child was registered previously on that server), in this step the server will locate the sick child on Google map and retrieve the child's information from the database which saved this information in registration stage, and base on these information will send succoring facility and at the same time informing the hospital, his parents, doctor, etc. about that emergency case of the child using the SMS mode through GPRS network again. The design and implement of the proposed system shows more effective cost than other systems because it used a minimum configuration (hardware and software) and works in economic mode.**

Keywords- GPS; GPRS; Mobile GIS; SMS; Tracking device; Emergency System.

I. INTRODUCTION

Increasing the rating of sick children in different diseases per year base on World Health Organization WHO and UNICEF, e.g. over the last 40 years asthma particularly in children approximately 300 million people worldwide currently have asthma and its prevalence increases by 50% every decade, in North America, 10% of the population has asthma [1].

Other disease is congenital heart defect over 1,000,000 babies born with this disease worldwide each year 100000 of them will die within the first year, i.e. one of every 100 infants they have congenital heart defect to some extent [2]. There are 70,000 children (less than 1 year -14 years) worldwide are expected to develop type 1 diabetes annually per year increase is estimated at around 3% [3].

From what advancement the sociality face to face a front of a big problem with this large numbers of sick children from view of three important diseases only, i.e. what about other diseases this problem become more effective if there isn't who offer succoring and helping at a suitable time for those children when they are in school, shopping, with their friends, etc, i.e. will lose a large number from those children if they needed succoring and there isn't someone with the child known which type and suitable help he needed.

To solve this problem sure will think about new track system for those sick children to succor them, not like these traditional tracking systems which are developed so far use a handheld GPS receiver device for tracking the location depend on real time tracking and continuity on the interval of tracking [4] for example, of these types of traditional tracking systems; may construct from n-tier as shown in Fig. 1 [6].

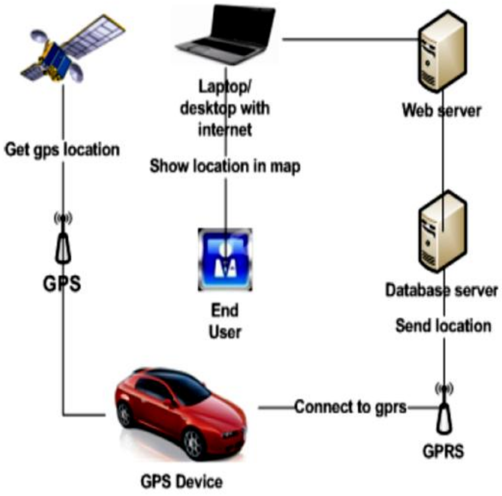


Figure 1. The N-Tier Tracking System Diagram [6].

Really there is needed for an emergency system can tracking the sick children only when they required help and there are no one knows what is the disease of each of them, at a suitable time and in the quality which the children are required using new techniques and modes to satisfy system in effective cost.

II. RELATED WORK

Katina Michael and others [2006] employed usability context analyses to draw out the emerging ethical concerns facing current human-centric GPS applications personal locators for children, the elderly or those suffering from Alzheimer’s or memory loss, and monitoring of parolees for law enforcement, security or personal protection purposes. The outcome of the study is the classification of the current state GPS applications into the contexts of control, convenience, and care; and a preliminary ethical framework for considering the viability of GPS location-based services emphasizing privacy, accuracy, property and accessibility [4].

Alahakone, A.U. and Veera Ragavan [2009] presented the development of a geospatial information system for path planning and navigation of mobile objects, The system involves a GIS implemented using Google maps to visualize the routes of mobile objects acquired from GPS receivers over a GPRS network [5].

Ruchika and BVR [2011] proposed a cost effective method of tracking a human's mobility using two technologies via GPRS and GPS, and further the cost is reduced by using GPRS rather than using SMS for communicating the information to the server, but the tracking system is design based on Android only, not for any mobile phone (general) can support GPS and General Packet Radio Service GPRS like iPhone, windows phone, iPad, etc [6].

The whole systems allow the user's mobility to be tracked using a mobile phone which is equipped with an internal GPS receiver and a GPRS transmitter, i.e. most of the applications developed so far use a handheld GPS receiver device for tracking the location [6], and real time tracking and continuity on the interval of tracking may be very high cost specially when the server, IP network, and ISP are busy in the interval of tracking [4].

To overcome the problems above, an emergency system must be contain the following techniques and modes:

A. Mobile GIS:

As expansion of GIS technology from the office into the field, a mobile GIS enables field-base personnel to capture, store, update, manipulate, analyze, and display geographic information. Mobile GIS integrates one or more of the following technologies:

1) Mobile devices.
2) Global Position System (GPS).
3) Wireless communications for Internet GIS access.

There is wide using of mobile GIS to complete the multiple tasks one of them the tracking for persons, vehicle, etc [7].

B. An Emergency System Works Via GPRS:

GPRS is the widely acknowledged successful application to adopt; it cannot be denied that high costs are involved in both setting up as well as maintenance of the application. Also, in order to full integrate and fully tap upon the efficiency of the system, and

C. Short Message Service SMS:

The mobile terminal sends data through SMS to the receiving terminal, compare to the modem solution the SMS solution is more economical because the tracking system will work in an emergency cases only (when really the child needs the succoring and help), i.e. to overcome the time of tracking systems in general, which are used to maintain long time continual tracking system it would therefore, result again in a high cost in maintaining a continual tracking system.

III. The Motivating

The objective of this paper reach to an emergency tracking system offers succoring (when the children needed the help), i.e. make the system works really only when the sick child required the succoring and in economic mode that will reduce the time and delay as well as the cost, to satisfy this aim must be thinking about new techniques and modes, so the proposed of an emergency system will involve the following characteristics:

A. The Mobile GIS Technique:

The system involves a GIS implemented using Google maps to visualize the location of mobile or track device for a sick child without needing to use GPS receiver, but the proposed system will base on supporting of a mobile build-in GPS technique on devices like iPhone, Windows phone, iPad, etc.

B. Modes Of Transmission As Follow:

1) *GPRS mode:* The mobile terminal sends data through GPRS data channel to a special TCP/IP server linked to the Internet, or a PC with a fixed Internet IP address. In this case, the GPRS is always online and billed only on the bytes transmitted, rendering it to be a much cheaper alternative to any current systems, and

2) *SMS mode:* The mobile terminal sends data through SMS to the receiving terminal, compared to the modem solution; the SMS solution is more economical.

C. Server of TCP/IP mode:

Have a fixed IP address, as well as a dynamic IP address as soon as the receiving server or receiving terminal get its IP address on boot up, the user will need only to reconfigure the IP address setting of the mobile terminal to align it to any of the users desired output this remote setting mechanism makes all this possible without any hassle.

IV. Architecture of Proposed an Emeregncy System

The main tasks for the suggested design of an emergency system to succor the sick children summarize in the flowchart as show in Fig. 2.

The architecture of suggested an emergency system involves multiple modules as shown in Fig. 3, these modules are:

A. Registration module:

This module referring to any sick child needed to be serve using this an emergency system must be register via web interface constructed for the system one time only (no duplicate) and saved the child's information in a database created for registration. Without registration, this system can't recognize the child which needs succor/help.

GIS module:

In this module the sick child will use his mobile (a child who is equal or less than 15 year), this mobile can support GPS technique (built-in) to locate his coordinates (Latitude and Longitude) or using tracking device like GlobalSat TR-203 (a child who can't use mobile). Every time the child need succoring from this an emergency system will send SMS contains only the coordinates of location and child ID (number of child's mobile or his sequence number in database, etc) which specified in registration module (A above) using GPRS network, and this SMS will be received by the web server of an emergency system.

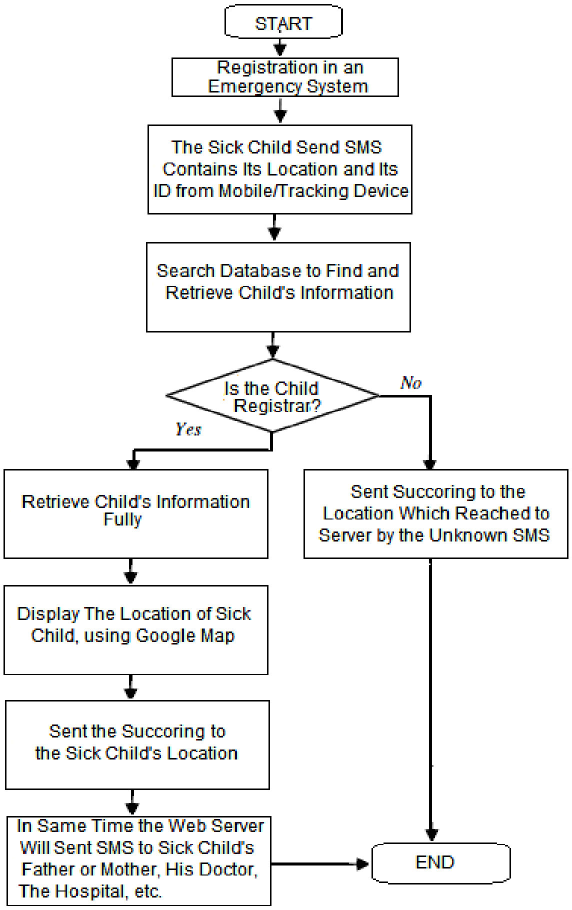

START
Registration in an Emergency System
The Sick Child Send SMS Contains Its Location and Its ID from Mobile/Tracking Device
Search Database to Find and Retrieve Child's Information
Is the Child Registrar?
Yes
No
Retrieve Child's Information Fully
Sent Succoring to the Location Which Reached to Server by the Unknown SMS
Display The Location of Sick Child, using Google Map
Sent the Succoring to the Sick Child's Location
In Same Time the Web Server Will Sent SMS to Sick Child's Father or Mother, His Doctor, The Hospital, etc.
END

Figure 2. Flowchart for the main tasks of proposed an emergency system.

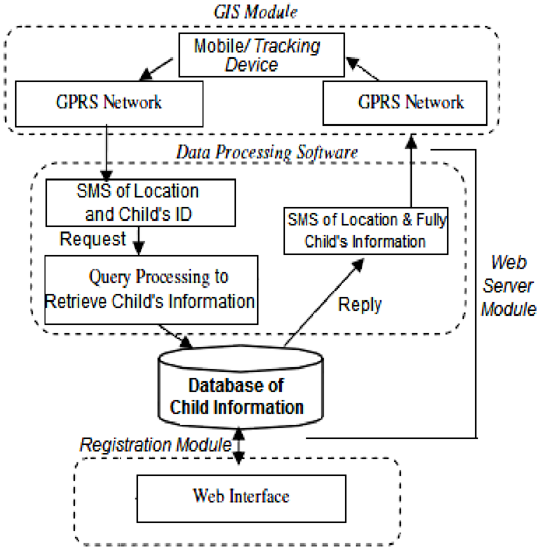

GIS Module
Mobile/ Tracking Device
GPRS Network
GPRS Network
Data Processing Software
SMS of Location and Child's ID
Request
SMS of Location & Fully Child's Information
Query Processing to Retrieve Child's Information
Reply
Web Server Module
Database of Child Information
Registration Module
Web Interface

Figure 3. Architecture of proposed an emergency system.

C. Web server module:

That module will work when an emergency system received SMS, the web server will use the Child ID within received SMS to search the database to find and retrieve the fully information of sick child, then send succoring facility (Car, Helicopter, Lifeboat, etc), and at the same time send SMS to informing the emergency hospital, the father or mother of sick child, etc.

Fig. 4 shows a diagram of serving or supporting sick child which is done in two stages by the proposed an emergency system, the first stage start when the child request the succoring/help from an emergency system, i.e. the serving start base on request of sick child, so not like a natural tracking system which was worked in continuity, while the proposed an emergency system offers serve or support for a sick child (at any time and online), but the real serve start when SMS of request reached from child to the web server as shown in Fig. 4; a.

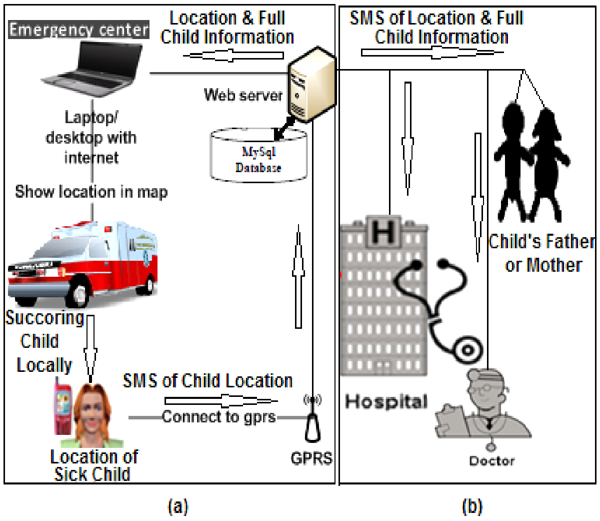


Figure 4. Serving/Supporting diagram of proposed an emergency system.

This feature will reduce and minimize the cost because the proposed system will need only web server based on a database not two servers one for web server and another for database, again this feature (non continuity of emergency system) will use SMS solution which is more economical because an emergency system works when the child needed not like the traditional tracking systems which were continuous, Fig. 4; b shows the second stage of serving an emergency system starting when retrieve the fully information of childlike name, type of diseases, father or mother, his doctor, etc the technique of finding and retrieving the child's information written in the flowchart as shown in Fig. 5.

V. Experimental Results

Implementing the proposed design of an emergency system reveals below:

A. The requirement of configuration

The configuration for suggested system in this research can be divided into:

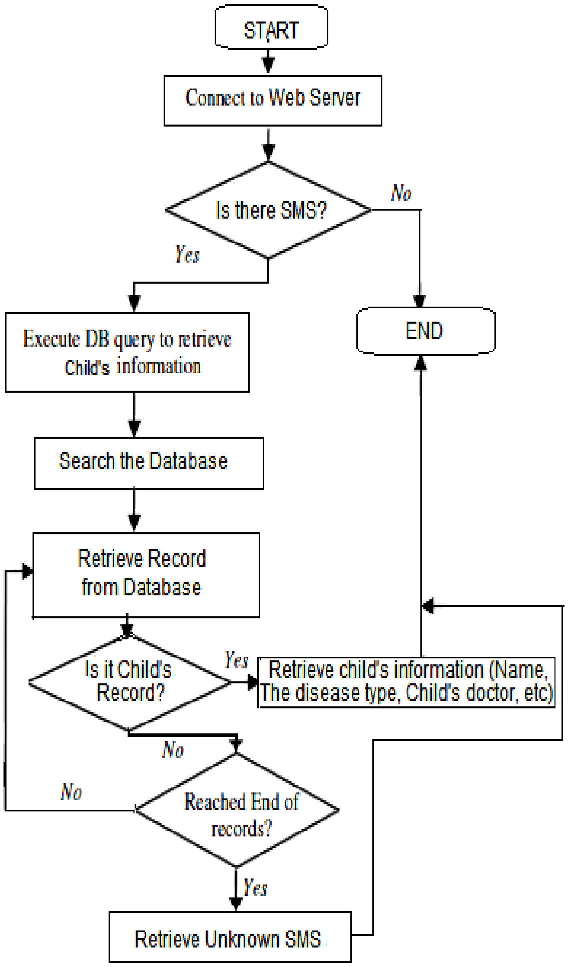

START
Connect to Web Server
Is there SMS?
No
Yes
END
Execute DB query to retrieve Child's information
Search the Database
Retrieve Record from Database
Is it Child's Record?
Yes
Retrieve child's information (Name, The disease type, Child's doctor, etc)
No
No
Reached End of records?
Yes
Retrieve Unknown SMS

Figure 5. Flowchart technique of finding and retrieving child's information.

1) *The Software:* Tools which are needed windows server 2008 to setup and install web server, the other important software C# as programming language under Visual studio 2010 package which is used in the web interface for registration and for implement the all other techniques like the connection between the web server and database, then the search (find and retrieve) technique information of sick child, etc. the last software needed MySQL package use to construct the database involve table call info contains the fields as shown in Table 1.

TABLE I. FIELDS OF INFO TABLE

Field	Type
ChildId	Int(10)
Name	Varchar(20)
Age	Varchar(7)
Father-No	Int(10)
Mother-No	Int(10)
Disease-Name	Varchar(20)

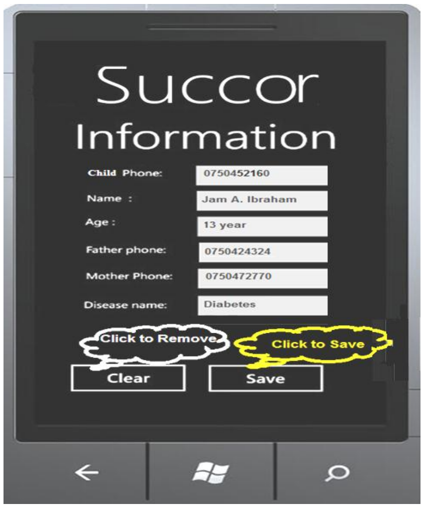
2) *The Hardware:* As the first device will need a server friendly with windows and Microsoft packages selected type HP server, second need a mobile can support GPS technique (built-in), here selected windows phone which is friendly with Microsoft packages (the compatibility make avoid conflict) and cheap comparing to iPhone, iPad and tracking device (which can support GPS/GSM/GPRS technology like GlobalSat TR-203) as shown in Table 2. Finally, will need a resource to connect with Internet Service Provider ISP cross GPRS network.

TABLE III. COMPARING THE WINDOWS PHONE WITH OTHER TYPES OF MOBILES AND TRACKING DEVICE

Feature	Windows Phone	iPhone	Tracking Device (TR-203)
Cost	Cheap	Expensive	Relatively Expensive
Zone	Unlimited	Unlimited	Limited
Multiple usage	Yes	Yes	No

B. Implement the proposed System

For An emergency system, which called Succor will select environment for implementation Erbil city and the first step are the registration of the child using GUI (web interface) from any PC connected to internet or from his mobile directly as shown in Fig. 6.



Child Phone: 0750452160
Name : Jam A. Ibraham
Age : 13 year
Father phone: 0750424324
Mother Phone: 0750472770
Disease name: Diabetes

Figure 6. Registration of sick child in Emergency system using his mobile.

After registration the Succor system can offers serving/support to the sick child when request that by click one button (e.g. help button) and the location is determined automatically (Longitude and Latitude) by windows phone which support GPS functions as shown in Fig. 7 then the mobile will send SMS containing the coordinates and ChildId (phone number) to the web server.

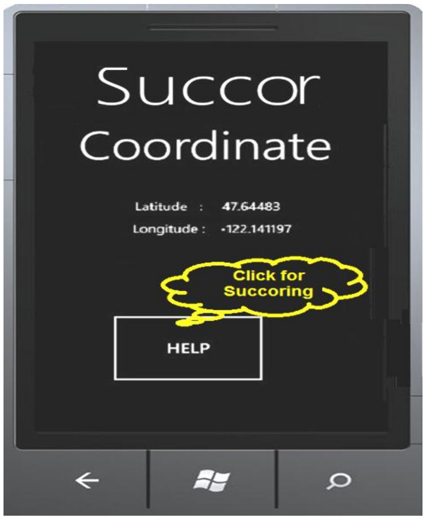
The web server must be done a sequence of tasks as shown in Fig. 8:

1) *Search:* The web server will search the SMSs of requiring a succouring online/automatically or manually by click search button and these SMSs appear in the description location (see Fig.8).

2) *Find:* this button for finding fully information using ChildId (as key for finding) from the database of Succor system and appearing at information location as shown in Fig.

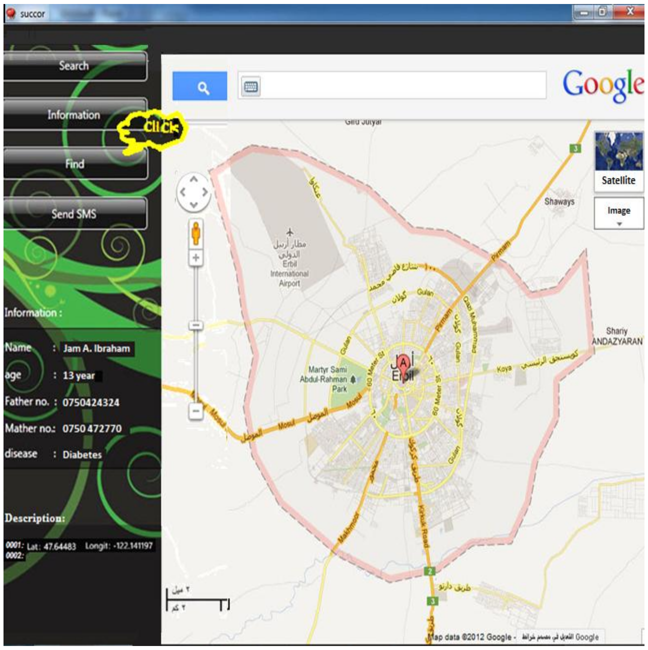
8, at the same time will determine the real location of sick child on Google maps.

3) Send SMS: the web server after sending succoring facility (e.g. car, helicopter, etc) to the real location of child, this button will use to send SMSs for father/mother of child (or any other persons from child's Family), and also to the emergency hospital (to become ready for receiving child), see Fig. 8.



Latitude : 47.64483
Longitude : -122.141197

Figure 7. The sick child when need serve from succour (emergency) system.



Information :

Name : Jam A. Ibraham
age : 13 year
Father no. : 0750424324
Mather no.: 0750 472770
disease : Diabetes

Description:

0001: Lat: 47.64483 Longit: -122.141197
0002:

Figure 8. The tasks/buttons of web server of succouring system.

C. *Discussion of the results:*

The important results of the proposed an emergency system comparing with other systems shown in Table 3.

TABLE IIIII. COMPARING THE PROPOSED SYSTEM WITH OTHER SYSTEMS

Feature	**Proposed an Emergency System**	**Ruchika and BVR System**	**Alahakone and Veera System**
Request of Hardware	Relatively minimum request, e.g. no need extra GSM network, separated database server, etc	Relative middle request	Relatively maximum request
Serving continuity or at the request	At the request however, the serving is online 24 hours	Continuity	Continuity
Using SMS	Yes, it becomes economic with noncontinuity	No	Yes
Need the support of GPS receiver	No, because it supported by build-in GPS	Yes	Yes
Using specific mobile	No (General), which are support GPS	Yes	No

V. CONCLUSIONS AND FUTURE WORK

A. *Conclusions*

The conclusions which are obtained from proposed system summarized as follow:

1) The design of proposed an emergency system is an effective cost more than other systems as referring to in Table 2; especially the proposed system is better than Ruchika and BVR system which is characterized based on cost effective [6].

2) The proposed system really works when the sick child send SMS, i.e. an emergency system is not continuity however, it works online and can accept any request at any time. This approach of succoring which is used the SMS mode relatively economic from other modes like modem solution.

3) This an emergency system can use as tracker for a child or any other type of ages whose make the registration or not in case of unregistered will send the succoring facility without any others information to the location from it the server received the SMS.

B. *Future work*

Can extend the proposed system to achieve the follow tasks:

1) Make the an emergency system can do more GIS computing to determine the real location and distance (base on an update image satellite for the zone or location), this distance computed between an emergence center (which send the succoring facility) and the location of sick child as well as between the child's location and emergency hospital which take care of child, and then base on real location and the optimal path (distance) will decide to send a suitable type of succoring facility (helicopter, car, lifeboat, etc) and in the shortest path because selecting the suitable succoring facility and in shortest path are playing an important role in succoring the sick child.

Involve an emergency system knowledge about all emergency hospitals (e.g. by extended or added database) in the zone or location to inform the succoring facility to take the correct direction to the suitable hospital (proper to the type of the child's disease) to serve and help the child in a suitable an emergency hospital and at the shortest time (without any loss of time in this emergency case of child).

ACKNOWLEDGMENT

Thanks to the team members (my students), who worked on my idea (project of succor group), and I was their mentor in competition of Imagine Cup for Microsoft in 2012.

REFERENCES

[1] Sidney S. Braman, MD, FCCP, "The Global Burden of Asthma", © 2006 Official journal of the American College of Chest Physicians, DOI 10.1378/chest.130.1_suppl.4S, downloaded from chestjournal.chestpubs.org by guest on June 9, 2012, http://chestjournal.chestpubs.org/content/130/1_suppl/4S.full.html.

[2] Todd L. Lowary, Jillian M. Buriak, Lori J. West, "Infant Heart Transplants and Nanotechnology", Faculty of Medicine and Dentistry, University of Alberta, 2009.

[3] Type 1 diabetes incidence, http:// type 1 diabetes incidence.html.

[4] Katina Michael, Andrew McNamee, and MG Michael,"The Emerging Ethics of Humancentric GPS Tracking and Monitoring", Faculty of Informatics, University of Wollongong, Australia [2006], http://ro.uow.edu.au/infopapers/385.

[5] Alahakone, A.U. "Geospatial Information System for tracking and navigation of mobile objects ", **Page(s):** 875 – 880, paper appears in Advanced Intelligent Mechatronics, 2009. AIM 2009. IEEE/ASME International Conference on 14-17 July 2009.

[6] Ruchika Gupta and BVR Reddy, " GPS and GPRS Based Cost Effective Human Tracking System Using Mobile Phones ", VIEWPOINT Volume 2 • No. 1 • January-June 2011.

[7] ESRI, "GIS BEST PRACTICES", ESRI • 380 New York Street • Redlands, Pages 1-57, Copyright © May 2007, WWW.ESRI.COM/MOBILEGIS.

AUTHOR PROFILE

Ayad Ghany Ismaeel received MSC in computer science from the National Center of Computers NCC- Institute of Postgraduate Studies, Baghdad-Iraq at 1987, and Ph.D. computer science in qualification of computer and IP network from University of Technology, Baghdad- Iraq at 2006.

He is professor assistant, at 2003 and currently in department of Information Systems Engineering in Erbil Technical College-Iraq, His research interest in mobile, IP networks, Web application, GPS, GIS techniques, distributed systems and distributed databases. He is lecturer in postgraduate of few universities in MSC and Ph.D. courses in computer science and software engineering from 2007 till now in Kurdistan-Region, IRAQ.

Ayad Ghany Ismaeel is Editorial Board Member at International Journal of Distributed and Parallel Systems IJDPS http://airccse.org/journal/ijdps/editorial.html, Program Committee Member of conferences related to AIRCC worldwide, and reviewer in IJCNC (which is listed as per the Australian ARC journal ranking http://www.arc.gov.au/era/era_2012/era_journal_list.htm), IJDPS, IJCSIT journals (http://airccse.org/journal.html), and Conference CCSEIT-2012 within http://airccse.org/, as well as he adviser and reviewer in multiple national journals. The last published papers were in International Journal Distributed and Parallel System (IJDPS) as follow:
NEW TECHNIQUE FOR PROPOSING NETWORK'S TOPOLOGY USING GPS AND GIS, published in (IJDPS) Vol.3, No.2, March 2012. http://airccse.org/journal/ijdps/papers/0312ijdps05.pdf; NEW METHOD OF MEASURING TCP PERFORMANCE OF IP NETWORK USING BIO-COMPUTING, Published In (IJDPS) Vol.3, No.3, May 2012. Http://Airccse.Org/Journal/Ijd

Billing System Design Based on Internet Environment

Muzhir Shaban Al-Ani
Collage of Computer Science
Anbar Uiversity
Anbar, Iraq

Rabah Noory
Collage of Computer Science
Anbar Uiversity
Anbar, Iraq

Dua'a Yaseen Al-Ani
Collage of Computer Science
Anbar Uiversity
Anbar, Iraq

Abstract— **This paper deals with the design of Internet billing system, in which it is possible pay invoices electronically. This approach is implemented via virtual banks, in which the process of money transfer can be implemented. In other hand many applications can be realize such as; deposit e-money, withdrawal e-money and determine account balance. A Gate way translator is used to apply authentication rules, security and privacy.**

Keywords- Billing System, Internet Billing system , E-Commerce, E-bank, bill payment, Authentication, Security.

I. Introduction

Paper bills are now the primary channel of communication between companies and their customers. However, their potential for personalization is limited, and they are not interactive. If a customer wants to react to something in his paper bill – for example, to make a customer service inquiry or to order a new service – he must make a telephone call. Internet Billing promises far more than a new and inexpensive way to deliver billing information. Industry experts predict that Internet Billing will fundamentally change the way companies interact with their customers. Eventually, the Internet Bill will be an interactive entry to a host of additional services including customer self-care, automated sales one-to-one marketing. The Internet Bill will become the gateway through which customers and companies have electronic one to one dialogs[1].

Businesses and consumers are banking on the Internet in more than one sense. Despite the early proliferation of electronic banking applications on private networks through dial-up services, most electronic banking applications have migrated to the Internet. Consumers will not be tied to one particular bank and its software, nor to a single terminal where the bank's own software must be installed. Banking on the Internet provides the flexibility of banking from any Internet access terminal using the now ubiquitous Web browser. Banking on the Internet can reduce the number of staff banks must maintain without having to make the investment in establishing private networks. The World Wide Web, or the Web, and its user-friendly, graphically rich browsers have made the Internet both friendly and accessible to the common desktop user at home and in the office [2].

The advancement of electronic banking or commonly known as e-banking, began with the use of ATMs and has included telephone banking, Direct bill payment, electronic fund transfer, online banking and other electronic transactions[3].

Banking services offered to consumers over the Internet will allow consumers to generate bank statements, check balances, transfer money between accounts, and authorize fund transfers to deposit money, to pay monthly bills, and to write personal checks. The Internet will provide a very competitive medium for banks to woo consumers. Consumers will be able to quickly and easily scan savings and loan rates and banking fees without having to interact with bank personnel.

Beyond home banking, consumers will be able to write electronic checks to online merchants that draw value directly from the consumer's own bank account rather than use a line of credit. The Internet will make banking a much more competitive environment in another critical aspect. Local banks will now be competing with national and international banks whose Internet presence removes barriers of physical distance. In addition, a number of "virtual" banks have now entered the market to compete with traditional banks for clients. The environment created by

Internet banking will present the vast array of services currently offered by banks in a form that is very convenient to consumers Commerce [2].

II. E-Commerce

Deep penetration and spread of Internet, lead to more electronic applications are becoming available. Electronic commerce (E-commerce) is one such enabling technology, which has wide spread utility touching almost everybody in society. It helps buyers and sellers, individuals and business, retail and bulk suppliers. In fact, e-commerce has very attractive features like anywhere, anytime shopping / banking (24 hours x 365 days) and no holidays, zero inventory, no middlemen, and so forth.

It helps customers to compare various products in the range and class, study their features/performance and make an informed decision about the emergence of e-commerce has created new financial needs that in many cases cannot be effectively fulfilled by the traditional payment systems. Recognizing this, virtually all interested parties are exploring various types of electronic payment (E-payment) system and issues surrounding e- payment system and digital currency [5].

The earliest example of e-commerce is Electronic Funds Transfer (EFT). This allows financial institutions to transfer funds between one another in a secure and efficient manner. Later, Electronic Data Interchange (EDI) was introduced to facilitate inter business transactions. However, early EDI systems were typically operated over special networks that are complex to set up and costly to administer. For these reasons, EDI has not been as widely deployed as expected. With the advent of Internet technologies and advanced cryptographic techniques, it is now feasible to implement e-commerce over a public network – the Internet. The development of the World Wide Web (WWW) greatly accelerates the development of e-commerce and expands its scope to cover different types of applications [6].

E-commerce includes activities such as establishing a Web page to support investor relations. In brief, e-commerce involves the use of information technology to enhance communications and transactions with all of an organization's stakeholders. Such stakeholders include customers, suppliers, government regulators, financial institutions, mangers, employees, and the public at large. E-commerce is a revolution in business practices.

If organizations are going to take advantage of new Internet technologies, then they must take a strategic perspective. That is, care must be taken to make a close link between corporate strategy and e-commerce strategy. E-commerce, in a broad sense, is the use of computer networks to improve organizational performance. Increasing profitability, gaining market share, improving customer service, and delivering products faster are some of the organizational performance gains possible with e-commerce. E-commerce is more than ordering goods from an on-line catalog. It involves all aspects of an organization's electronic interactions with its stakeholders, the people who determine the future of the organization [7].

There are different types of e-commerce from perspective of the buyer and seller relationship, according to this relationship, e-commerce applications can be divided into the following four categories:

A. *Business-to-Consumer (B2C)*

In this case, the seller is business organization, whereas the buyer is consumer. This emulates the situation of physical retailing and so it is commonly called electronic retailing or consumer-oriented e-commerce. It frequently involves a temporary relationship and has relatively low volume of transactions and small payments.

B. *Business-to-Business (B2B)*

In this case, the vendor and the buyer of the goods or services involved in a transaction are both organizations rather than individual customers.

In contrast to B2C e-commerce, B2B is characterized by a number of features and these include high volumes of goods trade, prior agreements or contracts between the partners involved requiring a much higher level of authorization, taxation, and documentation and information exchange.

C. *Consumer-to-Consumer (C2C)*

This refers to situations where both the seller and the buyer are consumers. On line auctions provide an effective means for supporting C2C e-commerce.

D. *Consumer-to-Business (C2B)*

This has perhaps been the area in which there has been the biggest growth in e-commerce. In this type of applications, a customer specifies his requirements in relation to the product or services he wants to the business represented by an e-commerce site which does a search over the Internet to explore the web sites that match these requirements and return the result to the customer [8].

III. BILLING SYSTEM

Billing systems are key competitive weapons for telecommunications companies [9]. A billing system is a combination of software and hardware that receives call detail and service usage information, grouping this information for specific accounts or customers, produces invoices, creating reports for management, and recording (posting) payments made to customer accounts. Billing systems are composed of interfaces (Network, Marketing, Customer Care, Finance, etc.), computers, software programs and databases of information. Computers are the hardware (computer servers) and operating systems are used to run the programs and process. Network interfaces are the hardware devices that gather accounting information (usage) from multiple networks, convert it into detailed billing records, and pass it on to the billing system.

Billing system use databases to hold customer information; usage call detail records, rate tables, and billing records that is ready to be invoiced. The key functional parts of a billing system include creating usage records, event processing, bill calculation, customer care, payment processing, bill rendering and management reporting. In addition to the basic billing system functions, billing systems share information with many other business functions such as sales, marketing, customer care, finance and operations.

Billing charges are determined by events that occur in a communication system. Billing events can originate from many sources: a media gateway, a media server, a content aggregator or a visited partner’s network and they must be converted into a standard format. A typical billing process involves collecting usage information from network equipment (such as media servers, access devices and set top boxes), translating and formatting the usage information into records that a billing system can understand, transferring these records to the billing system, assigning charge fees to each event, creating invoices, receiving and recording payments from the customers[10].

Telecommunication companies need an effective and accurate billing system to be able to assure their revenue. Billing systems process the usage of network equipment that is used during the service usage into a single Call Detail Record (CDR). The billing process involves receiving billing records from various networks, determining the billing rates associated with the billing records, calculating the cost for each billing record, aggregating these records periodically to generate invoices, sending invoices to the customer, and collecting payments received from the customer. Billing system is very complex starting from network elements that generate usage to the billing system to usage collection, mediation, rating, and invoicing [11].

IV. On-Line Billing System

Electronic billing is one of the fastest growing technologies for corporate law departments. Recent surveys indicate that roughly 15 percent of corporate legal departments require electronic bills from their law firms, and another 15 percent are considering it. If the person is a law firm with corporate clients, the person have probably seen acceleration in the number of requests from clients who want their bills submitted electronically. Choosing electronic billing and matter management systems are among the most important technology decisions that a law department can make, with significant potential consequences both positive and negative [12].

The concept of electronic billing is not new. Since the advent of the Internet, a small number of consumers have been using this electronic medium to pay bills online after receiving standard paper invoices via regular Postal Service. What is new in the electronic billing arena is the concept of electronic bill presentment. With electronic bill presentment, companies that send bills (billers) post consumers' statements to the Internet, enabling consumers to view the statements and make e-payments [1].

With ever increasing spread of Internet, Bill presentment and payment is becoming a new type of service area for periodic billers like Telephone Companies, Electricity etc. Internet based bill presentment and payment system converts billing centers from cost centers to revenue centers and for customers (payer) the system is a personalized service. Internet based bill presentment and payment system provides direct personalized communication channel between Billers and Payers, opens a new revenue channel by cross-selling advertisements. Drastic reduction of costs that are associated with paper based billing system. For customers or payers, receiving bills to payment of bills at one window through a Personal Computer, figure1 show on-line billing system.

Figure 2 Depicts overall workflow of the system step by step:

1. Customer gets an Electronic Cheque Book (e-Cheque Book) from his/her bank.

2. Customer sends registration request for online billing through biller's World Wide Web site.

3. Biller verifies credentials of the application and grants a subscription for online billing and sends user-id and password through e-mail or immediately when credentials are submitted. This enables the customer to view and pay bills.

4. Customer logs in to his/her online billing account of the biller's web site, verifies the bill details and pays with an electronic cheque (generated from the e-Cheque Book). The electronic cheque or e-cheque is sent to the Biller.

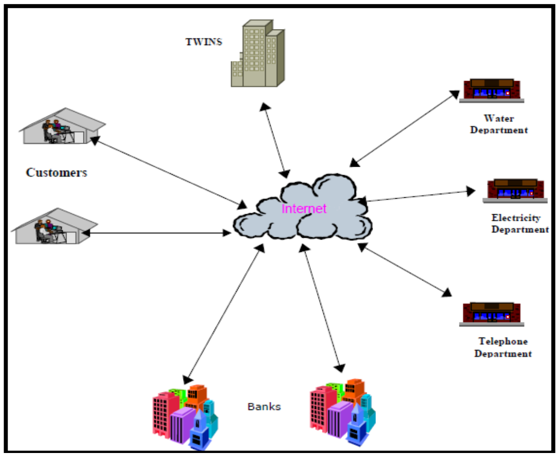


Figure1. The entities involved in on-line billing system

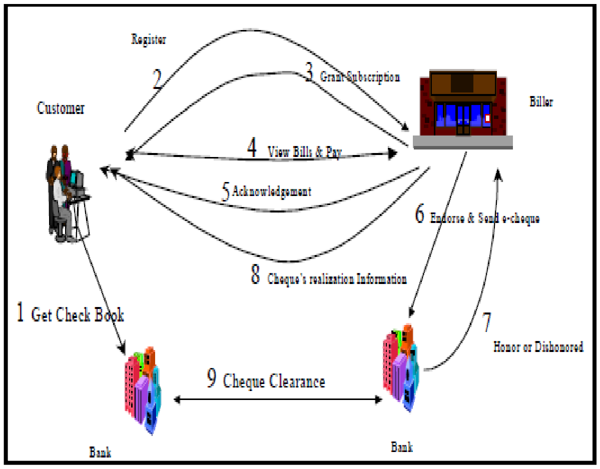


Figure 2: Internet billing system workflow

5. Biller system receives e-cheque and sends an acknowledgement for the received e-cheque.

6. Biller checks for validity of electronic cheques (received from customers), endorses, and sends them to bank for processing.

7. Bank (Biller's) validates the received e-cheques and sends them to the Customer's bank for clearance. Honored or dishonored information is sent to the Biller.

8. Biller's billing system updates the customer billing data based on cheque clearance status (as received from the bank) and sends appropriate information to the customer through electronic mail. The steps 1-3 above are done for registration, which is a one-time activity for a given customer, whereas steps 4-8 are used for viewing/paying bills, which is an on-going activity [4].

V. Literature Survey

There are many previous studies in the field of Internet billing system, below are some of these studies and their result are referred to:

- **J Crookes (1996)**, the term adopted for the system is multiservice billing system (MSBS). The strategic business issues which have shaped the design of

MSBS. It describes the scale and complexity of the problem which makes the construction of a multiservice platform such a difficult feat of software engineering. The concept of a common product model, which underpins the system's design, is introduced [9].

- **NN Murthy, et al.(2000),** In their paper, the authors presented a brief description of the technologies for e-commerce The authors also present TWINS (Twin Cities Information Network Service) test-bed application being developed as part of this project. TWINS, operational at twin cities of Hyderabad-Secunderabad, facilitates payment of various utility bill payment (like water, electricity, etc.) through a single window system. Payment of water bills through Internet using E-Cheque (Electronic Cheque) will be operational soon. This enables customers to pay their bills from anywhere, anytime. Thus, realizing the benefits of e-commerce to the citizens [4].
- **Yang Bo, Liu Dongsu and Wang Yumin (2001)**, In their paper, the authors improved the e-payment system with a smart card proposed by S.Brands, and present an anonymity-revoking e-payment system. On the one hand, the customer‘s privacy cannot be compromised by the bank or by the payee. On the other hand, anonymity can be removed by a TTP with the help of the bank. In this case, the third party can link a payment to a corresponding withdrawal and prevent money laundering and blackmailing [13].
- **EWB Team (2000),** This document provided information regarding the use of the Extra Work Billing System (EWB). The document is organized with step-by-step instructions for each task to be accomplished using the EWB system. The EWB System may be accessed through the Internet using either Netscape Navigator or Internet Explorer [14].
- **P.S. Barreto, et al. (2005).** In their paper, the authors presented a discussion concerning the performance of four network scenarios for billing purposes. Using the results of packet losses in an experimental platform simulating a NGN (Next Generation Network) environment, the authors evaluate on each scenario the impact in the billing process with different traffic flows comparing the total revenue calculus for two billing schemes: (1) charging per packet and (2) reducing the value corresponding to undelivered packets. Our results show that the environments that use Differentiated Services are both convenient for costumers and service providers [15].
- **Shiqun Li , et al. (2008).** In their paper, the authors first identified some vulnerability in the mobile billing system. Then, the authors propose a fair and secure billing system based on a proper combination of digital signature and hash chain mechanism. The proposed system can achieve authentication, non-repudiation, and fairness, which are desirable security requirements for an undeniable mobile billing system. [16]**.**
- **Albert Levi, Cetin Kaya Koc (2009),** In their paper the authors proposed a new Internet e-payment protocol, namely CONSEPP (Convenient and Secure E-Payment Protocol), based on the account authority model of ANSI (American National Standards Institute) X9.59 standard. CONSEPP is the specialized version of X9.59 for Internet transactions (X9.59 is multi-purpose). In CONSEPP the authors propose a lightweight method to avoid the need for merchant certificates. Moreover, the authors propose a simple method for secure shopping experience between merchant and consumer. Merchant authentication is embedded in the payment cycle. CONSEPP aims to use current financial transaction networks, like Visa Net, Bank Net and ACH (Automated Clearing House) networks, for communications among financial institutions. No certificates (in the classical sense) or certificate authorities exist in CONSEPP [17].
- **Giannakos Antoniou, et al. (2009)**, In their paper, the authors proposed an online payment scheme which uses the traditional e-payment infrastructure but which reveals no payment information to the seller. This is done with only an incremental increase in computational power [18].

VI. IMPLEMENTED BILLING SYSTEM

The implemented system is intended to support all the banking operations (Direct bill payment; determine account balance, money transferred, withdrawal and deposit). Figure 3 shows the architecture of the implementation system.

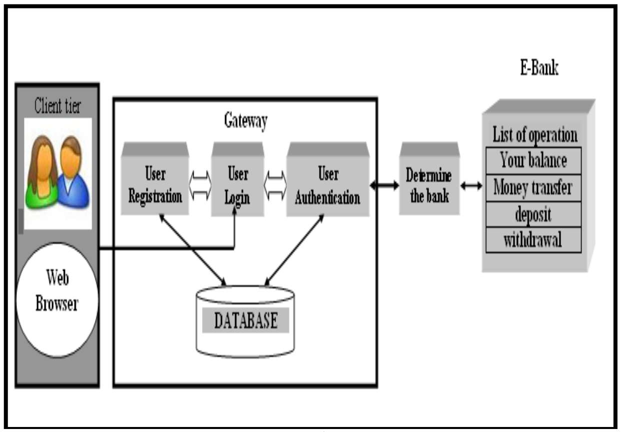


Figure 3: Architecture of the implemented system.

The implementation system consists of the following components:

A. Client

Clients (i.e. users and customers) can access to the implemented system from web browser (Internet explorer, firefox, etc) that provides a user interface that customer interacts this interface. The user must enters the name and the password to introduce to e-bank group .

B. Gateway

Gateway is used to control several operations including user access to the system, check the validity and reliability of the user from use of the system and make sure that the user is one of the participants in the e-bank group. In addition make the registration process in the e- bank group for new customer. After all this, the user can log into the e- bank group. The gateway must authenticate the user before allowing him to enter any bank want to deal with it. User's authentication required the correct user name and password that must be entered at the login step. They will be checked against the stored ones in designated database, upon match, the user will grant the access to its account, hence, the provided services. Passwords will be stored at the designated database as plain text; an unauthorized access from local or remote user to the database can have an catastrophic damage.

She/he can use any username along with correct password to access, transfer, etc…. Therefore, Hash function MD5 (Message-Digest Algorithm) is used to generate message digest for all the passwords that will be stored in the database. At the authentication, the Hashed password will be checked rather than the plain text.

In the case that the user is a new customer, the information must enter which are full name, phone, E-mail, address, username, password and limitation. The limitation is that the amount of money determined by the user in the registration process, so the user cannot exceed this amount during withdraws or transfer funds. This mechanism used to add more protection for the process of withdrawal and transfer of funds. In case of exceeding the limitation specified, the proposed system making stop for the withdrawal operation and make sure the reliability and validity of user.

This information must entered by the user to be registered in the gateway and the bank chosen by the user. The Card number is resulting from taking the hash function MD5 and CRC32 function (Cyclic Redundancy Check) for some of information which is (ZIP for the country (This system applied three example of countries, Iraq, Paris, America), Full Name for the user, and code number for the bank). The implemented system produced the account number. The first customer is given the account number equal to one; the second customer is given the account number equal to two, etc. Figure 4 illustrates the user's registration.

1) C. E-bank

The user can choose any bank willing to deal with it. There are four operations for in the bank that is (account balance, withdrawal, deposit and money transfer) that the user want to make them and our DB is sensitive to the changes.

2) Account Balance Operation

The user needs to determine the his/her account balance, therefore he must enter the card number .when the server is matching between the user's card number and the card number in our designated database , the account balance is produced, and these operations as listed below:

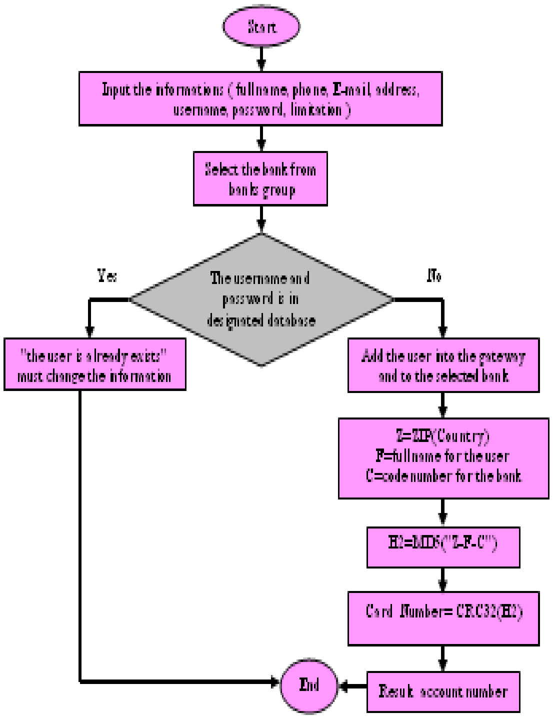

Start
Input the informations (fullname, phone, E-mail, address, username, password, limitation)
Select the bank from banks group
The username and password is in designated database
Yes
No
"the user is already exists" must change the information
Add the user into the gateway and to the selected bank
Z=ZIP(Country)
F=fullname for the user
C=code number for the bank
H2=MD5("Z-F-C")
Card Number= CRC32(H2)
Result account number
End

Figure 4: Flowchart of user's registration

3) Money Transfer

If the user wants to transfer money from any bank to another, then he must select the operation which is money transfer. The user can transfer the money from his/her account balance to another consumer by enters the several inputs (card number for the sender, account number for the Recipient, the amount of money to be transferred and determine the bank that receives the money).

The user can pay the bills for water, electricity, Telephone, etc., through the use of this proposed system and benefit from the money transfer service. Through a financial transfers between the user's bank and the banks that deal with Telephone Companies, water, Electricity, etc.

In the process of transferring funds from one bank to another, the mechanism is needed to convert the currency; where the process of conversion from one currency to another is through the program to determine Currency Exchange. Figure 5 illustrates the Money transfer.

4) Deposit operation

The user can select this operation, when he wants deposit the money in his/her account balance. He enters the card number and the amount of money which is wanted to add to his/her account balance.

5) withdrawal operation

The user must enter the card number and the amount of money who wants to withdraw from his/her account balance when he selects the withdrawal operation.

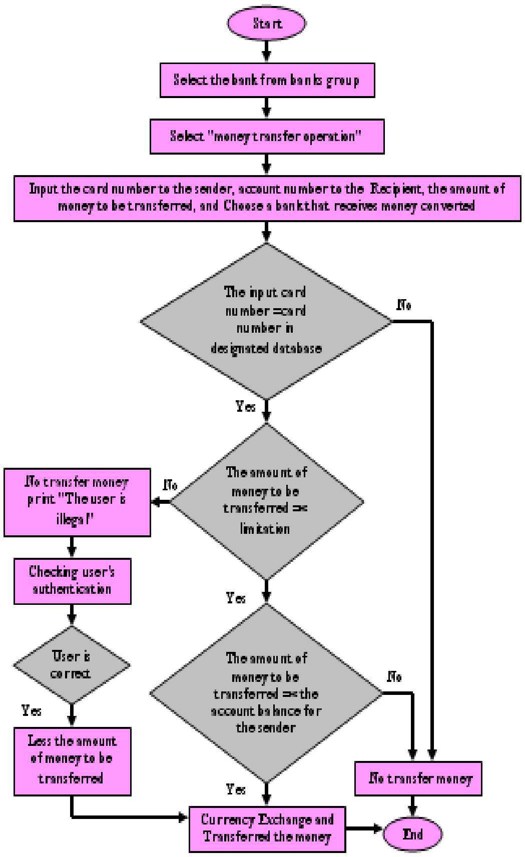

Start
Select the bank from banks group
Select "money transfer operation"
Input the card number to the sender, account number to the Recipient, the amount of money to be transferred, and Choose a bank that receives money converted
The input card number =card number in designated database
No
Yes
The amount of money to be transferred =< limitation
No
No transfer money print "The user is illegal"
Checking user's authentication
User is correct
Yes
Less the amount of money to be transferred
Yes
The amount of money to be transferred =< the account balance for the sender
No
No transfer money
Yes
Currency Exchange and Transferred the money
End

Figure 5: Overview of "money transfer" process

VII. Results And Analysis

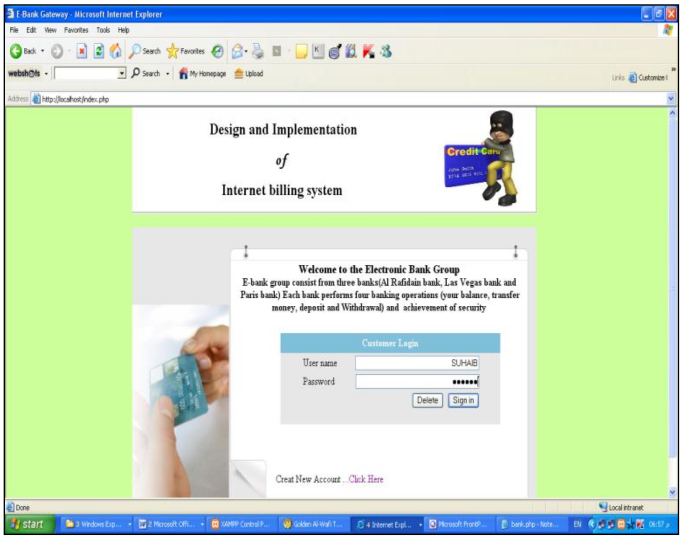
The Internet billing system is implemented to satisfy the security requirements. The authentication process is done using hash function, CRC 32 function. The payment system work 24 hour a day, 7 days a week and any time anywhere.

The Implemented system leads to increase flexibility and efficiency of the payment process by reducing transaction process time and reducing cost.

The implemented system introduced many flexible interfaces such as main system interface (figure 6), user registration interface (figure 7) and E-bank services interface (figure 8).

VIII. Conclusion

In this paper, we implemented Internet billing system; by construction of virtual banks which perform the processes of banks. Some of the concepts of security have been applied in this system to protect the system from unauthorized access. The security issue is implemented via; encrypted passwords using hash function (MD5), the hash function (MD5) and CRC32. These functions are used to generate the card number; the amount of money transferred cannot exceed a certain imitation. Users can do their payment via E-bank any time anywhere, in which access time is reduced as possible.


E-Bank Gateway - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
websh@ts Search My Homepage Upload
Links Customize
Address http://localhost/index.php
Credit Card
Done
Local intranet
start

Design and Implementation

of

Internet billing system

Welcome to the Electronic Bank Group

E-bank group consist from three banks(Al Rafidain bank, Las Vegas bank and Paris bank) Each bank performs four banking operations (your balance, transfer money, deposit and Withdrawal) and achievement of security

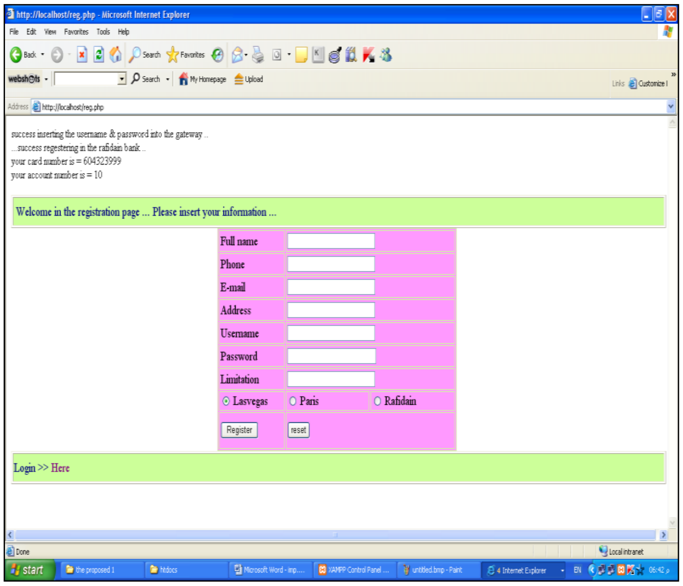
Customer Login

User name	SUHAIB
Password	••••••

Delete Sign in

Creat New Account ...Click Here

Figure 6: Main system interface


http://localhost/reg.php - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
websh@ts Search My Homepage Upload
Links Customize
Address http://localhost/reg.php
Done
Local intranet
start

success inserting the username & password into the gateway ..
...success regestering in the rafidain bank ..
your card number is = 604323999
your account number is = 10

Welcome in the registration page ... Please insert your information ...

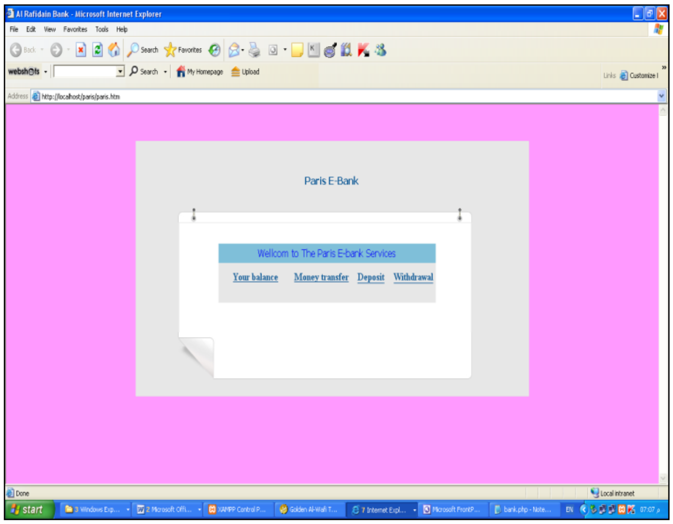
Full name	
Phone	
E-mail	
Address	
Username	
Password	
Limitation	

● Lasvegas ○ Paris ○ Rafidain

Register reset

Login >> Here

Figure 7: User registration interface


Al Rafidain Bank - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
websh@ts Search My Homepage Upload
Links Customize
Address http://localhost/paris/paris.htm
Done
Local intranet
start

Paris E-Bank

Wellcom to The Paris E-bank Services

Your balance Money transfer Deposit Withdrawal

Figure 8: E-bank services interface

REFERENCES

[1] Assimakopoulos Nikitas A., Anastasis N. Riggas & Giorgos K. Kotsimpos, "A Systemic Approach for an Open Internet Billing System",2003, http://www.afscet.asso.fr/resSystemica/Crete02/Assimakopoulos,%20Riggas,%20Kotsibos.pdf

[2] Ghosh Anup K., "E-commerce Security": Weak Links, Best Defenses, Wiley Computer Publishing,1998.

[3] Singh Abhishek, OM Shankar, Vikash Kumar and Tapanray," Risk in E-Banking", CC BY-NC 3.0,2009, available at

http://www.scribd.com/doc/22356535/Risk-in-E-Banking-PDF,visited on July 16,2011.

[4] NN Murthy, BM Mehtre, KPR Rao, GSR Ramam, PKB Harigopal, and KS Babu, "Technologies For E-Commerce: An Overview", CMC Center-R&D, CMC Limited Old Mumbai Highway, Gachibowli Hyderabad – 500 019, Andhra Pradesh ,2000.

[5] Sumanjeet Singh," Emergence of Payment Systems In The Age Of Electronic Commerce:The State Of Art ", Global Journal of International Business Research Vol, 2, No, 2, 2009.

[6] Chan Henry, Raymond Lee, Tharam Dillon and Elizabeth Chang ,"E-commerce Fundamental and Applications", Baffins Lane, Chichester, West Sussex, PO19 lUD, England,2001.

[7] Watson Richard T. , Pierre Berthon,Leyland F. Pitt and George M. Zinkhan, "ElectronicCommerce :The Strategic Perspective", Creative Commons Attribution 3.0 License, 2007.

[8] Media Abdul Razak Ali,M.Sc ,In a computer and software engineering department of the University of AL-mustansiriya, "Design and Implementation of SET Enabled E-commerce System",2005.

[9] Crookes J ,"Multiservice Billing System - a platForm for the future", BT Technol JVol 14 No 3 July 1996.

[10] Harte Lawrence, "Internet TV Billing Systems", Althos pupishing,2011, http://www.althos.com/tutorial/Internet-TV-station-tutorial-Billing-Systems.html.

[11] Mostafa hatem, "Billing System : Introduction", codeproject, 2005, http://www.codeproject.com/KB/architecture/billing.aspx#Introduction

[12] Thomas Rob, "Choosing an E-Billing System", published in ILTA - December, 2005.

http://www,serengetilaw,com/news/serengetimeasuretwice,pdf.

[13] Bo Yang,Liu Dongsu and Wang Yumin, "An Anonymity-Revoking E-payment System with a Smart Card", springer-verlag, Volume 3, Number 4, 4 December 2001, http://www.springerlink.com/content/2uhetkje7a1pkljk/.

[14] EWB Team," Electronic Extra Work Billing System: Online Step -By-Step Instructions", Revision 2, ISSC, EWB Release 1.1 Instructions, January 12, 2001, http://www,dot,ca,gov/hq/esc/tollbridge/BenMar/006034/MaterialsHandout/EWB,pdf.

[15] Barreto P.S. , G. Amvame-Nze, C.V. Silva, J. S. S. Oliveira, H.P. de Carvalho, H. Abdalla Jr, A.M. Soares, and R. Puttini,"A Study of Billing Schemes in an Experimental Next Generation Network", Springer-Verlag Berlin Heidelberg 2005, http://www.springerlink.com/content/r5nh3n0ebgf7w2h3/.

[16] Shiqun Li · GuilinWang · Jianying Zhou · Kefei Chen,"Fair and Secure Mobile Billing Systems", Springer Science+Business Media, LLC, 2008.

[17] Levi Albert and Çetin Kaya Koç "CONSEPP: Convenient and Secure Electronic Payment Protocol Based on X9.59", IEEE Computer Society Press, Los Alamitos, California, March 21, 2009, http://discuss.itacumens.com/index.php?topic=57564.0.

[18] Antoniou Giannakis, Lynn Batten, Shivaramakrishnan Narayan, and Udaya Parampalli, "A Privacy Preserving E-payment Scheme", Springer-Verlag Berlin Heidelberg 2009, http://www.springerlink.com/content/h2253738530vm061/ .

A Performance Evaluation of Multiple Input Queued (MIQ) Switch with Iterative Weighted Slip Algorithm

S N Kore, Sayali Kore, Ajinkya Biradar
Electronics dept.
Walchand college of Engineering,
Sangli, India

Dr. P J Kulkarni
Computer Science dept.
Walchand college of Engineering,
Sangli, India

Abstract— **Many researchers had evaluated the throughput and delay performance of virtual output queued (VOQ) packet switches using iterative weighted/un-weighted scheduling algorithms. Prof. Nick Mckeown from Stanford University had evolved with excellent iterative maximal matching (i-slip) scheme which provides throughput near to 100%. Prof. Kim had suggested multiple input queued architecture which also provide more than 90 % throughput for less number of input queues per port. (In VOQ N queues per port are used). Our attempt is to use MIQ architecture and evaluate delay, throughput performance with i-slip algorithm for scheduling. While evaluating performance we had used Bernoulli's and Bursty (ON-OFF) traffic models.**

***Keywords-** Network communications; Packet-switching networks;routing protocols; Sequencing and scheduling.*

I. INTRODUCTION

A High speed switches mainly classified as input queued (IQ) switch, output-queued (OQ) switch, and combined input-and output-queued (CIOQ) switch. An OQ switch buffers cells at the output ports. OQ switches guarantee 100% throughput since the outputs never idle as long as there are packets to send. An NxN OQ switch must operate N times faster than the line rate. Memory technology cannot meet that kind of high-speed requirement [1]. Therefore, IQ and CIOQ switches have gained widespread attention. The input queue switch has limitation of throughput equal to 58.6% [1] [2]. The most common architecture is the CIOQ switch in which buffering occurs both at the input and at the output. But CIOQ always need speedup high speed-up factor of two to provide 100% throughput. Both IQ and CIOQ switches use virtual output queuing in which each input maintains a separate queue for cells destined for each output [2][3].

Matching algorithms for Virtual output queuing removes head-of-line (HOL) blocking and overcomes limit on the throughput single FIFO queue [1]. In virtual output queued switches scheduling or selection of packets at HOL is critical issue. Many algorithms have been proposed for scheduling an IQ switch to obtain high throughput. All the algorithms find a matching between the inputs and outputs, but they were derived with different weighing techniques. Under the matching paradigm, the scheduler matches an input with an output and finds the maximal number of those pairs in a given time slot. This usually takes a few iterations in one time slot. Numerous algorithms work in iterative way and most of them are variants of i-slip algorithms [4] [7] [8]. The i-slip algorithm innovated by Prof. Nick-Mckeown had played vital role in development of switching architecture [4] [5] [6] [7] [8]. In multiple input queued (MIQ) architecture there are M queues per input port. Total NM queues are used in MIQ whereas N^2 queues are used in VOQswitches. Even with M=8 and N=64 throughput achieved is greater than 92%. In VOQ we need to handle 4096 queues and in case of MIQ only 512 queues need to be handled. It's quite interesting to analyze the performance of MIQ with i-slip. The i-slip algorithm have not been evaluated for multiple input queued switch (MIQ) where number of queues per input port is less than N if size of switch is N x N. We are reporting the performance of i-slip in MIQ under Bernoulli's arrival and bursty arrival.

II. SWITCH AND TRAFFIC MODEL

A. Switch Model

This section describes the switch model. Here number queues per port (M) used are less that size of switch (N x N) where $M \leq N$. In VOQ N^2 queues needs to be taken care where as in MIQ only NM Queues needs to be taken care. Our aim is to obtain throughput to be 100%, that restrict condition that every cell slot time we need to select non-conflicting N input-output matches among NM matches (N^2 in case of VOQ). Suppose M=2, indicate that there are two queues per port.

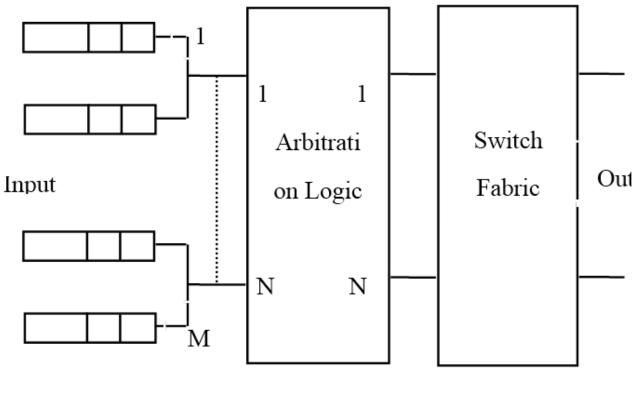


Figure 1. MIQ Switch

Arrivals, destined for output ports with even number are saved in one queue at input ports and others are saved at another queue. In general arrival to an output port N is saved in k^{th} queue at input port where k= N mod M where k=1,2...M.This approach introduces a new problem as there are now (a maximum of) N^2 packets at HOL in case of VOQ and NM packets in MIQ for selection.

The problem of selecting, N packets among NM packets to transmit becomes much more complex scheduling problem. The performance of such architecture is determined by the arbitration algorithm. This is illustrated in section 3.

B. Traffic Model

*Bernoulli's arrival:*In this arrival process the cell arrived in each time slot is identical and independent of other time slot. Assume that probability that cell arrives is p. Each arrived cell chooses output equally likely. Hence traffic is said to be uniformly distributed over output port Please do not revise any of the current designations.

Bursty arrival: Basically this type of modeling of traffic source is called as ON-OFF type. Here in ON-period (active) sourcesends packets & in OFF-period (silent) no packets are sent.

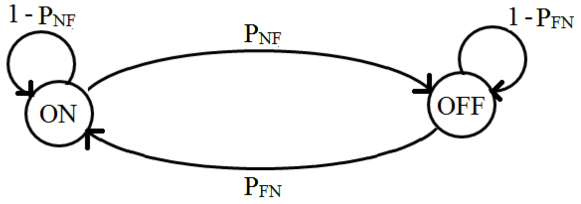


Figure 2. ON-OFF Traffic Model

Time is slotted and packets are generated in slot hence it is called as Markov Modulated Bernoulli's process (MMBP) with two states.

It is further classified as MMDP i.e. as Markov Modulated Deterministic process.

State transition matrix $T = \begin{bmatrix} 1 - P_{NF} & P_{NF} \\ P_{FN} & 1 - P_{FN} \end{bmatrix}$

Prob[ON state] = $\frac{P_{FN}}{P_{FN}+P_{NF}}$,

Prob[OFF state] = $\frac{P_{NF}}{P_{FN}+P_{NF}}$

P_n = Prob that ON state has length 'n' slot i.e being ON state it will remain for another (n-1) times in ON state and then goes to OFF state.

$P_n = (1-P_{NF})^{n-1}.P_{NF,}$ Its geometric distribution with Mean burst length $L_{b.}$

$L_b=\sum_{i=1}^{\infty} i\ P_{NF}\ (1 - P_{NF})^{i-1} L_b = \frac{1}{P_{NF}}$

Offered Load = $\rho = \frac{P_{FN}}{P_{NF}+P_{FN}}$

Burst length chosen is 16 and offered load ρ is 0.8 then $P_{NF} = 0.0625$ and $P_{FN} = 0.25$ which are used to change the state of the system. If system is in ON state it always generate packet uniformly distributed to any output port till system changes the state.

III. PREPARE YOUR PAPER BEFORE STYLING

A. Round-Robin Matching (RRM) algorithm

Before RRM is very similar to Prof. Anderson's Parallel Iterative Matching (PIM) [3] [4], where packet selection is done at random, it uses modulo N round robin arbiters, one for each input and one for each output. Each arbiter maintains a pointer, indicating the element that currently has highest priority. RRM operates as follows:

1. *Request*: Each unmatched input sends a request to each output if it has at least one packet at HOL.

2. *Grant*: Each output that has received at least one request selects one request to grant by means of its round-robin arbiter. It chooses the input that appears next in the round robin, starting from the input currently being pointed to. The pointer which is advanced (modulo N) to onebeyond the input just granted.

3. *Accept*: Similarly, each input that has received at least one grant will select one grant to accept by means of its round-robin arbiter. It chooses the output that appears just next in the round robin, starting from the output currently being pointed to. The pointer is advanced (modulo N) to one beyond the output just accepted. Unfortunately, RRM does not perform very well even under uniform i.i.d. Bernoulli arrivals; saturation throughput is merely 63%, which is close to that of PIM. The reason for reduction in throughput is because output arbiters tend to synchronize, causing multiple arbiters to grant to the same input, which leads to a waste of grants and thus poor throughput.

B. i-SLIP in VOQ with Bernoulli's arrival

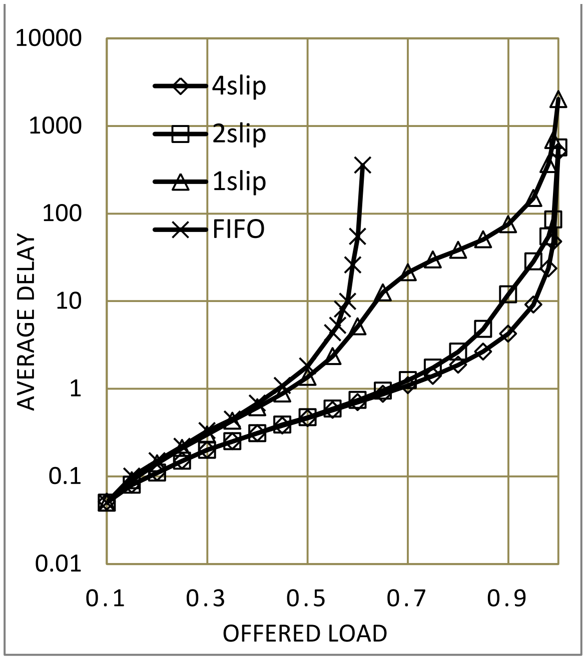
i-slip is an improvement on RRM, aimed at preventing synchronization of arbiters. Its operation is very similar to RRM, with only a modification in step 2 of how the pointers are updated:

1. Request: Same as RRM.

2. Grant: Each output that has received at least one request will select one request to grant by means of its round-robin arbiter. It chooses the input that appears next in the round robin, starting from the input currently being pointed. The pointer is advanced to one beyond the input just granted if and only if the grant is accepted in step 3.

3. Accept: Same as RRM. Note that this is almost identical to non-iterative SLIP, with the exception of the added condition in steps 2 and 3: the pointers are only updated in the first iteration for reasons of fairness. Compared to SLIP, iterative SLIP improves performance further when the numbers of iterations are increased. On an average i-SLIP appears to converge in about Olog(N) iterations, a result similar to PIM..

Fig.3 indicates that saturation throughput under VOQ (i.e for 16x16 switch with 16 queues per port) can be achieved to be 100% under 1, 2 or 4 slip. Increasing Number of iterations improves the delay performance.



Figer 3. Delay performance of switch size 16x16 with VOQ with Bernoulli's arrival and i-slip of 1,2,4.

C. i-SLIP in VOQ with Bursty arrival

Fig.3 shows the performance, evaluated for switch size of 16x16 with 16queues per port and bursty traffic (ON-OFF) with different burst size with multiple number of iterations. Burst size selected is 16 and 64.

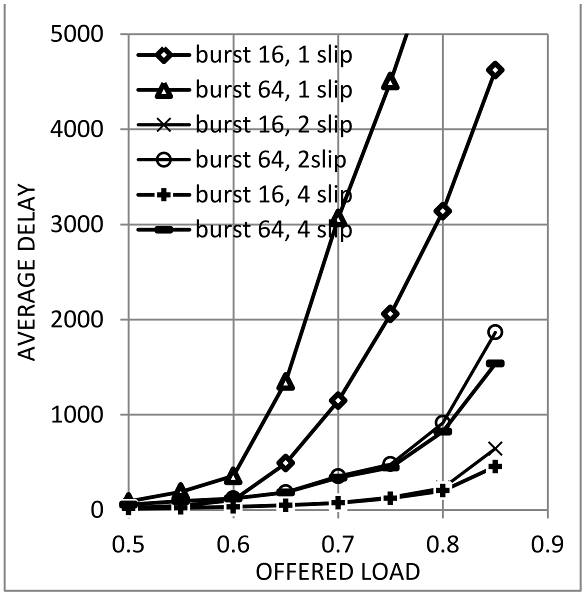


Figure 4. Delay performance of switch size 16x16 with VOQ with Bursty arrival and i-slip of 1,2,4.

Delay performance is degraded as burst size is increased. Such model is analytically analyzed by Prof. Kelinrock and Prof. Kim with restricted rule. Here Iterative slip is un-weighted i.e. matching of input output does not consider any bias such as length of queues or longest port first etc. With slip more than 2 does not improve the performance under lighter input load (less than 0.7) but it observed that under higher input load(more than 0.85) through and delay performance is improved.

Fig. 4 indicate the delay performance of i-slip for switch size 16x16 with number of queues per port are 4,8 with number of iterations are 1,2,4 slip. In 8 queues per port with 2 slip and 4 Queues per port with 4 slip has same performance. It's obvious that increase in number of queues per port and increasing number iteration in i-slip will give performance nearer to output queuing.

IV. I-SLIP IN MIQ

Our attempt is to evaluate the performance of i-slip algorithm if number of queues per port is reduced to M where M< N the model is identified as MIQ. Here16x16 with number queues per port reduced to 8 with 1-slip, then this model is equivalent and approximated to even and odd queues where throughput is saturated to 76.4 % [9].Iterative slip in VOQ suggested by Prof. McKeown is implemented in Cisco router1200 and giving best performance [5]. Our attempt is that the McKeown'si-slip implementation can be extended to MIQ where management of N^2 queues reduces to NM queues only. Even through the saturation throughput is limited in MIQ can be overcome by implementing iterative i-slip. Number of iteration of Olog(N) are sufficient for achieving throughput of 100%. In fig 2 it clearly shows that in 16x16 switch with 8 number of queues per port and slip 1,2 4 has increased saturation throughput from 76%, 91% , 98% respectively. As the number of iteration is increased delay performance is also improved.

A. Weighted i-SLIP in MIQ with Bernoulli's arrival:

Here in Fig.5 simulation graphs are drawn for number of queues per port 4 and 8 along with variation of slip. Arrival Traffic is Bernoulli's arrival with uniform distribution. Each input port will send requests to output port depending on HOL packet destination address along with queue length in that queue. Each input port can send maximum M requests to the output arbiter. In case of VOQ, there might be maximum N requests from each input port. Total Number of requests sends to output arbiters can be NM, which is reduced in MIQ (In VOQ it is N2).

Arbiter at the output port will receives number of requests. Arbiter at the output will grant one request among the received from various input ports which have highest queues length. Grants received at input port i from different output ports j are evaluated. If multiple grants are received then one is chosen which has highest queues length. Once the input arbiter accepts the jth port request then queue number, j mod M is evaluated to select queue from corresponding input port to remove cell from its HOL.

Let system be queues/port be M=8 and number of ports be N =16. In VOQ there is M=N, hence each input port can maximum send 16 requests to 16 arbiters at output ports if there is cell at HOL.In case of MIQ there maximum 8 requests will be sends from each input port to different 16 output arbiter.At input port 1 there are 8 queues and queue no.1 at input port 1 can store cells destine to output port 1 or 9.

TABLE I. DELAY PERFORMANCE UNDER BERNOULLI'S ARRIVAL FOR SWITCH 16X16 WITH 8 QUEUES PER PORT AND SLIP OF 1, 2, 4

16x16 switch with 8 queues per port					
4 slip		***2 slip***		***1 slip***	
Load	***Delay***	***Load***	***Delay***	***Load***	***Delay***
0.1	0.05	0.1	0.05	0.1	0.05
0.15	0.08	0.15	0.08	0.15	0.09
0.2	0.12	0.2	0.11	0.2	0.14
0.25	0.16	0.25	0.16	0.25	0.21
0.3	0.2	0.3	0.2	0.3	0.3
0.35	0.26	0.35	0.26	0.35	0.43
0.4	0.32	0.4	0.32	0.4	0.61
0.45	0.4	0.45	0.4	0.45	0.9
0.5	0.49	0.5	0.5	0.5	1.4
0.55	0.61	0.55	0.63	0.55	2.48
0.6	0.76	0.6	0.8	0.6	5.81
0.65	0.95	0.65	1.02	0.65	18.87
0.7	1.23	0.7	1.81	0.66	25
0.75	1.63	0.75	2.84	0.67	33.11
0.8	2.28	0.8	5.65	0.68	44.63
0.85	3.49	0.85	7.03	0.69	64.63
0.9	6.64	0.9	29.05	0.7	114.19
0.95	27.8	0.91	48.52	0.75	2329
0.96	67	0.92	100		
0.97	276	0.93	301		
0.98	652				
0.99	1115				
1	1537				

Hence input arbiter at port number 1 can send request for HOL at M=1 to outputs port 1 or output port 9 depending on current address in HOL cell.While sending queue length it is number of cells waiting in its queue which contains cells destine to output port 1 and 9.

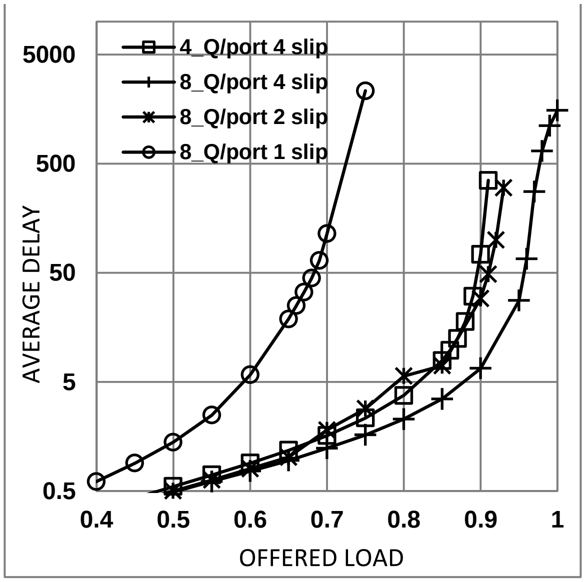


Figure 5. Delay Performance under Bernoulli's arrival for Switch 16x16 with 8 queues per port and slip of 1,2,4

The input output ports for which matching is obtained will not take any part in further iterations. It is observed that 4 iteration are sufficient to find maximal match and throughput to be 100%.

Fig.5 shows the graph of delay performance for 16x16 switch with 8 queues per port. Here total input queues are 128 instead of 256. In case of 8 queues per port with 1-slip limits maximum maximum throughput approximated to 76%. As the number of slips are increased then throughput increases to 84% & 98% with slip of 2& 4. Delay is also bounded under heavy traffic load conditions.

Fig. 5 indicate the delay performance of i-slip for switch size 16x16 with number of queues per port are 8 with number of iterations are 1,2,4 slip. In 8 queues per port with 2 slip and 4 Queues perport with 4 slip has same performance. It's obvious that increase in number of queues per port and increasing number iteration in i-slip will give performance nearer to output queuing.i-slip in MIQ with bursty arrival

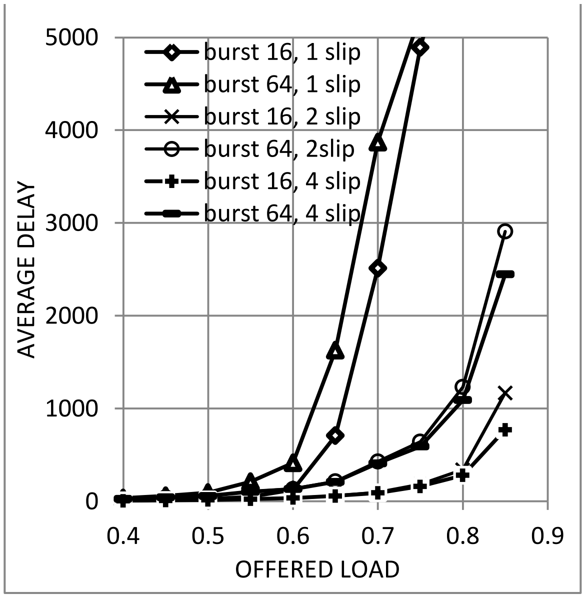


Figure 6. Fig. 6 Delay Performance under Bursty arrival for Switch 16x16 with 8 queues per port and slip of 1,2,4

In Fig.6 performance evaluation of MIQ switch with switch size of 16 and number of queues per port = 8 are taken with different burst size and slip is varied as 1, 2, 4. It is observed that as slip is increased throughput delay performance approaches output queuing. It's always recommended if traffic is bursty then increase slip for better performance.

V. CONCLUSION

Here performance of i-slip under MIQ structure with uniform Bernoulli's arrival and bursty (on-off) arrivals. Increasing number of iterations is more flexible than increasing number of queues port and is the key for obtaining good delay throughput performance. Increase in the burst size degrades the performance of switch even under virtual output queuing. Maximum Weight matching algorithm can be the better solution to provide good delay throughput performance. Such algorithms are computationally complex and have to be implemented on parallel architectures for real time application.

There are different variants of i-slip are available and our work can be extended to these algorithms to obtain better performance.

REFERENCES

[1] M.G.Hluchyj and M.J.Karol, "Queuing in high performance packet switching" IEEE J.Selected Area of Communication, Vol6, pp.1587-1597, December 1988

[2] S. N. Kore, V. B. Dharmadhikari, H. S. Jamadagni, "Cell Selection Policies in Multiple Input Queued ATM Switch", Centre for Electronics Design and Technology, IISC Bangalore-12. Published in The Canadian Conference on Broadband Research (International conference) CCBR -99

[3] McKeown, N., "The iSLIP Scheduling Algorithm for Input-Queued Switches," IEEE/ACM Trans. Networking, vol. 7, no. 2, Apr. 1999, pp. 188-201.

[4] C. Kolias, l.kleinrock "On odd even ATM Switch", IEICE Trans. On commun. vol.e81b, no. 2, Feb. 1998, pp. 244-250

[5] Saad Mneimneh, "Matching From the First Iteration: An Iterative Switching Algorithm for an Input Queued Switch" IEEE/ACM Trans. Networking, vol. 16, No. 1, Feb. 2008, pp. 206-217.

[6] Alessandra Scicchitano, Andera Binaco, Palol Giaccone, Emilio Leonardi, Enrico Schiattarella, "Disrtibuted scheduling in input queued switches" IEEE Comm. Society ICC, 2007

[7] Mohsen Bayati, Balaji Prabhakar, Devavrat Shah, Mayank Sharma, "Iterative Scheduling Algorithms" IEEE Comm. Society IEEE INFOCOM, 2007

[8] Kevin F. CHEN, Edwin H.-M. SHA, S.Q. ZHENG "Fast and Noniterative Scheduling in Input-Queued Switches" Int. J. Communications, Network and System Sciences, 2009, 3, pp. 169-247

AUTHORS PROFILE

Mr. S N kore: received his B.E degree in Electronics and Telecommunication from University of Pune at Pune and M.E. degree in electronics from Shivaji University at Kolhapur in 1983 and 1991respectively.He has worked as scientific officer at Tata Institute of Fundamental Research (TIFR) at Pune from 2nd Nov'83 to 1st Oct'84. Currently he is an Associate Professor of electronics engineering at Walchand College of engineering, Sangli, Maharashtra.

Miss Sayali S.Kore: she had done B.tech in Electronics Engineering from Walchand College of Engineering Sangli in 2012. She is recipient of gold medalist in 2012 at WCE Sangli. Her area of interest is Signal processing, Image processing, VLSI Engineering and Computer Networking.

Mr. Ajinkya Biradar: received his B.E. degree in Electronics and Telecommunication from University of Pune at Pune in 2010 and is pursuing M.Tech in electronics from Walchand College of engineering, Sangli

Dr.P.J. Kulkarni: He is currently working as Deputy Director of Walchand College of Engineering (WCE) Sangli. He is Prof. in Department of Comp. Sci. and Engineering at WCE Sangli. He is recipient of state level Ideal Teacher award winner in 2011-12. He has published more than 24 papers in international Journal and national journals. Also completed 9 research projects. He has guided more than 10 Phd in Shivaji University.

An integrated modular approach for Visual Analytic Systems in Electronic Health Records

Muhammad Sheraz Arshad Malik[1], Dr.Suziah Sulaiman[2]
Department of Computer & Information Science
Universiti Teknologi PETRONAS
Bandar Seri Iskandar, 31750, Tronoh, Perak, Malaysia.

***Abstract*— Latest visual analytic tools help physicians to visualize temporal data in regards to medical health records. Existing systems lack vast support in the generalized collaboration, a single user-centered and task based design for Electronic Health Records (EHR). Already existing frameworks are unable to mentor the interface gaps due to problems like complexity of data sets, increased temporal information density and no support to live databases. These are significant reasons for a single model to comply the end user requirements. We propose an integrated model termed as CARE 1.0 as a future Visual analytic process model for resolving these kinds of issues based on mix method studies. This will base on different disciplines of HCI, Statistics as well as Computer Sciences. This proposed model encompasses the cognitive behavioral requirements of its stake holder's i.e. physicians, database administrators and visualization designers. It helps in presenting a more generalized and detailed visualization for desired medical data sets.**

Keywords- Visual analytic Systems; EHR; Information visualization; CARE 1.0.

I. Introduction

Visual Analytic System **(VA)** is a combination of automated analysis techniques with information visualizations for an effective knowledge derivations, relationship of data and decisions on very large and complex data sets[5][6][7]. Visual analytic systems help in relating the information to a simple, easy and understandable form for corragurated, adjunct, multi-dimensional and complicated data. Most salient features within a VA system are earlier detection of expected data anomalies, easier understanding of results and accelerate efficiency of decision support for various data sets belonging in medical to engineering domains.

VA systems play a vital role in presenting the Electronic Health Records **(EHR)** for improving health care operations in various countries around the globe. These tools enables the users, designers and collaborators to better understand the temporal and non-temporal queries , differential analysis and validated decisions As data repositories increased both in sizes and intangibilities, existing visual analytic applications reflect poor implementation of Information visualization in terms of exploratory analysis, effective user cognitive representation and user driven process modeling[8][9].

This research work focuses on temporal categorical EHR data including past hospital visits of patients, diagnosis and drugs care plan etc.

Same strategy can be used for temporal numerical data as that of blood pressure readings of patient, pulse rates, heart beats etc. Current EHR systems mostly deal with data entry, retrieval, availability and query based numerical representations. But these systems are lacking the abstract temporal analysis, lesser information density as well as poor exploratory processes. Existing visualization tools Midgaard [14], Lifelines [12][13] , Web-based interactive Visualization Systems[11] and VIE-VISU [9] help physicians to represent EHR data using interactive visualization for different particular scenarios as like.

As in medical health records, visualization application like lifeline2 [12] already tried to solve the various temporal data problems by providing colored triangular dots on a screen. But due to variance in temporal data i.e. both categorical and numerical, it is hard to generalize the requirements of physicians at user interface level. Major reasons are misalignment data formats, mismatch between stake holders requirements and limited scope of datasets. These effects on data integration for analytical representation of information as well as co relation with active live data base systems. As there is a clear difference arises between the user requirements understanding between physicians, data base administrators and visual analysts based on different temporal queries. This leads in generating a potential gap to attain a fully functional analytic system required for easy understandability of information from EHR database.

In this work, a detail studies have been carried out on different available existing VA frameworks to identify user interface problems that hinders understanding of visualization. Misalignment in representation of data fields, results in temporal data exploration issues within existing EHR systems. This provokes in reduced capability in formulation of representation of analyzed data based on the user cognitive experiences, queries input and results exploration. So resulting information details lack with reference to end user perspective as in case of doctors who required particular information for a particular portion of existing record. Existing Visual analytic systems do not provide any kind of live or dynamic support for online databases. This is most importantly needed for latest kind of applications based on reducing the time factor to minimize the information processing for analytical representation in front of physicians.

Our model proposes a simplified and dynamic integrated VA solution from different disciplines e.g medical data base sets, HCI, Information visualization and Statistical analysis and databases etc for improving health care interactive facilities. This will help physicians to improve the diagnostic approaches for best patient care facilities as well as other users for oncoming needs and trends based on different temporal data repositories. This will not only leverage the service standards in medical fields but also in other disciplines of social sciences for future research and exploration. In addition, it will work to help designers to align a visual analytic system based on collaboration of adjunct user's requirements who can be from different disciplines as well as formulate a base for researchers in information visualization.

This model integrates the stake holders inputs, their information exploring trends for both temporal and non-temporal queries that will help to process the information both offline and online. The resulting framework will help to determine the requirements for a multi varied application tool. This model will facilitate designers for the knowledge management of physicians and medical specialists to align data representation on the basis of their desired EHR data sets. In a normal VA system, physicians demand a visualization based on their requirements using queries, backend queries are prepared in SQL or any database language by DBA and design of visualization is controlled by IV designer. So integration of these stake holders is considered very much important within a single model that is our real research motivation.

A systematic literature review methodology is used to present related work section in this paper regarding user interface problem in existing VA systems. Proposed Model section is going to explain the details of the various portions of this framework. Conclusions and future work sections represent perspective ongoing development in it.

II. Related Work

"A picture is worth a thousand of words"[3] really depicts the values of pictures in any information communication processes and facilitate the understanding of information using pictorial shapes.

Different VA systems frame works studies have been carried out in past few years in Information visualization to better visualize the patient data records. LifeLine is one of the first implemented frameworks that presents patients data like problems, allergies, diagnosis, labs, imaging, medications, and immunizations in the form of lines. Thickness and color of lines represent different conditions of severity as well as termination of any symptoms [13].Temporal and casual relationships within these facets are not focused within this framework. This was further addressed in using object based scenario within visual representations using Prefuse Toolkit **[4].** Dendrogram, Knowledge tables/ Hierarchies, Scatterplot/histograms and Parallel coordinates are used in HCE 3.0 tool to understand the variability in data using clustering algorithmic techniques. This focused on realizing the importance of multi dimensional data visualizations based on clustered gene technology [1] to segregate it on basis of similar gene groups**.**

Visual data analysis is used to manage meta-knowledge to handle vast amounts of extracted knowledge. Existing graphical approaches representing association rules in data sets are not offering global and detailed views at same time and that is crucial in HCI [15] proposed prototype CBVAR using FEV (Fish Eye View) focuses on using reduction of association rule sets including generic base sets plus generic association rules in XML file format. This prototype is not providing enough support for user interface contexts for processing user information. However , there is another approach of using Visual analytic systems was proposed on time oriented data concept using computational analysis methods on diversified data[20]. This proposition, still requiring the implementation in task oriented domains as in medical data not all temporal queries can be mapped as time oriented data and it may differ in representing characteristics of time. Similar but different in approach concept in the form of data aesthetics that could be intrinsic or extrinsic were used on VA systems with termed as data focus and mapping formats for representation [21]. This work represented the role of data aesthetics in information visualization representations in VA systems but it lacks the knowledge encompassing aesthetic factors validation.

A sketching-oriented design is proposed for information visualization tools like InfoViz [22].According to Craft, cartoon like representation is a better and more visualized idea within the visualization schemas to present the complex 2D and 3D designs. A varied set of sketches should be taken regards to be taken and compared as with reference to visualize the temporal categorical data e.g. structure of kidneys elongations or in case of skull 2D views in cases of MRI. Thus a varied set of change in shapes can be presented in the form of already available set of sketches and then it can better be configured with existing databases by designers to match the most suitable data set presentation based on that using any information visualization kit mostly used 2D visualization kits.

InfoVizModel, is another emerging framework is another approach to address the solution of representation of web based data using IV models using Information Architecture (IA)**[**19**]**.This work closely related to our present work as it is strongly works in division of small level and big level architecture and currently used as a part for information retrieval from CNKI, Baidu and Google. In this system, information architecture is divided into four phases' navigation, organization, labeling and retrieval where each phase works independently as well as dependently. Former architecture focuses on creation of and management of personal information content and laters is responsible for the construction of a webplatform and web content.

Another approach to visualize the web based information was presented in the form of separate APIs in the form of separate stages where one API translates and keeps records of other API[18]. In this way only those icons and data facets are represented that are required by the user as active and inactive icons category. This approach is mostly and widely adopted by Google, Yahoo and MSN web portals in these days for representation of different data sets.

Two stage Visual Analytic framework was proposed by [23] based on combination of different disciplines to solve heterogeneous data visualization issues. This model tried to sort out major problems in user interface issues, data transformations by using HCI, Computer Sciences, and Mathematics etc based two stages. This work is also an extension of nested model design [16][14][17] and contextual design as combination in separate stages. Designers can derive tangible artifacts based on two requirements **a)** these should be concrete basis for practical VA system **b)** Artifacts must be usable by users without introducing cognitive overhead. Primary stage have domain observation, analytical requirements that are further divided into analysis base on each layer i.e user, context,task and organization. Secondary stage refers to user centric refinement that is referring to logging of information about how data generated. This further leads to user pattern analysis and customization that is substituting usage collection, annotation tracking and content sharing and interaction logging.

Last but not the least most closest and up to date work that actually urges to work visualization analytics was the LifeLine2 model as one of the latest approach to address EHR data representation issues. LifeLine2 is providing a model based solution is proposed as one step next version to previous LifeLine due to its increased performance, temporal event based data visualizations and a bit approach to user alignment issues. It does not focus on temporal numerical data so only temporal categorical data e.g past hospital visits, diagnosis etc not the blood pressure readings or such other data sets [1][9][12][8]. One of the key features are granularities with the data sets by keeping them as events to better represent using triangular dots and a graphical user interface based on time stamps.

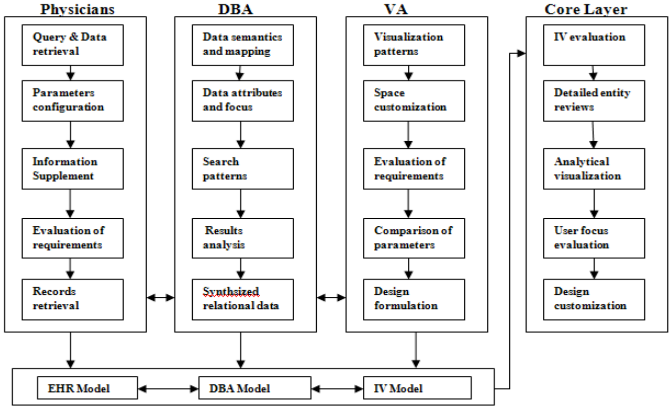
All the above solutions presented till yet are lacking still lots of areas of research that leads to poor exploratory processes, incomplete information representation and differential visualization. This results in frustration and un necessary delay in representations different perspectives in cases of a complete temporal health record data base that is very much important to improve our existing health care systems not in Malaysia but in the whole world.

Our conceptual integrated model will focus on solution side of these issues, as in all previous models by the help of integration of VA, Database and physicians query modular structures together used within the latest models to better grasp the visualization for its stake holders.

III. PROPOSED INTEGRATED MODEL CARE 1.0

The newly proposed model is comprised of three major portions that are pertaining to its three major stake holder's i.e. Data base administrators & holders for data sets, Physicians query structures and finally the VA designers using IV modules.

This model is integration of three base models along with one core layer that is providing a base line for visualization on the basis of most important stake holder i.e physicians. As with the previous related works of different researchers in this domain that focuses on each individual portion on single model structure [12][16][21]. This work goes into an extension of these previous work to accumulate the positive aspects of covering predicted problems by physicians with temporal data i.e both categorical and numerical data. Temporal categorical data sets are tested already been used at limited way to explore the use of IV [12] but still it lacks the ability to cover support to backtrack history of an event of interest in results exploration.



Integrated Modular structure of Framework

Three components are described separately as mentioned in the above model.

A. EHR Model:

Each EHR data set model associates with the ***type of query*** for the data it requires e.g. in case of finding the number of patients entered in an Intensive Care Unit (ICU) suffering heart attacks would have a previous history of visiting hospitals , medicine dose recommendations, symptoms and any possible smoking habits or not. So if a physician realizes to utilize the previous record then it's easier to understand the patient history and potential reasons for the attack causes. Parameters configuration, information supplement, evaluation of requirements and records retrieval within the available datasets classification are associated components within this model. As correct parameters settings are required based on personal experience, training and background of a physician and specialist to help in determining the clearest data set in a complex dataset. Information supplement is any additional info associated within a desire query and evaluation of requirements is focusing on the validity of commended operation. Each of these factors effect on the usability of other portions of the model and dependent on each other and if there is any issue occurs that can be tracked on that portion of the model easily.

B. DBA Model:

Data semantics and mapping is mostly carried out on the basis of the temporal queries in already existing VA applications but still these mappings are not considered with reference to search patterns and data focus[9][14][21] i.e either intrinsic or extrinsic. This integrated modular framework presenting the same mapping with relation to search patterns like index searching within the desired datasets selection based on physicians query. As the medical terms, is a different area of understanding for Database teams and people so there is a mutual framework of understanding can be developed in conjunction with continuous feedback and data linkage with previous phase. So database administrator gets query request clearly understandable to the context what kind of exact data modules are required based on all these previous inter related activities and this will yield a more closer relational data form. This data form will then be testified both at physicians requirements analysis side and after matching to the right demand set it will be processed to formulate a visualization.

C. IV Model:

Each visualization model comprises of its various components that are developed on the basis of various directly and indirectly related factors for its resulting visualization could be a tree, histogram, lines, bars or any match of colors etc. IV model within this framework is comprised of its influencing factors that are characterized as patterns, space, evaluation of requirements, comparison of parameters and design evaluation that are interlinked with each other. Various VA models tried to follow either a few components or fewer integration [4][8][16] but as temporal data sets generate complex variable visualizations that integrations at a wider range is required. Current VA model used within this proposed work is using all these features as integration in a patterned way so that even physicians requirements should not be un addressed and in the same way it also measures the validity of a visualization existence for a given data set based on its relationship within its own and associated parameters. Most significant feature within this area is the collaboration of this phase with DBA phase as visualization will directly be impacted by the data it requires to formulate that is already associated with resulting query.

D. Core layer:

Core layer is the most integrated and vibrant portion of this model. This is comprising of components that work on validation and testing side of the resultant data, its visualization , analysis linkage and user focus disciplines by the interconnection of all three modular layers of the integrated model. This portion of the model facilitates its all stake holders on user specific areas of exploration, design customization based on different data sets and anomlies. These features help to co relate the inputs and outputs of each portion of the stake holders and provides an inter linkage between them thus trying to remove errors, measure the level of flexibilities within a given data sets and all possible visualizations with maximized level of analysis and co relations of entities .

Normally this portion will help not only the physicians but also the health policy makers to identify the trends analysis of various health issues e.g outbreak of any disease, drug effects on a particular age group in a geographical area or multi root cause analysis of symptoms, diagnosis and suggested measures.

IV. Conclusion And Future Work

Information visualization in EHRs is one of the significant areas of interest but due to lack in collaboration of its different stake holders that are from different backgrounds it leaves a wider gap for its complete implication. This conceptual model tries to focus on removing the gaps as created by the requirements of physicians that need a kind of visualization that can help them to fulfill their diagnostic requirements flexibly. While DBAs and Visualization designers are two other different disciplines groups that leave the potential gap of data evaluation based on former's requirements.

This model tried to bring the three stake holders work together by adjuncting their outputs and validating their inputs. Still there are further studies required at micro level conjunctionalities of mapping of user inputs, design re-structuring flexibilities side effects and encompassing the variabilities in EHR datasets with reference to exploration spectrum.

Acknowledgement

Authors will like the opportunity to say thanks to Computer and Information Sciences Department of Universiti Teknologi PETRONAS for providing technical support. Moreover, special thanks to Dr. Suziah Bt.Sulaiman for her guidance and tiring efforts at every instance of this piece of research proposal. We are thankful to The Universiti Teknologi PETRONAS for providing logistic and financial support for this study.

References

[1] J. Seo and B. Shneiderman, "A knowledge integration framework for information visualization," in From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments, H. Matthias, et al., Eds., ed: Springer-Verlag, 2005, pp. 207-220.

[2] P. Hastreiter and T. Ertl, "Integrated registration and visualization of medical image data," in Computer Graphics International, 1998. Proceedings, 1998, pp. 78-85.

[3] D. Pfitzner, et al., "A unified taxonomic framework for information visualization," presented at the Proceedings of the Asia-Pacific symposium on Information visualisation - Volume 24, Adelaide, Australia, 2003.

[4] J. Heer, et al., "prefuse: a toolkit for interactive information visualization," presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, Portland, Oregon, USA, 2005.

[5] D. B. N. Christopher, A. Harle and R. Padman, "AN information visualization approach to classification and assessment of diabetes risk in primary care," in Proceedings of the 3rd INFORMS Workshop on Data Mining and Health Informatics, J. Li, D. Aleman, R. Sikora, eds., 2008.

[6] G. A. Daniel Keim, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melancon. , "Visual Analytics: Definition, Process, and Challenges in Information Visualization." vol. 4950. Heidelberg: Springer-Verlag, Berlin, 2008.

[7] J. C. Granda, et al., "Design Issues in Remote Visualization of Information in Interactive Multimedia E Learning Systems," in Visualisation, 2008 International Conference, 2008, pp. 70-76.

[8] W. Xiaoyu, et al., "A two-stage framework for designing visualanalytics system in organizational environments," in Visual Analytics Science and Technology (VAST), 2011 IEEE conference on, 2011, pp. 251-260.

[9] W. Horn, Popow, C., and Unterasinger, L., "Support for fast comprehension of ICU data: Visualization using metaphor graphics.," Method Info. Med, vol. 40(5):, pp. 421–424, 2001.

[10] D. L. McGuinness, et al., "Towards Semantically Enabled Next Generation Community Health Information Portals: The PopSciGrid Pilot," in System Science (HICSS), 2012 45th Hawaii International Conference on, 2012, pp. 2752-2760.

[11] D. S. Pieczkiewicz, Finkelstine, S. M., and Hertz, M. I., "Design and evaluation of a web-based interactive visualization system for lung transplant home monitoring data.,". Proc. Am. Med. Inform. Assoc. Ann. Symp.,, pp. 598–602, 2007.

[12] T. D. Wang, et al., "Visual information seeking in multiple electronic health records: design recommendations and a process model," presented at the Proceedings of the 1st ACM International Health Informatics Symposium, Arlington, Virginia, USA, 2010.

[13] C. Plaisant, Mushlin, R., Snyder, A., Li, J., Heller, D., and Shneiderman, B.,, "Using visualization to enhance navigation and analysis of patient records.," Proc. Am. Med. Inform Assoc., pp. 76–80, 1998.

[14] R. Bade, Schelchweg, S., and Miksch, S, " Connecting timeoriented data and information to a coherent interactive visualization.," ACM Int Conf on Human Factors in Comp Syst.,, vol. Proc 22nd,New York, NY, USA. , pp. 105–112, 2004.

[15] O. Couturier, et al., "A scalable association rule visualization towards displaying large amounts of knowledge," in Information Visualization,2007. IV '07. 11th International Conference,2007,pp. 657-663.

[16] T.Munzner, "A Nested Model for Visualization Design and Validation," Visualization and Computer Graphics, IEEE Transactions on, vol. 15, pp. 921-928, 2009.

[17] H. C. Purchase, et al., "Theoretical Foundations of Information Visualization," in Information Visualization, K. Andreas,et al.,Eds.,ed:Springer-Verlag, 2008, pp. 46-64.

[18] K. Matsui, et al., "A Proposal of Framework for Information Visualization in Developing of Web Application," in Applications and the Internet (SAINT), 2011 IEEE/IPSJ 11th International Symposium on, 2011, pp. 457-462.

[19] W. Jiaxin, "WIVF: Web information visualization framework based on information architecture 2.0," in Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on, 2010, pp. 734-738.

[20] W. Aigner, et al., "Towards a conceptual framework for visual analytics of time and time-oriented data," presented at the Proceedings of the 39th conference on Winter simulation: 40 years! The best is yet to come, Washington D.C., 2007.

[21] A. Lau and A. V. Moere, "Towards a Model of Information Aesthetics in Information Visualization," in Information Visualization, 2007. IV '07. 11th International Conference, 2007, pp. 87-92.

[22] B. Craft and P. Cairns, "Directions for Methodological Research in information Visualization," in Information Visualisation, 2008. IV '08. 12th International Conference, 2008, pp. 44-50.

[23] W. Xiaoyu, et al., "A two-stage framework for designing visual analytics system in organizational environments," in Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on, 2011, pp.251-260.

Developing Backward Chaining Algorithm of Inference Engine in Ternary Grid Expert System

Yuliadi Erdani
Politeknik Manufaktur Bandung
Bandung, Indonesia

Abstract— **The inference engine is one of main components of expert system that influences the performance of expert system. The task of inference engine is to give answers and reasons to users by inference the knowledge of expert system. Since the idea of ternary grid issued in 2004, there is only several developed method, technique or engine working on ternary grid knowledge model. The in 2010 developed inference engine is less efficient because it works based on iterative process. The in 2011 developed inference engine works statically and quite expensive to compute. In order to improve the previous inference methods, a new inference engine has been developed. It works based on backward chaining process in ternary grid expert system.**

This paper describes the development of inference engine of expert system that can work in ternary grid knowledge model. The strategy to inference knowledge uses backward chaining with recursive process. The design result is implemented in the form of software. The result of experiment shows that the inference process works properly, dynamically and more efficient to compute in comparison to the previous developed methods.

Keywords- expert systems; ternary grid; inference engine; backward chaining.

I. Introduction

There is no official definition for the term of expert system but there are some descriptions for it created by people working in the field of expert system. With the term of expert system we refer to a computer system or a program, into which several procedures of artificial intelligence are integrated. Expert system can be understood as vehicles for Artificial Intelligence (AI) techniques.

Expert system is also applied artificial intelligence. Expert systems are programs for storing and processing knowledge of a special area, that why they are able to answer to questions and solve problems, with which experts normally deal [6]. In the current situation, expert system is an intelligent computer program that uses knowledge and inference procedures to solve problems that are difficult enough to require significant human expertise for their solution.

The expert knowledge must be obtained from specialist or other sources of expertise, such as texts, journal, articles, and database [8]. This type of knowledge usually requires much training and experience in some specialized field such as medicine, geology, system configuration, or engineering design. Once a sufficient body of expert knowledge has been acquired, it must be encoded in some form, loaded into a knowledge base, then tested, and refined continually throughout the life of the system.

Some task that can be performed by expert system are difficult tasks to be specified, the task that may have incomplete or uncertain data, there may not always be an optimum solution, the task cannot be solved in a step-by-step manner, and solutions are often obtained by using accumulated experience [11]. An example of applied expert system is web-based consultation system [1]. Benefit of expert systems is the ability to preserve valuable knowledge which would otherwise be lost when an expert system is no longer available. Expert system also can allow an expert to concentrate on more difficult aspect of the task. It can enforce consistency, and they can perform dangerous tasks which would otherwise be carried out by humans.

One of known and very popular expert system type is production rule. Production rule are simple but powerful forms of knowledge representation providing the flexibility of combining declarative and procedural representation for using them in a unified form. The term production rule came from production system which is developed by A production system is a model of cognitive processing, consisting of a collection of rules (called production rules, or just productions). Each rule has two parts: a condition part and an action (conclusion) part. The meaning of the rule is that when the condition holds true, then the action is taken. A typical production rule is given below:

***IF** there is a flame **THEN** there is fire*

The statement of the rule above means that fire is caused by a flame. If anything happens with a flame, it will lead to fire production. It is the idea of production system. The production system or production rule provides appropriate structures for performing and describing search process. A production system has four basic components as enumerated below [9]: A set of rules following the classical IF-THEN construct. If the conditions on the left-hand side are satisfied, the rule is fired, resulting in the performance of action on the right-hand side of the rule, A database of current facts established during the process of inference, A control strategy which specifies the order in which the rule are selected for matching of antecedents by comparing the facts in the database. It also specifies how to resolve conflicts in selection of rule or selection of facts, and a rule firing module. An expert system that impalements production rule is known as rule-based expert system.

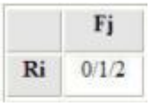
Building or construction of rules can easily be done in most rule-based expert system. Knowledge expert or knowledge engineer does not have to do any work specifying rules and how they are linked to each other. Sometime the knowledge expert or knowledge engineer can reference rules or facts that have not yet been created. It seems to be a simple and an instant work. The problem due to the performance of the knowledge will not occur until the number of rules is getting higher. Some problem may appear in the form of inconsistent rules, unreachable rules, redundant rule and closed rule chain of rules.

The mentioned problems above have been solved with so called Ternary Grid [1][4][5]. Since the ternary grid can solve some problem concerning knowledge bottleneck, there is no any developed inference method, technique or machine working on ternary grid knowledge model. As consequence of it, all ternary grid knowledge must be converted into production system format, so that the knowledge can be processed by rule-based inference machine to deliver solution. The inference engine of an expert system interprets and evaluates the facts in the knowledge base in order to provide an answer. By the numerous methods of problem solution, which can be implemented in a rule interpreter, only the representatives of the concatenation strategies are to be treated here: forward chaining and backward chaining

The developed inference machine of expert system can work in ternary grid knowledge model. The strategy to find solution previously uses forward chaining with iterative approach [12]. As another alternative solution, the inference machine can be implemented by using backward chaining method. The emphasis of this paper is to describe the backward chaining method that is implemented in the inference engine of ternary grid expert system. The backward chaining method should bring more benefit than developed forward chaining method, e.g. dynamic answers of inference engine, efficient computation effort, etc.

II. METHODS

Before talking the inference engine, we must first regard the knowledge representation. The Ternary Grid represents the production rule in the following structure (Fig. 1):



	Fj
Ri	0/1/2

Figure 1. Ternary Grid basic structure

Ri: Rule i (i is the number of rule)

Fj: Fact j or logical term (j is the number of fact)

$$i = \{1,2,3,...,I\}$$

$$j = \{1,2,3,...,J\}$$

$$J > I + 1$$

The Value of every grid box is 0, 1 or 2

0 = unused, is represented by empty grid box.

1 = Fact Fm belongs to the condition part of rule Rn (LHS= Left Hand Side).

2 = Fact Fm is part of the conclusion part of Rn (RHS = Right-Hand Side).

In the beginning of the development, the Ternary Grid was only used for knowledge acquisition system. The basic feature of the system architecture is to organize the independent and sequential obtaining process of the factual knowledge and the elicitation process of judgmental knowledge using Ternary Grid. The overall systems architecture is presented in terms of collection of functions providing effective acquisition, processing, transferring and flexible transformation of knowledge. This section gives an overview of the system that shows the design approach of the system and the concept of acquisition process.

The Ternary Grid acquisition system has task to organize the knowledge base, to obtain the factual knowledge, to elicit the judgmental knowledge, and to transfer the knowledge into knowledge base. The system was able to improve the performance of expert system knowledge [5]. Meanwhile the Ternary Grid has been used for other part related the expert system, such as knowledge representation, knowledge based system, inference engine, etc.

Even the Ternary Grid has been applied in an inference engine, but the approach of existing inference engine uses only forward chaining method. The backward chaining method has not been used in any inference engine. The developed backward chaining method will be implemented in inference engine of expert system based on Ternary Grid. Inference engine of expert system is computer program that answers questions from user. It processes all information from the knowledge base by firing rules and facts [9].

Backward chaining is a strategy of inference process which is the opposite of forward chaining. The strategy of backward chaining is started from a goal and ended with a fact that leads to the goal. Backward chaining method is also called as goal driven strategy of inference engine. In other literature, the backward chaining is a chaining process that begins with the last element in the chain and proceeds to the first element. This is often a very effective way of developing complex sequences of behavior

There are two search algorithms, which are normally used by backward chaining method, i.e. depth-first and breath-first search algorithm. Both algorithms search data in a tree structure. Depth-first search algorithm searches a data in a tree as deep as possible before backtracking. Breath-first search algorithm searches a data in neighbor nodes before it moves deeper to the bottom of the tree. Using depth-first search algorithm, the process of backward chaining can be illustrated as follows:

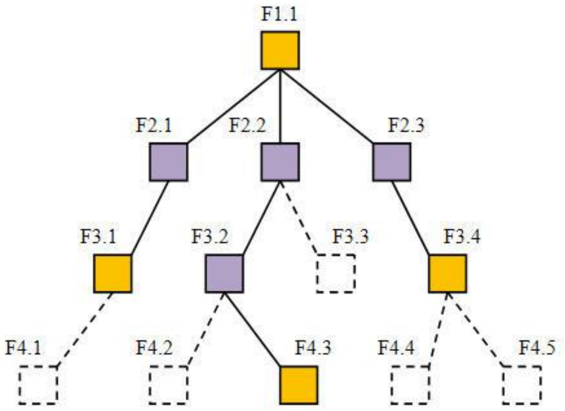


Figure 2. Backward chaining illustration

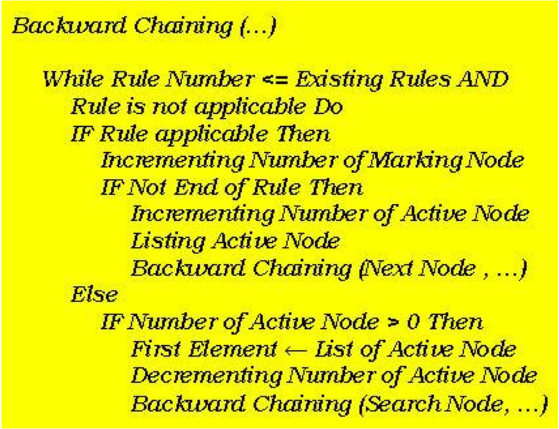
Figure 3 shows the illustration of backward chaining process of inference engine of expert system. The known facts are F1.1, F2.1, F3.2 and F3.4 in the beginning of process. The inference process begins from fact F1.1. That fact F1.1 is called as goal. The inference process moves then backward to other facts behind goal.

It is so called condition part of rule. The inference engine tries to applied fact F2.1, F2.2 and F2.3, etc. The developed backward chaining uses depth-first search algorithm. The following steps describe the inference process of expert system using developed backward chaining algorithm as follows:

Goal: F1.1
Fetch: F2.1 → unknown
Fetch: F3.1 → known
Rule (F3.1, F2.1) → fired
Rule (F2.1, F1.1) → fired
Fetch: F4.1 → unknown
Fetch: F2.2 → unknown
Fetch: F3.2 → unknown
Fetch: F4.2 → unknown
Fetch: F4.3 → known
Rule (F4.3, F3.2) → fired
Rule (F3.2, F2.2) → fired
Rule (F2.2, F1.1) → fired
…
Etc.

The inference process continues until all possible facts have been asked (tested) to be fired. The developed algorithm searches all data deeper into the bottom of tree structure in the knowledge base of expert system.

The designed and implemented backward chaining algorithm is explained in the following algorithm:

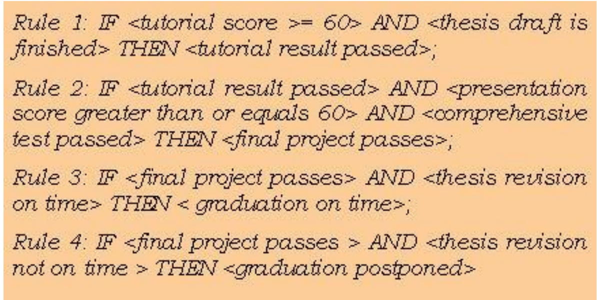


```
Backward Chaining (…)

    While Rule Number <= Existing Rules AND
        Rule is not applicable Do
        IF Rule applicable Then
            Incrementing Number of Marking Node
            IF Not End of Rule Then
                Incrementing Number of Active Node
                Listing Active Node
                Backward Chaining (Next Node , …)
        Else
            IF Number of Active Node > 0 Then
                First Element ← List of Active Node
                Decrementing Number of Active Node
                Backward Chaining (Search Node, …)
```

The process continues as far as the number of rules is less than the number of existing rules and there is rule that is not applicable. If a rule is applicable (fired) then the program search the next facts that lead to applicable rule until there is no fact found anymore. If a rule is not applicable (fired) then the program search other possible rule in other paths. The process continues until all possible facts have been tested or fetched.

III. Results

The same data as [12] [13] is used in this experiment.



Rule 1: IF <tutorial score >= 60> AND <thesis draft is finished> THEN <tutorial result passed>;

Rule 2: IF <tutorial result passed> AND <presentation score greater than or equals 60> AND <comprehensive test passed> THEN <final project passes>;

Rule 3: IF <final project passes> AND <thesis revision on time> THEN < graduation on time>;

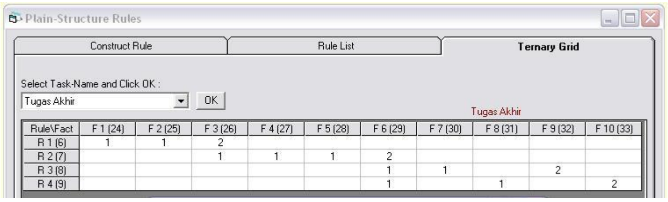
Rule 4: IF <final project passes > AND <thesis revision not on time > THEN <graduation postponed>

According to Ternary Grid acquisition technique [5], the mentioned rules are inputted into ternary grid knowledge base as it is shown in figure 4. Using the developed concept, the rule-based format must not be converted into ternary grid. The inference process of the expert system in ternary grid uses backward chaining with recursive approach. All fact inputs are stored in set of facts Fk. The inference engine searches all rules that are possible to be executed and stores them in set of rules Rx:

$$R_x = \left\{ p \,\middle|\, p \to q,\ p \in F_k, p \in F, F_k \subset F \right\} \qquad (5)$$

The inference engine determines then rules that are able to be applied and stores in the following set of rule Ryn.

$$R_{y_n} \subset R_x \qquad (6)$$

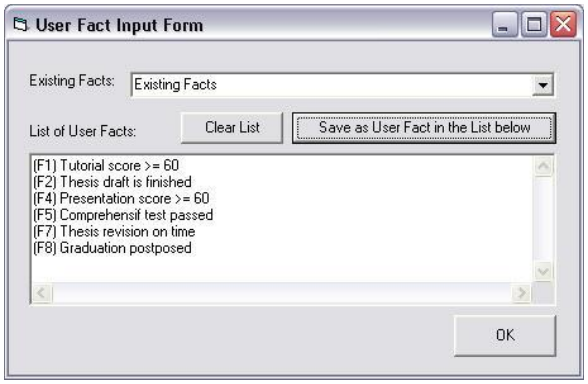


Select Task-Name and Click OK : Tugas Akhir

Tugas Akhir

Rule\Fact	F 1 (24)	F 2 (25)	F 3 (26)	F 4 (27)	F 5 (28)	F 6 (29)	F 7 (30)	F 8 (31)	F 9 (32)	F 10 (33)
R 1 (6)	1	1	2							
R 2 (7)			1	1	1	2				
R 3 (8)						1	1		2	
R 4 (9)						1		1		2

Figure 3. Given facts and rules in ternary grid



Existing Facts: Existing Facts

List of User Facts:

(F1) Tutorial score >= 60
(F2) Thesis draft is finished
(F4) Presentation score >= 60
(F5) Comprehensif test passed
(F7) Thesis revision on time
(F8) Graduation postposed

Figure 4. Kknown facts

The application does then the inference task by processing all facts that are given before. The result of inference process can be shown in figure 6.

Inconsistent rules can be detected and eliminated using the following processes:

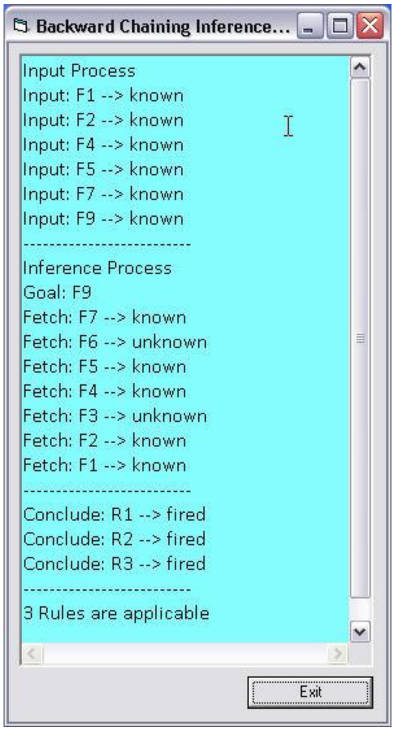
- Find rows, in which value 3 appears:

$$B = \left\{ i \mid a_{ij} = 3 \right\} \qquad (7)$$

- Remove row duplication

$$C = \bigcup_{b \in B} b \qquad (8)$$

The result of inference process shows the effectiveness of the developed algorithm. In comparison to the method of [7] [12] and [13], the developed inference method can work directly in ternary grid without having to be converted to rule-based format. In comparison to [12] and [13], the developed method work more dynamic and efficient to compute. The implemented recursive approach in inference process reduced the number of required iteration. The result of inference process can also show other facts that weren't known before. These all facts could lead to give more conclusions that will bring more information to expert system.



```
Input Process
Input: F1 --> known
Input: F2 --> known
Input: F4 --> known
Input: F5 --> known
Input: F7 --> known
Input: F9 --> known
------------------------
Inference Process
Goal: F9
Fetch: F7 --> known
Fetch: F6 --> unknown
Fetch: F5 --> known
Fetch: F4 --> known
Fetch: F3 --> unknown
Fetch: F2 --> known
Fetch: F1 --> known
------------------------
Conclude: R1 --> fired
Conclude: R2 --> fired
Conclude: R3 --> fired
------------------------
3 Rules are applicable
```

Figure 5. Result of inference process using backward chaining algorithm

The following data are taken from several conducted experiments

TABLE I. EXPERIMENT DATA

Number of rule	Number of fact	Number of Iteration
10	35	62
20	70	134
30	100	295
50	180	673

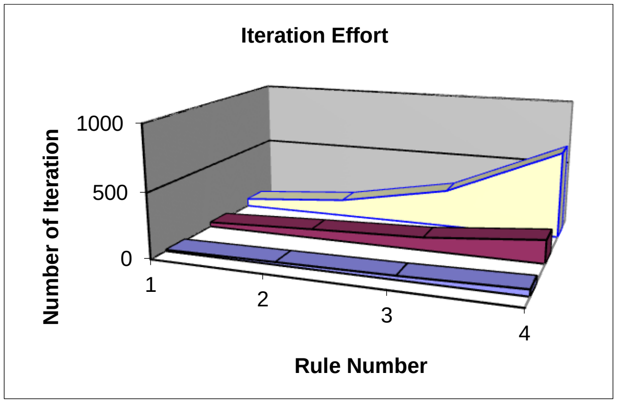


Figure 6. Recursion effort

Figure 6 show the effort of recursion that is influenced by increasing the number of facts and rules.

IV. CONCLUSION

The developed inference engine using backward chaining method in ternary grid works properly. It can determine all applied rules that lead to the goal fact. In comparison to the previous work using iterative approach and forward chaining algorithm, the developed method works more dynamic and more efficient to compute. The inference process could also detect and lead to other rules that previously unknown and brought new solutions. Referring to some literatures concerned expert systems; the developed method is novel and will give contribution in developing inference method of expert systems.

REFERENCES

[1] David J.C. McKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press. 2003.

[2] Erdani, Y., Hunger, A., Werner, Stefan., Mertens, S. Web-Based Consultation System with Expert System, IASTED CST 2003 – International conference (International Association of Science and Technology for Development – Computer Science and Technology), May 19-21, 2003, Cancun, Mexico, 2003, ISBN – 0-88986-349-0, page 61-64

[3] Erdani, Y., Hunger, A., Werner, S., Mertens, S. Ternary Grid as a Potentially New Technique for Knowledge Elicitation/Acquisition, 2nd IEEE Conference on Intelligent System, vol I: pp. 312-315. ISBN 0-7803-8278-1, Varna - Bulgaria, June 22-24, 2004. This paper has been accepted also by SEKE 04 (Software Engineering and Knowledge Engineering) conference in Banff, Canada.

[4] Erdani, Y., Hunger, A., Werner, S., Improving the Knowledge Performance using Ternary Grid Knowledge Acquisition and Model, WSEAS Transactions (Journals) on Information Science and Application, Issue 2, Volume 2, February 2005. ISSN 1790-0832

[5] Erdani, Yuliadi, Acquisition of Human Expert Knowledge for Rule-based Knowledge-based Systems using Ternary Grid, Verlag – Dissertation.de, Berlin, 2005, ISBN 3-89825-000-8

[6] Horn, Christian; Kerner, Immo O. : Lehr- und Übungsbuch Informatik, Band 3. Fachbuchverlag Leipzig, im Carl Hanser Verlag, München Wien, 1997. ISBN 3-446-18699-9

[7] Hunger, A., Werner, S., CONGA: A Course Online/ Offline Information and Guidance System to support an International Degree Course, Proceeding of ICCE 99 (Chiba-Japan, 1999, ISBN 1 58603 027 2, Page 577-583)

[8] Joseph C. Giarratano, Gary Riley. Expert Systems, Principles and Programming, 2005, ISBN 0-534-38447-1

[9] Krishnamoorthy, C. S., Rajeev, S., Artificial Intelligence and Expert Systems for Engineer, CRC Press, Boca Raton – Florid, 1996, ISBN 0-8493-9125-3

[10] Lunze, Prof. Dr.-Ing. Jan. Künstliche Intelligenz für Ingenieure, Band 1: Methodische Grundlagen und Softwaretechnologie, R. Oldenbourg Verlag, München Wien, 1994, ISBN 3-486-22287-2

[11] Sascha Mertens, Marius Rosu, Yuliadi Erdani. An intelligent dialogue for online rule based expert systems, Proc. of the 2004 International Conference on Intelligent User Interfaces, January 13-16, 2004, Funchal, Madeira, Portugal. ACM 2004, ISBN 1-58113-815-6

[12] Yuliadi Erdani, Pengembangan Metode Inferensi untuk Sistem Pakar berbasis Ternary Grid, Jurnal P&PT Vol. VIII, No. 2, Desember 2010, ISSN 0854-5766

[13] Yuliadi Erdani, Developing Recursive Forward Chaining Method in Ternary Grid Expert Systems, IJCSNS (International Journal of Computer Science and Network Security), Vol. 11 No. 8, August 2011, ISSN 1738-7906

AUTHORS PROFILE

Dr. Ing., Yuliadi Erdani, M.Sc., Dipl. EL. Ing., HTL received the B.S. degree in Electrical Engineering from the Ingenieurschule Burgdorf (ISB) Switzerland in 1995 and the M.S. degree in Computer Science and Communication Engineering (CSCE) from the University of Duisburg, Germany in 2002. He received PhD degree (Dr.-Ing.,) in Informationstechnik from the same university i.e. University of Duisburg-Essen, Germany in 2005. He did a lot of work with expert systems. He has been continuing his research work in the area of expert systems from 2006 until now and every year he always got research fund/grant. He works now as lecture in a state Polytechnic (university of applied science) in Bandung, Indonesia.

Comparison Study of Commit Protocols for Mobile Environment

Bharati Harsoor[1]
[1]Dept of CSE, University College of Engg,
OU, Hyderabad, India

Dr.S.Ramachandram[2]
[2]Dept of CSE, University College of Engg,
OU, Hyderabad, India

Abstract—**This paper presents a study of protocols to commit the transactions distributed over several mobile and fixed units and provides the method to handle mobility at the application layer. It describes the solutions to defeat the dilemma related to principle implementation of the Two Phase Commit (2PC) protocol which is essential to ensure the consistent commitment of distributed transactions. The paper surveys different approaches proposed for mobile transaction and outline how the conventional commitment are revisited in order to fit the needs of mobile environment. This approach deals with the frequency disconnections and the movement of mobile devices. This paper also proposes Single Phase Reliable Timeout Based Commit (SPRTBC) protocol that preserves the 2PC principle and it lessens the impact of unreliable wireless communications.**

Keywords- Mobile Transactions; Transaction Log; Transaction Recovery; Network disconnection; handoff; ACID properties.

I. Introduction

Due to the mobile computing standard, the mobile users can access information independent of their physical location through wireless connections. However, accessing and manipulating the information without confining the users to definite locations complicates the processing of data. Mobility and disconnected computing are two major issues in such environment. With the advancement in the distributed technology, consistency mechanism for a mobile transaction has become easier and manageable with more than one participant. To preserve data consistency *all* or *nothing* effect of transaction execution is usually enforced at commit time.

To ensure consistent termination of distributed transactions regardless of communication and site failure we use the following various commit protocols for mobile transactions. Along with study, we propose a new execution framework providing an efficient extension that supports the reliable execution of mobile transactions called Single Phase Reliable Timeout Based Commit (SPRTBC) protocol.

It is one phase commit protocol, which makes use of only *decision phase* to perform commitment of transactions. With the proposed model, during first step, no resources are blocked due to timeout approach. It confirms ACI (Atomicity, Concurrency, Isolation) properties, during second step, it preserves durability property; hence it supports disconnections and handoff having reduced blocking situations.

II. Comparison Study Of Commit Protocols For Mobile Environments

A. Two Phase Commit Protocol (2PC)

In distributed systems, an Atomic Commitment Protocol is required to terminate the distributed transactions. The most commonly used and standardized mechanism dealing with the commitment problem is Two Phase Commit protocol (2PC) [4][7]. It is the simplest and most used Atomic Commit Protocol. Generally, it follows two phases, voting and decision phase. In voting phase the coordinator requests all the participants to prepare to commit the transaction, if any of the participant responds No, the coordinator decides to abort and inform every participant to abort their local transaction, otherwise if all the participants votes Yes then CO decides to commit and informs all the participants to make their local transaction durable or permanent. The participants acknowledge the coordinator.

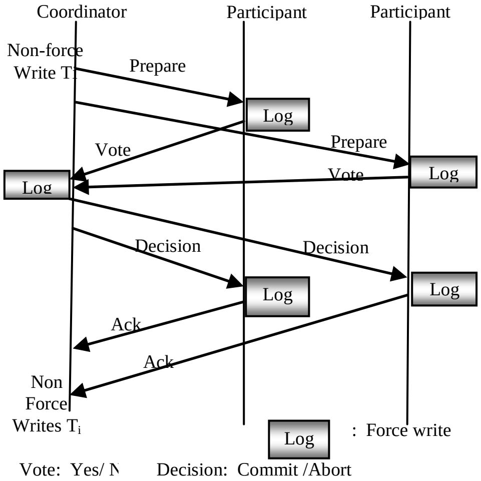


Figure 1. 2PC-Two-Phase Commit Protocol

Fig. 1 illustrates sequence of operations carried out in Two Phase Commit Protocol. The issues related to the Two Phase Commit Protocol with the mobile environment are, Inaccurate global decision ; means if the coordinator does not receive all votes before its timeout expiration it may decide to terminate globally in case where global commit is possible and Blocking situations; mean the following blocking conditions may arise in Two Phase Commit Protocol.

- The coordinator waits until the reception of every acknowledgement messages from the participants. Here no data is blocked.
- A participant waits after voting commit until reception of global decision. This situation may be constrained where the participant's local resources remain locked during that time. Such a participant is not permitted to unilaterally terminate the local transaction.

B. Mobile - 2PC (M - 2PC)

The aim of M-2PC (Mobile-2PC) [7] protocol is to globally commit mobile transaction T_m which is being executed over more than one host. Suppose that a transaction T_m is issued at MH called as Home-MH and which is attached to a BS called Home-BS. As MH transfers from one cell to next cell and it joins to a new BS that is called Current-BS. At commit time a commit demand is issued from the Home-MH, hence its current-BS (either it will be the Home-BS) becomes the commit-BS. The M-2PC protocol may terminate either with same cell or in a new cell protected by new BS. Fig. 2 illustrates the sequence of transaction executions used by M-2PC. The transaction execution is split into two phases; the initial one is almost equal to traditional 2PC, while the next phase controls the mobile wireless part.

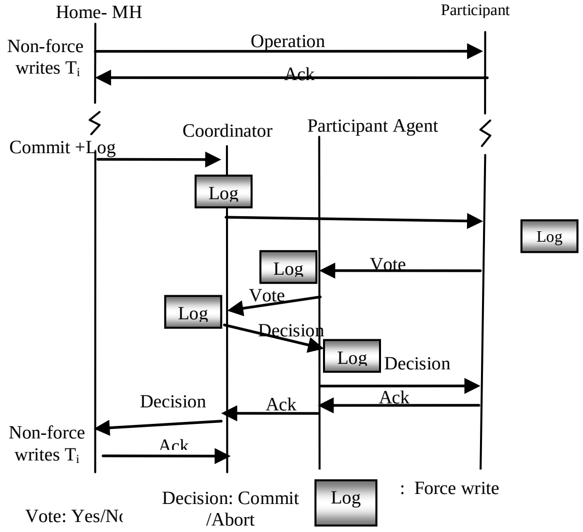


Figure 2. M-2PC – Mobile 2PC Protocol

The Home-MH (participant) sends the transaction to be carried out in batches to BS (Coordinator). When hand-off occurs the incomplete transaction information is transferred to the new BS which becomes the new coordinator. With each change in location, the MH may require to send the message to inform to the old BS that handoff may need to be achieved. While carrying out the process the participants and the coordinator may communicate between each other, so that all are involved during commit.

This solution may give a way to deal with mobility at application layer and embeds the mobility mechanism in the protocol. In M-2PC no message concerned to the protocol execution must be lost during a disconnection or a handoff. During disconnections the continuity of service is guaranteed because of three-tier architecture, where the agents (Coordinator for the mobile client and participant-agent for the mobile server) execute on behalf of the MHs.

During handoff, the MHs are in charge of telling their correspondents about the new location by transferring them a message after registering in a new cell. Also no loss of messages appears during a period of handoff processing (supports the disconnection handling and mobility control). This solves the problem of the address change. The MH must record the identity and location information of the correspondent as it needs when it registers in new BS. Also, there will be no loss of messages, while on the handoff processing. The drawback of this protocol is that it is not capable to handle disconnection and handoffs simultaneously.

C. Unilateral Commit Protocol (UCM)

The Unilateral Commit Protocol for Mobile (UCM) [8] environment supports off-line transaction execution and decreases the risk of abortion of such transaction during reconnection moment. It also supports the disconnection of one or more participants while commitment of the execution of the protocol and is particularly designed for mobile environment, which is based on the idea of single phase commit protocol.

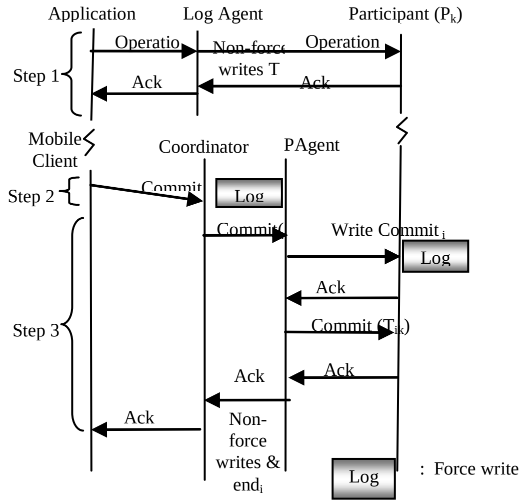


Figure 3. UCM – Unilateral Commit Mobile Protocol

Its message complexity is quite low (a single phase to commit the transaction), thereby saving an essential communication cost in wireless environment. UCM removes the voting phase of 2PC during which the coordinator verifies that participants can guaranty ACID properties or not. Having these properties assured at commit time at each participant site p_K, these operations are logged by log register (force write) and locally executed.

Each operation is acknowledged up to the request. Once all the acknowledgements are received by the application, it issues a commit request. The transaction operations and their acknowledgements are commonly logged to make sure the atomicity. If the transaction reaches validation phase then the global decision is commit. If any problem arises with the global transaction it is immediately aborted

At this point, ACID properties are locally guaranteed by the participants for all the local transaction branches. It reduces the cost of wireless communication by reducing the message complexity. A global commit is performed in a single phase, the decision phase. It is initiated by the transaction's operation log transfer from the application to the coordinator. Fig. 3 shows the sequence of operations carried out by the UCM.

The major issues related to UCM are, blocking situations: UCM coordinator waits if at least single Ack message is missing. Handoff/Mobility: with the UCM the handoff/Mobility problem was not particularly taken care.

D. Timeout Based Commit Protocol (TCOT)

"Transaction Commit on Timeout (TCOT) [4]," is based on a "timeout" approach for Mobile Database Systems, which is generally used to reach a final transaction termination decision (e.g. commit, abort, etc) in any message-oriented system. The transaction is being initiated and fragmented by the MH; the initial fragment is executed at MH while the remaining will be sent to coordinator. The coordinator distributes these left over fragments among the relevant DBSs (Data Base Server).

Let E_t being an upper bound of the execution time, just long enough to allow a fragment to successively finish its execution on participant site and S_t be the upper bound of data shipping time from MH to DBS. If timeout (Max $(E_t + S_t)$) occurs before the log arrives or not all the commit messages are received, the coordinator informs to all the participants about a global abort decision. A participant can unilaterally abort and inform the coordinator. A global commit is decided by the coordinator if it receives the updates log from MH before S_t expires and the commit messages from all participants. Moreover a static or moving coordinator is feasible in case of mobility. Fig. 4 Illustrates the sequence of transaction execution carried out by TCOT.

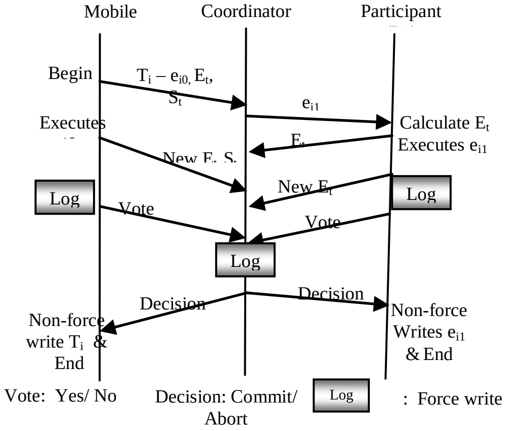


Figure 4. TCOT –Timeout Based Commit Protocol

TCOT is specifically suited for wireless environment; timeout mechanism is the only way to reduce the impact of slow and unreliable wireless link. In case if it's quickly moving and frequent disconnection the abort rate increases, also the message rate can increase. With increase of MPL (Multi Programming Level) the performance degrades. TCOT performs well in an environment where the communication over wireless connection which is highly available and reliable. Timeout not only enforces the termination condition but also the entire execution period as well.

TCOT commits transaction in minimum number of uplinks (user to server direction) by permitting every processing host participating in the transaction to have independent decision making capability based on the timeout mechanism. TCOT is designed for a system offering a connectivity mode known as, Mobile connectivity which permits the users to remain connected all the time to the network by the wireless channel. In Intermittent connectivity mode, the user voluntarily decides as to connect/ disconnect from/to network.

E. Reliable Timeout Based Commit Protocol (RTBCP)

We have proposed the commit protocol to implement the mobile transactions called Reliable Timeout Based Commit Protocol (RTBCP) [9], which is based on the "timeout" approach for mobile database systems to reach transaction global decision. This execution model has the Mobile Host (MH**)** and the Base Station (BS) communicating with each other through messages**.** The Mobile Transaction (MT) is initiated by the MH and is executed either at mobile host or at the fixed hosts. Hence it uses distributed mode of execution between MH and the data base servers (DBS) available at wired network, henceforth these data base servers are called as Fixed Cohort Units (FCUs).

The designed algorithm in [9] depicts Transaction execution at Mobile Host (MH), initiates transaction T_i and split T_i into set of fragments, the first fragment is executed at MH and the remaining fragments of T_i are sent to coordinator (CO) available at the base station. The CO distributes these fragments among various fixed cohorts (FCU) at the wired network. Let E_t being an upper bound of the execution time, i.e. just long enough to allow a fragment to successfully finish its entire execution on participant site. Upon receipt of their respective fragment, each participant calculates e_{tk} (time required to execute fragment at site k), Since we use logs and databases locally it is not necessary to calculate data shipping time S_t [as in TCOT] from the MH to the DBS. If the timeout (Max (E_t)) expires before all the commit messages are received, the coordinator informs all the participants about global abort decision. A participant can unilaterally abort and inform the coordinator. A global commit is decided by the coordinator if it receives commit messages from all the participants.

Determining the value of E_t practically need more rational verification. In case of handoffs and frequent disconnection the abort rate increases, accordingly the message rate also increases. With increase in number of transactions the performance degrades. RTBCP mainly suits for a wireless environment with highly available and reliable wireless link, with timeout mechanism it avoids blocking situations and by maintaining logs locally it reduces the commit time which produces good performance over slow and unreliable wireless link.

At this point, ACI properties are locally guaranteed by the participants for all the local transaction branches. However, since participants are not aware of the termination of the transaction, they cannot guarantee the Durability property.

Durability is ensured by the coordinator itself based on the messages received from all the participants within timeout duration, which gets the T_i log produced by all the nodes and force-writes on stable storage and at the same time, the coordinator then broadcasts the *Commit* decision to all participants and forgets the transaction and will not wait for their acknowledgments. Hence there is no problem of blocking I/O. Based on the global decision MH & FCU's update their databases. Once this is achieved, the ACID properties are guaranteed altogether for all the transaction branches. If T_i fails to commit, then it may initiate cascade rollback.

The absence of an abort message after E_t expires indicates a global abort. Thus, the commit time is the time indicated by E_t. A premature abort is indicated by an abort message. In 2PC [7], the FH waits for messages from participants to make any decision. If the wait is over unsuccessfully, then it aborts the transaction. In TCOT, the absence of a message is enough to make a decision, thus no additional phase is necessary. The RTBCP also works on timeout, all the participants (MH & FCUs) decide to abort and they will not wait for any coordinator decision. Fig. 5 shows the transaction's execution at MH & FCUs.

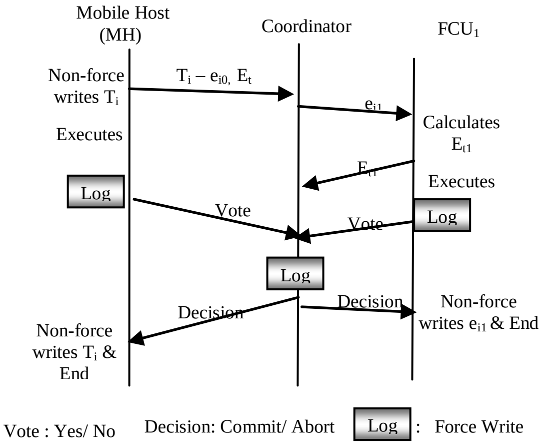


Figure 5. RTBCP -Transaction executions at MH

F. *Modified Reliable Timeout Based Commit Protocol (MRTBCP)*

The Modified Reliable Timeout Based Commit Protocol is one-phase commit protocol which is an extension to the RTBCP. It supports off line execution, disconnection and mobility. It eliminates voting phase of 2PC during which the coordinator verifies that the participants can guarantees ACID properties.

The MRTBCP initiates and fragments the transaction (T_i) at transaction manager at MH(TM-MH), the first fragment e_{i0} is being executed at MH and the remaining fragments of T_i i.e. T_i - e_{i0} are sent to the various participants (MH and part-FHs) for execution. Once the participants receive their respective fragments, they compute and send E_t to the TM-MH, after receiving all E_ts , the MH calculates maximum Time T_m= Max (E_{t0},E_{t1},…..,E_{tn}) required to execute the transaction at MH & Part-FHs.

While executing at Participant, if time expires before it acknowledge TM-MH, then the TM-MH decides to abort and issues an abort request to CO. The participant can unilaterally abort the transaction, if it does not receive the acknowledgement for commit before time expiry.

A global commit is decided by the MH, if it receives an acknowledgement from all the participants before time expiry. Once all the acknowledgments are received by the participants, the TM-MH issues a commit request to CO. The coordinator force-writes and delegates the commit messages to participants at wired network and waits for an acknowledgement. After receiving all acknowledgments the coordinator informs the TM-MH, about the decision.

The transaction's commit and acknowledgment messages are continuously logged at Log Agent to ensure atomicity. If the transaction reaches validation phase then the global decision is to commit or else the transaction is immediately aborted. The maintenance of databases and log autonomously at each participant insures the proper recovery in case of failure. At this point, ACID properties are guaranteed by the participants for all the local transaction branches. Fig. 6 outlines the sequences of execution transaction at MH and FH using MRTBCP.

The MRTBCP also takes care of handoff problem, specifically to handle the situations in case of mobility. Although the handoff process leads to decrease in the performance with MH, increasing the frequency of handoffs does not introduce an additional degradation in performance.

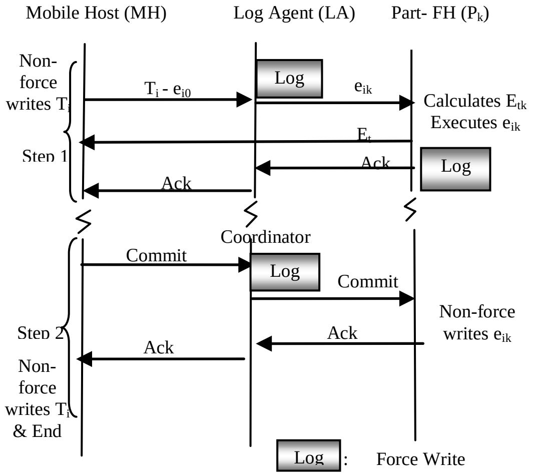


Figure 6. MRTBCP- Modified Reliable Timeout Based Commit Protocol

The performance with participants is not affected by the frequency of the handoff process. This is because, after the submission of a mobile transaction to its participant, there is no need to transmit messages between the coordinator site and the participant until the transaction is completed. Therefore the search process for the coordinator site is required too rarely to affect the performance. It is mainly suited for a wireless environment; a timeout mechanism avoids the blocking situations, since we use logs and databases locally, it reduces the commit time producing good performance metrics over slow and unreliable wireless link. Figure 6.3 shows the handoff management approach in case of mobility with MRTBCP.

G. Single Phase Reliable Timeout Based Commit Protocol (SPRTBCP)

The basic idea of SPRTBCP is to eliminate the voting phase of the 2PC by introducing the properties of the local databases. In this context, a transaction is initiated by the TM-MH (Transaction Manager at MH) and this transaction is assured to be committed in a failure free environment by distributing the fragments at various participants (Part-FHs and MHs). When the acknowledgments for all fragments of a transaction T_i are received by the TM-MH, it means, the transaction fragments i.e. $e_{i0,}$ $e_{i1,}$ $e_{i2,\ldots,}$ e_{in} have been successfully executed till completion. At this point, TM-MH submits its positive commit message to the CO which can directly ask each participant host accessed by the transaction T_i to commit, with no synchronization between the sites. If a transaction fragment, say e_{ik} is aborted by Participant$_k$ during its execution for any problem, the CO simply asks each accessed participant to abort that transaction. Assume that Participant$_k$ crashes during the one-phase commit of transaction T_i during which T_i may have been committed at other hosts. To ensure T_i's atomicity, the effects of the transaction branch e_{ik} have to be forward recovered in Participant$_k$.

The participants executing their respective fragments launch positive acknowledgement and also update their local logs that contain physical redo log records generated during the execution of this operation along with the respective Log Sequence Number (LSN). The CO registers the commit decision in its own log. Once Participant$_k$ recovers from its crash, it redoes set of operations using local log records with highest LSN and reinstalls them in the database.

To enforce transaction atomicity with the site autonomy, SPRTBCP utilizes logging schemes introduced in their respective participant's database systems. On each participant site, local logs keep up each operation sent to it before its execution. During the decision phase, when a participant receives the commit decision, it updates the local database.

If the local database crashes before completing the commit, it will abort the transaction. After the database recovery, the Participant re-executes all operations found in its log and belonging to the globally committed transaction. This approach guarantees global atomicity while preserving site autonomy. To achieve high performance and throughput, transactions are to be interleaved and executed concurrently. We assume that the concurrent executions of transactions are coordinated such that there is no interference among them.

In order to recover from failures, SPRTBCP maintains logs locally with each of the participants. Indeed, maintaining the logs locally, the CO must guarantee that the decision must be non-force written in stable storage before broadcasting its decision. In case of a participant crash during the one-phase commit, the failed transaction branches will be re-executed due to the operations registered in their respective redo logs.

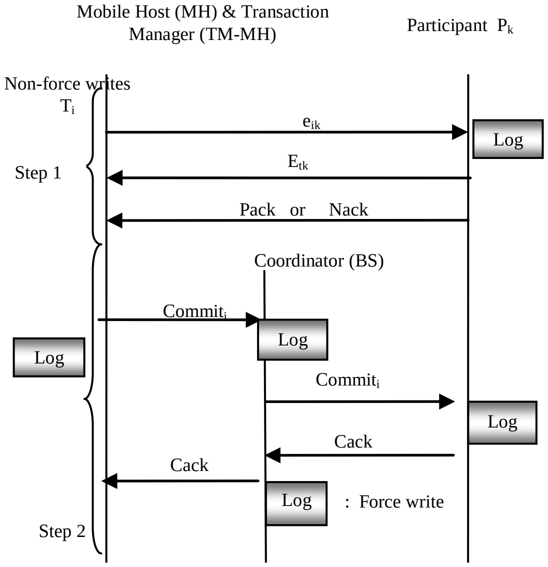


Figure 7 Single Phase Reliable Timeout Based Commit Protocol (SPRTBCP)

Fig. 7 gives the sequence of executions carried out during the transaction processing and shows the series of operations scenario introduced by the SPRTBC protocol.

1) Disconnection

In the presence of disconnections, a large number of transactions can be expected to abort. While executing any transaction, the coordinator mainly concerns with the wireless network, MHs and its current location. During processing, the MH might be either connected to the network fully or totally disconnected or partially connected or weak connection (very low bandwidth).

Specifically, all transaction fragments which undergo a disconnection will be either aborted by the coordinator when it fails to receive a response from a participant site before timeouts or blocked if coordinator cannot proceed until it has collected all the necessary responses**.**

The designed model proposes the system to operate independently even during total disconnection. In case, a mobile host physically detaches from the network, SPRTBCP has enough information locally available for its autonomous operation during disconnection. SPRTBCP maintains log and database locally at each host supporting offline execution providing less number of aborts during disconnection.

2) Handoff/ Mobility Management

The mobile host on movement from one mobile cell to another, it connects to the new MSS. The movement of the local transactions execution should support the mobility across different mobile cells; while the shared transactions support the mobility of standard transactions across different mobile sharing areas when the mobile host is moving across different mobile cells.

III. Performance Analysis Of Various Protocols

This segment represents the presentation study of above mentioned commit protocols used for mobile environment. A study provides the performance metrics used to evaluate the performance of commit protocols viz. by means of number of message transfers, writing disk of log records (force writes), blocking property, and real-time atomic property, impact of frequency disconnection, latency, handoff association.

The performance of SPRTBCP is compared with 2PC, M-2PC, UCM, TCOT, RTBCP and MRTBCP. The 2PC is the most well-known blocking commit protocol, while M-2PC is improved version of 2PC. UCM is one phase commit protocol that has been proposed for light weight processing and TCOT is timeout based non-blocking commit protocol.

The number of messages presented includes the execution and the commitment phases. In TCOT with normal execution only 2 message rounds are required whereas M-2PC requires minimum of 3 message rounds but RTBCP message complexity is almost similar to TCOT. The UCM and MRTBC Protocols use only one phase for commitment of transaction, since they add logging messages at base station, during commitment phase of transaction, due to which the failures are handled effectively but there is a small increase in the message complexity.

To overcome the problem, we have designed an extended version of MRTBCP called "Single Phase Reliable Timeout Based Commit Protocol (SPRTBCP)". It maintains log locally at each mobile and fixed host participant. SPRTBCP proposes one phase, reliable, efficient and non-blocking atomic transaction commit protocol.

Table I summarizes the principal properties of protocols studied over. To commit a transaction, the best protocol in terms of wireless messages is UCM. This is obtained at price of making strong assumptions about the local concurrency and recovery mechanisms. This may limit its usability in arbitrary heterogeneous systems. TCOT adopt the latest approaches which are completely different from 1PC or 2PC protocols. The other protocols preserve 2PC principles and try to optimize it to fit mobile environment requirements.

When the number of nodes increases in a cell, the performance of the transmission channel decreases as soon as flow threshold is reached. The conflict rate also increases leading to decreasing throughputs. The TCOT protocol gives very good performance when the MPL (Multiple programming level) is low. With an increasing MPL its performance deteriorates significantly, hence it becomes the less powerful protocol. With an increase in MPL the throughput of 2PC becomes closer to UCM & M-2PC.

TCOT has the best latency in case, with and without mobility of nodes, because the timeout limits the processing time of a transaction in all cases. But high value timeouts may lead to good throughput but increase latency.

A small value of Δ_t (Extension requests) may generate a large number of T_i aborts or request for the extension of Δ_t may affects throughput and message cost. Table 1 describes the impact on various commit protocols due to the disconnection.

Impact of handoff on M-2PC cannot be included as it is designed for supporting mobility management. UCM does not support mobility. TCOT handles mobility and disconnections but compared to RTBCP, it increases commit time [9] hence reduces overall throughput. MRTBCP also supports for the mobility, disconnections and handoff. It is one phase and timeout protocol hence produces better performance by reducing latency, message complexity and increased throughput over other protocols.

TCOT, RTBCP, MRTBCP and SPRTBCP are semantic based commit protocols, they eliminate the uncertainty period of transaction termination and the blocking effects, where they allow a participant to unilaterally commit transaction and release the resources it holds.

If the final decision is global abort, compensation is used semantically to undo the aborted transaction effects. The protocols like UCM, 2PC follow strict atomicity where they follow traditional ACID requirements. Table 1 gives the detailed analysis of the performance of various protocols used for the commitment of mobile transactions.

A. Performance metrics

The proposed technique is simulated and performance metrics are analyzed. The results of experiment are presented to verify the performance of the SPRTBCP protocol

1) Effect of Disconnection on Commit Rate

Disconnection may also occur involuntarily and unpredictably. Figures 8(a) and (b) depict the effect of disconnection on commit rate. SPRTBCP commit rate is compared based on timeouts with TCOT, RTBCP and MRTBCP (Figure 8(a)). SPRTBCP makes use of single phase operation and maintains log and database locally at each host supporting offline execution. Because of which, SPRTBCP completes transaction execution efficiently in case of disconnections producing higher commit rate compared to other protocols.

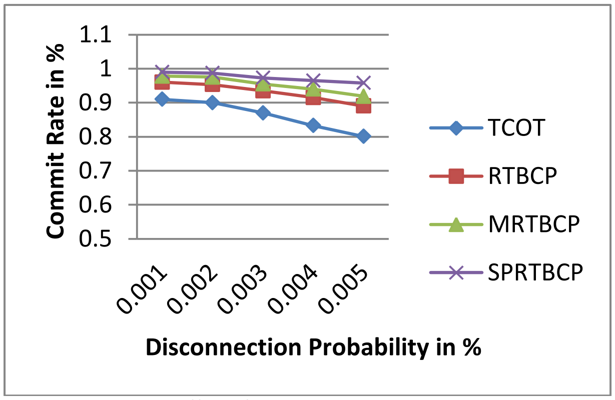


Figure 8(a) Effect of Disconnection on Commit Rate

Figure 8 (b) shows the commit rate compared with 2PC, M-2PC, UCM, RTBCP and MRTBCP considering disconnection as abort. The results show that SPRTBCP provides almost more than 95% of commit rate with the disconnection probability of 0.005%. If the traditional 2PC is executed in mobile environment, the number of disconnections increase, leading to transaction aborts i.e. the CO tries to communicate with a disconnected MH will cause blocking of resources. As a result, the commit rate reduces. In case of M-2PC and UCM, message overheads lead to decrease in the commit rate.

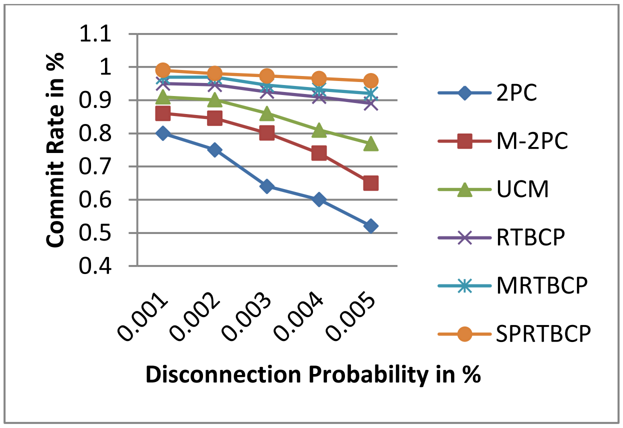


Figure 8(b) Effect of Disconnection on Commit Rate

2) *Effect of Disconnection on Abort Rate*

As frequent are disconnections, as transaction abortions are. This is not acceptable in mobile environments because frequent disconnections are not exceptions but rather are part of the normal mode of operation, so they should not be treated as failures. Contrary to the traditional 2PC, a protocol must not account on MHs to be continuously available to participate in the transaction commitment. Due to maintenance of log and database locally SPRTBCP tolerates disconnections providing less number of aborts.

Figure 9(a) shows the effect of disconnection on abort rate in comparision with TCOT, RTBCP and MRTBCP. It is observed that with the single phase commit operation and having offline execution without Log Agent at the wired network, SPRTBCP produces less number of aborts compared to the other protocols.

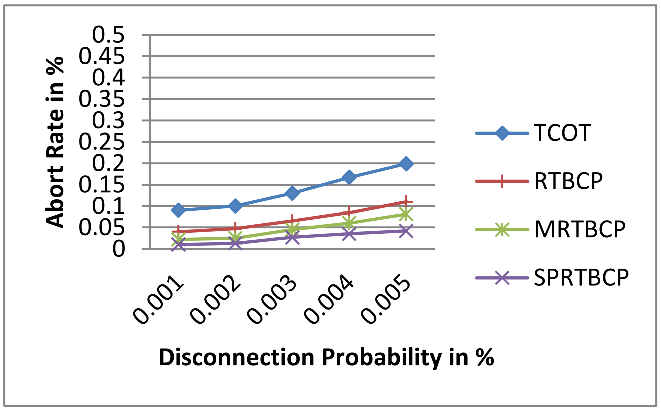


Figure 9(a) Effect of Disconnection on Abort Rate

Figure 9(b) shows the effect of disconnection on abort rate compared to the other protocols. We can observe that as the probability of disconnection increases the abort rate will also increase.

It is verified that SPRTBCP produces more efficient results (having very less abort rate) compared to 2PC, M-2PC, UCM, TCOT, RTBCP and MRTBCP and proved that SPRTBCP is more reliable in case of disconnections.

TABLE I: PERFORMANCE OF COMMIT PROTOCOLS

Protocol	**No. of phases**	**Atomicity**	**Site of transaction execution**	**Message complexity**	**Impact of frequency disconnection**	**Impact of latency**	**Handoff management**
2PC	2	Strict	FH	4n	Increase in number of aborts	Bad	Registration level
M-2PC	2	Strict	MH & FH	4n-1	Increase in number of Aborts until resources released	Medium	Protocol level
UCM	1	Strict	MH & FH	2n	Delay local transactions	Good	-----
TCOT	1	Semantic	MH & FH	$(2n-1) + \Delta_t$ Δ_t: no. of timeout extensions	Increase in number of Aborts	Good	Registration and protocol level
RTBCP	1	Semantic	MH & FH	$(2n-1)+ \Delta_t$	Increase in number. of Aborts	Good	Registration and protocol level
MRTBCP	1	Semantic	MH & FH	$(2n-1) + \Delta_t$	Increase no. of Aborts	Good	Registration and protocol level
SPRTBCP	1	Semantic	MH & FH	$(2n-1)+ \Delta_t$	Increase in number of aborts	Good	Registration and protocol level

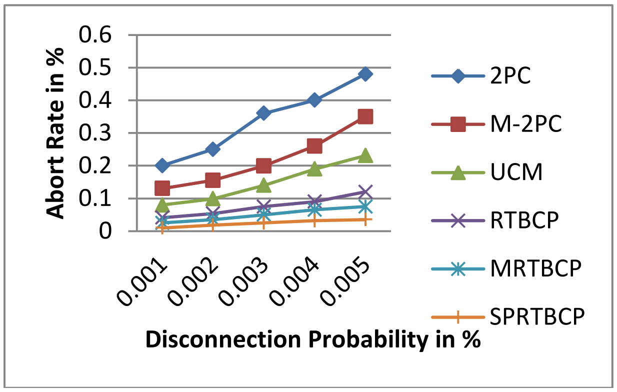


Figure 9(b) Effect of Disconnection on Abort Rate

IV. Conclusion

Transactions are main building blocks to have reliable systems. It could be provided by guaranteeing data consistency, concurrency control and recovery in case of failures due to disconnections and handoff.

This paper presents an overall comparison analysis of commit protocols for mobile environment. The study is mainly based on the performance metrics Viz. impact of frequency disconnection, latency, handoff management etc.

This revise presents the ongoing research which consists of the design and experiment of dedicated protocol that satisfies as several requirements of mobile surroundings as possible. To therapy to this condition a new atomic commitment protocol devoted to mobile and distributed computing is extremely desirable.

Hence, we proposed Single Phase Reliable Timeout Based Commit Protocol as extension to Modified Reliable Timeout Based Commit Protocol that increases a new commitment protocol, which suits for mobile transactions, designed to preserve all the above performance metrics. The protocol aims at handling new challenges including site failures and message loss blocking-free manner.

In addition to MRTBCP, it is single phase commit protocol without the Log Agent, due to which it reduces message complexity and average commit time. It is proved that, even in case of disconnections, failures and during mobility SPRTBCP produces better performance and reliable execution of transactions compared to the existing ACPs like 2PC, M-2PC, UCM and TCOT.

References

[1] L. Amanton, B. Sadeg, S. Saad, Bouzefrane, “Disconnection Tolerance in Soft Real Time Mobile Databases,” Proceedings of 16th International Conference on Computers and Their Applications, USA. 2001,pp. 98-101.

[2] B. Bose, S. Sane, “Distributed Timeout Based Transaction Commit Protocol for Mobile Database Systems,” in Proceedings of International Conference and Workshop on Emerging trends in Technology (ICWET-2010),2010, pp. 518-523.

[3] C. Bobineau, C. Labb’e, C. Roncancio, P. Serrano- Alvarado, “Comparing Transaction Commit Protocols for Mobile Environments,” in Proceedings of the 15th International Workshop on Database and Expert Systems Applications (DEXA’04) 1529-4188/04 IEEE.

[4] V. Kumar, N. Prabhu, M. H. Dunham, and A. Y. Seydim, “TCOT- A Timeout-Based Mobile Transaction Commitment Protocol,” *IEEE Transactions on Computers*, 51(10), 2002.

[5] P. Serrano, C. Roncancico, M. Adiba, “A Survey of MobileTransactions,” DAPD Jnl., 16(2), 2004.

[6] M. M. Goreyand , R. K. Ghosh, “The Recovery of Mobile Transactions,” in Proceedings of the 11th International Workshop on Database and Expert Systems Applications (DEXA’00) 2000 pp. 23, IEEE.

[7] N. Nouali, A. Doucet, and H. Drias. “A two-phase commit protocol for mobile wireless environment.” in H. E.Williams and G. Dobbie, editors, *Sixteenth Australasian Database Conference (ADC2005)*, volume 39 of *CRPIT*, pages 135–144, Newcastle, Australia, 2005. ACS.

[8] C. Bobineau, P. Pucheral, and M. Abdallah. “A Unilateral Commit Protocol for Mobile and Disconnected Computing”, In *PDCS*, USA, 2000.

[9] B. Harsoor, S. Ramachandram, “Reliable Timeout Based Commit Protocol”, in Proceedings of 2nd International Workshop on Trust Management in P2P Systems (IWTMP2PS-2010) CNSA-2010, Springer Verlag 2010, pp. 417-423.

Authors Profile

Mrs. Bharati Harsoor received her bachelor’s degree in CSE (1995), Masters in Computer Science (2001). She is a Research Scholar at Osmania University; Hyderabad.She is presently working as Associate Professor, Department of Information Science, PDA College of Engineering, Gulbarga, Karnataka State, India and Published papers in various National/International Conference and Journals. Her areas of interests include Mobile Computing, Databases, and Software Engineering. She is member of Institute of Electronics and Telecommunication Engineers (IETE).

Dr. S. Ramachandram (1959) received his bachelor’s degree in Electronics and Communication (1983), Masters in Computer Science (1985) and a Ph.D. in Computer Science (2005). He is presently working as a Professor and Head, Department of Computer Science, University College of Engineering, Osmania University, Hyderabad, India. His research areas include Mobile Computing, Grid Computing, Server Virtualization and Software Engineering. He has authored several books on Software Engineering, handled several national & international projects and published several research papers at international and national level. He also held several positions in the university as a Chairman Board of Studies, Nodal officer for World Bank Projects and chair of Tutorials Committee. He is a member of Institute of Electrical and Electronic Engineers (IEEE), Computer Society of India (CSI) and Institute of Electronics and Telecommunication Engineers (IETE).

Optimized Pessimistic Fibonacci Back-off Algorithm (PFB)

Muneer Bani Yassein
College of Computer and
Information Technology
JUST Amman, Jordan

Mohammed Ahmed Alomari
College of Computer and
Information Technology
JUST Amman, Jordan

Constandinos X. Mavromoustakis
Department of Computer Science
University of Nicosia
Nicosia, Cyprus

***Abstract*— MANET is a self-directed system consisting of mobile nodes, which can be either routers and/or hosts. Nodes in MANET are connected by wireless links without base stations. The Backoff algorithm considered as a main element of Media Access Control (MAC) protocol, which is used to avoid collision in MANET's. The Fibonacci Backoff algorithm and the Pessimistic Fibonacci Backoff are proposed to improve network performance depending on contention window size. This research introduces a new hybrid Backoff algorithm called Pessimistic Fibonacci Backoff (PFB) Algorithm which merges the two previous algorithms in order to find the most proper contention window sizes that reduce collisions as much as possible. This research takes into consideration and evaluates each of the following main measurements: Packet delivery ratio, normalized routing load and end-to-end delay. Based on the extracted simulation results, PFB algorithm outperforms Pessimistic Linear-Exponential Backoff (PLEB) by up to 76%,40.41%, 31.88% in terms of Packet delivery ratio, end-to-end delay and normalized routing load respectively, especially in the sparse environments. All of the simulation results are obtained by the well-known NS-2 Simulator, version 2.34, without any distance or location measurements devices.**

Keywords- Back-off; collision; end-to-end delay; normalized routing load; packet delivery Ratio; MANET's; PLEB, PFB; and MAC.

I. INTRODUCTION

Wireless networks (WN) use radio signals to communicate among computers and other network devices. WN's are getting popular nowadays due to easy setup feature. World is now moving to this type of communication. Today, this vision is being challenged by various forms of mobility, which are effectively reshaping the landscape of modern distributed computing. Mobile wireless networks consist of two kinds of mobile networks, infrastructure-based and ad-hoc wireless networks.

Wireless sensor networks (WSN) corresponds as a communication way and supports the random movement of the nodes. The main features and challenges of WSN are [8]: low cost devices, large scale of deployment, end-to-end quality of service, energy-efficient devices, and secure operation.

A Mobile ad-hoc Network (MANET) is a self-directed system consisting of mobile nodes (including routers and hosts) connected by wireless links. MANETs have received substantial attention due to their strong features, such as Multi-hop Routing, distributed operations, dynamic topology, capacity and light-weight terminals [1,2].

The Media/Medium Access Control (MAC) protocol is a sub-layer of the data link layer, which provides a control mechanism to allow packet transmission through the wireless network. Within MANETs, MAC protocol consists of several key parts, such as the Backoff mechanism which solves the collision problem which occurs when more than one node transmits data simultaneously due to single transmission restrictions through the channel [3,4]. To realize how the infrastructure wireless networks function, if there are two computers, each one contains a wireless adapter to connect with an access point. When data is transmitted using a wireless router as a binary data, then the operation at the receiver will follow a vice procedure [4, 5]. The base stations are the bridges within these networks, having a role of [1, 6]: linking each mobile node with the nearest base station (Within node Range) and allowing the communication between them.

This research proposes a Backoff algorithm called Pessimistic Fibonacci Backoff (PFB) Algorithm to reduce the differences between successive contention window sizes using delay parameter, normalization and Packet Delivery Ratio. Its efficiency measures by merging more than one increment behavior in order to have long waiting times when a collision suddenly occurs. This algorithm uses the Pessimistic Linear Exponential Backoff algorithm introduced in [7] where the structure replaces the linear and exponential waiting time with an exponential, cubic, and a Fibonacci series.

The rest of this paper is organized as follows; Section II covers the motivations behind this work, Section III covers the methodologies and presents the related work in order to give a better understanding for the Back off algorithms utilized in the past by other researchers. Section IV presents the proposed Pessimistic Fibonacci Back off algorithm proposed whereas Section V covers the results and introduces the analysis of these results. Finally, the last Section concludes the paper in the last section.

II. MOTIVATIONS

In MANETs, the Backoff field considered to be one of the researcher's main areas in order to achieve an efficient Backoff algorithm [3]. In literature, many Backoff algorithms were proposed such as: Binary Exponential Backoff (BEB), Fibonacci Increment Backoff (FIB), Logarithmic Backoff (LOB), Pessimistic Linear-Exponential Backoff (PLEB) and

Optimistic Linear-Exponential Backoff (OLEB). Binary Exponential Backoff is considered as standard Backoff algorithm in MANETs.

The PLEB Algorithm used in IEEE 802.11 MAC protocol provides exponential and linear increments in contention window values depending on Backoff timer (BOT). These two incremental behaviors make BOT increasing quickly. Therefore, the need to develop a powerful Backoff algorithm is more timely in Wireless Networks in order to enable more features such as the desired stability, the minimum network overhead, the power consumption and the maximum network throughput. All these factors lead to achieve the better network performance.

III. Related Work

The Fibonacci Backoff Algorithm reduces the differences between concurrent contention window sizes using a mathematical Fibonacci series as in equation 1 to reduce the difference between successive contention window sizes.

$$F(n)=Fn-1+Fn-2 \ where\ F0=0\ and\ F1=1 \qquad (1)$$

F(n) represents the new contention window size, leading to a smaller increment on large window sizes. It checks if the channel is idle and then the Backoff is reduced one unit. Else, it sends the packets if the channel is idle for more time until the Backoff time has a zero value as examined in [10, 11].

This mechanism leads to decrease the expected wait time by reaching a large window size for a specified node, allowing this node to access the shared medium and thus to avoid increasing in channel idle times. The Fibonacci series has a valuable characteristic to provide the precise value which obtained by the ratio of successive terms in the Fibonacci series [15].

The Backoff algorithms were proposed in order to avoid the collision and to resolve contention among different nodes as well as to improve the network resources utilization. Once there is a collision, retransmission delay occurs which indicates the nodes' needs to make a distinction in terms of time for a period time. Backoff time value is chosen randomly from bounded contention window which varies according to the consequence of the latest tries of the transmission based on the number of active nodes and traffic load in the network.

The Backoff algorithm takes place and it is performed in each of the next cases:

- If the channel is busy before the first transmission.
- After each attempt of retransmission
- After the node is successfully transmits a package.

Most of proposed Backoff algorithms behave unsatisfactory in case of failure to transmit rapidly the data in case of increasing the contention window; this clearly appeared in binary exponential Backoff algorithm. Other algorithms do not enable the node with enough time before retransmitting, which results in more power usage and more network overhead. For example, the exponential increment of BEB algorithm which is used in standard IEEE 802.11 MAC does not achieve the best performance due to large Contention Window (CW) gaps produced. Another example is a linear increment of Linear Multiplicative Increase and Linear Decrease (LMILD) [9]; it does not allow a sufficient Backoff time before data retransmission.

The Backoff Algorithms are divided into two categories [1, 10]:

- Static Backoff Algorithms

In this category, Backoff algorithms proposed using a fixed Backoff wait time period which all nodes have a fixed Backoff waiting time based on the equation 2.

$$Backoff\ Timer{=}I,\ where\ I\ is\ a\ \text{fixed}\ integer \qquad (2)$$

- Dynamic Backoff Algorithms

In this category, Backoff algorithms proposed using a variant Backoff wait time period by randomly choosing the Backoff timer value depending on equations 3 and 4 [11]:

$$CWnew\begin{cases} Max(f(CW)\,,(Cwmax)\,,after\ I\ transmission \\ Min(g(CW),(Cwmin)\,,after\ a\ collision. \\ Min(h(CW),(Cwmin)\,,after\ hearing\ a\ collision \end{cases} \qquad (3)$$

$$BackoffTimer = b, b\ is\ random\ integer \qquad (4)$$

The rest of this section presents the most related work to represents the importance of pessimistic and Fibonacci Backoff algorithms.

S. Manaseer [1] introduces about some Backoff Mechanisms for Wireless Mobile ad-hoc Networks, focused on presenting and illustrating the importance of the two proposed Backoff algorithms [1] called Pessimistic Linear-Exponential, and Optimistic Linear-Exponential. Experimental results demonstrate that PLEB improves the network throughput, and reduces packet delay for large numbers of nodes and large network size with low mobility speed. On the other hand, OLEB has the same experimental results at high-traffic rates.

S. Manaseer, M. Ould-Khaoua and L. M Mackenzie in [13] introduce Fibonacci Backoff Algorithm for Mobile ad-hoc Networks. A Backoff Algorithm called Fibonacci Increment Backoff algorithm was proposed to reduce the differences among successive contention window sizes. Results from simulation experiments revealed that Fibonacci Increment Backoff algorithm achieves a higher throughput than the Binary Exponential Backoff algorithm used in MANETs. N. Song [14], enhanced the IEEE 802.11 distributed coordination function with an exponential increase and exponential decrease Backoff algorithm, whereas [14] also studied the effects of increasing and decreasing the waiting time intervals using exponential Backoff algorithm. Results representing the exponential increase algorithm have a good results using the coordination function. J. Deng, et al. [9] proposed the Linear Multiplicative Increase and Linear Decrease (LMILD) Backoff algorithm. This algorithm had shown a best performance and aims to achieve best Contention Window (CW) size. If failure transmission occurs it uses a factor multiplicative and linear increment; firstly, multiplicative the contention window by colliding nodes when there is a collision, other nodes hearing this collision make a linear increment to their contention window. On the other hand all nodes decrease their contention windows linearly when there is a transmission success.

V. Bharghavan, et al. [10] proposed Multiplicative Increase and a Linear Decrease (MILD) Backoff algorithm. This algorithm aims to solve the unfairness problem by reducing the probability of successful users to access the channel. In this algorithm the contention window size is incremented by a factor multiplication when a transmission failure occurs.

In our work, there are some important issues that should be taken into account when trying to design a Backoff algorithm that aims to improve the performance over the network such as determining the methods used to increase and decrease the CW and selecting suitable increase and decrease factors.

IV. Proposed Pessimistic Fibonacci Backoff (PFB) Algorithm

This section presents the proposed Pessimistic Fibonacci Backoff (PFB) algorithm which tries to organize node transmissions in time to avoid collisions. The PFB algorithm proposed is implemented to operate in collaboration with the IEEE 802.11 MAC protocol in order to reduce the collision in MANETs. This is achieved by incrementing the Backoff time by using a combination of Fibonacci, cubic and exponential increment behaviors.

In general, Backoff algorithms should use an appropriate increase for the contention window size in order to gain the best performance, because it should be incremented after each failure occurs in transmission. Some of the mentioned increment behaviors are appropriate when using a small or medium network dimensions but seems not enough good in large ones [1].

PLEB Algorithm

```
0  Set Backoff timer to initial value
1  While BOT ≠ 0 do
2      For each time slot
3          If channel is idle then BOT = BOT-1
4  If channel is idle for more than IDFS then
5      Send
6  If send failure then
7      If NOB <= N then
8          CW = CW * 2
9          BOT = Rand (x);
10     Else
11         CW = CW + T
12         BOT = Rand (x);
13 Else
14     CW= Initial value
15     BOT = 0
16     Go to 1
17 Stop
Note : 1 ≤ X ≤ CW-1, NOB: Number of Backoffs, N,M: CW Threshold
```

Figure 1. Pessimistic Linear-Exponential Backoff (PLEB) Algorithm

The Pessimistic Linear-Exponential Backoff algorithm as in Figure1, assumes that congestion in the network will take more time to be resolved; it combines linear and exponential increment behaviors by using linear before using exponential at the first stages. PLEB algorithm adopts that a transmission failure is due to the network congestion resulted from the network high traffic load or a larger number of nodes located in a given network area.

The PFB algorithm proposed in next section aims to improve overall network performance mainly to achieve the best data delivery ratio with fewer routing load in the network. This algorithm uses a combination of Fibonacci, cubic and exponential increment behaviors.

Figure 2 presents the new proposed Backoff algorithm which referred as the Pessimistic Fibonacci Backoff (PFB). It assumes that there is congestion which is high enough and cannot be resolved in the near future.

PFB increases the contention window size exponentially and a transmission failure occurs to give a longer waiting time before retransmission, then PFB increases the timer cubically in order to avoid increasing Backoff extremely. After a fixed number of cubic increments PFB starts to increase the timer, using a Fibonacci series to achieve a less dramatic growth of the contention window size, allows nodes to perform more tries to access the channel.

PFB Algorithm

```
0  Set Backoff timer to initial value
1  While BOT ≠ 0 do
2      For each time slot
3          If channel is idle then BOT = BOT-1
4  If channel is idle for more than IDFS then
5      Send
6  If send failure then
7      If NOB <= N then
8          CW = CW * 2
9          BOT = Rand (x);
10     Else
11     If  N < NOB< M
12         CW = CW^3
13         BOT = Rand (x);
14     Else
15         Use Fibonacci CW BOi+1=fib(I)   --- CW = next fib
17 Else
18     CW= Initial value
19     BOT = 0
20     Go to 1
21 Stop
Note : 1 ≤ X ≤ CW-1, NOB: Number of Backoffs, N,M: CW
```

Figure 2. Pessimistic Fibonacci Backoff (PFB) Algorithm

When the congestion is found in the network the PFB adopts the corresponding transmission failure which is caused by a larger number of node's high traffic load within a limited area. When there is a transmission failure, PFB algorithm firstly increases the contention window size, mainly to give a long waiting time before starting the next transmission along the network paths. After that, PFB starts to increase the time using an exponential increment behavior in order to provide adequate values exploiting other paths, and then utilizing a cubically increasing. Finally, it uses the Fibonacci increment behavior in order to increase the CW size more exceptionally.

The intention of using these increment behaviors is to decrease the possibility in breaking paths which increased in a very sparse network, due to mobility and because there is single network path. This algorithm aims to achieve less growth in the contention window size to allow the nodes to access the channel after incrementing Backoff time.

V. Experiments Set-Up

This section explains the main methodology points used to direct this research, such as: implementing PFB, testing it and primary justification. Many parameters were used to measure the network efficiency such as delivery ratio, delay time, throughput, overhead, traffic load, and number of hubs. This work, uses three: end-to-end delay, normalization, and packet delivery Ratio.

A larger size of data was successfully received by nodes over the network due to the reduced amount of contention window size. During the simulation experiments the total number of nodes, the nodes pause time, and the maximum node speed are varied. Results are compared to both standard Backoff algorithm and Pessimistic Linear Exponential Backoff algorithm. NS-2 simulator [16] used to implement Pessimistic Fibonacci Backoff algorithm, and each figure represents the average of twenty autonomous runs. Our simulation is done using a personal laptop with 64 bits Ubuntu open-source operating system, the CPU is AMD phenom™ II N830 Triple-Core of speed 2.10GHz, and 8 GB of RAM.

TABLE I. List Of Parameters

Parameter	Value
Nodes Number	25,50,75,100, 125
Area(x*y)	(900m*900m)
Connection Number	10
Pause time	0,150seconds
maximum speed	1,4,15 meters/second
Packet Size	512 bytes
Number of Packets	10, 20
Simulation Time	900 seconds
Bandwidth	2 Mbps

As mentioned above NS-2 simulator used and we implement proposed algorithm in it. We used the constant Bit Rate (CBR) traffic generator, therefore the sources and destinations are distributed in the network area randomly. The packet size is 512 bytes, sending 10 and 20 packets/second with a 10 number of connections. Nodes move at a random speed distributed in 900 meters X 900 meters mainly to stabilize the network, and to record more accurate results.

In this simulation, the value of proposed algorithm achieved improvement over standard and Pessimistic Fibonacci Backoff algorithm using equation 5.

$$\textit{Improvement} = (\textit{Old Value} - \textit{New value})/\textit{Old}*100\% \qquad (5)$$

VI. Simulation Results and Analysis

The proposed topology simulations assume the following points:

- For the full length of simulation, nodes have sufficient power supply. At no point of the simulation lifetime, a node goes to offline because of lacking power.
- External network interference or noise does not exist. All the data that exist in the network is originated from within the network.
- Each node is equipped with a transmitter/receiver, or transceiver, IEEE 802.11 devices.
- The number of nodes over the network is constant for the length of simulation time. No nodes join nor leave the network for the duration of simulation.

Performing experiments and simulations, ideally is to achieve minimum delay, maximum packet delivery Ratio and minimum normalized routing load. Hence, the following factors will be used to measure the performance of the Pessimistic Fibonacci Backoff algorithm:

- End to End Delay [5]: is the average delay taken to transmit a packet through a network from source to destination. This delay may suffer from media access control retransmission delays and buffering the same time routing discovery.
- Packet delivery Ratio [6]: the number of received data packets by the destination nodes to the total number of transmitted data packets by the source node.
- Normalized routing load (overhead) [6]: the average number of sent routing packets to the total number of received packets.

Figure 3 represents the relation between end to end delay and network density of the STD (STD is the standard used algorithm which is Binary), PLEB and PFB scenarios, while each node has a movement speed up to 1 meter per second, and the transmission rate is 10 packets per second.

Figure 3 show that PFB algorithm achieves the lowest delay with specified area over all numbers of nodes. In case of dense environments (Maximum = 25 nodes), PFB outperforms STD and PLEB by 56.83% and 50.59%, respectively. In contrast, in sparse environments, PFB Outperforms STD by 40.41% and PLEB by 25.49%.

This is due to the fact that in the PFB, the number of increments and slow decrement behavior produced by exponential factor which leads to generate a longer Backoff values, is greater than PLEB, which leads to have a longer waiting times, and drive the contention window size to be larger instead of the need to achieve a minimum delay.

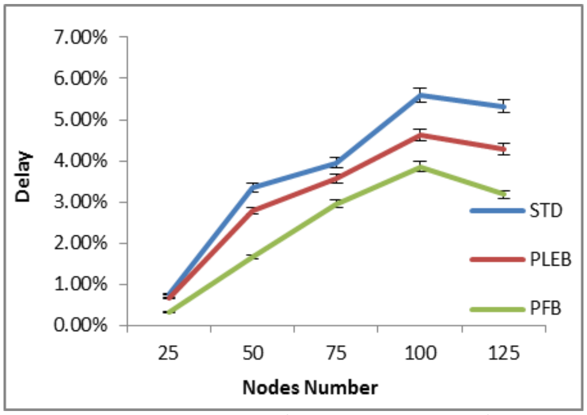


Figure 3: End To End Delay of STD, PLEB and PFB for Sending 10 packet/second, 0 second pause time, and 1 meter/second maximum speed.

Figure 4 shows end-to-end delay of the STD, PLEB and PFB scenarios over increasing number of nodes until reaching 125 nodes, 150 second pause time, while each node has a movement speed up to 4 meters per second, and the transmission rate is 10 packets per second.

This graph shows that PFB algorithm achieves the lowest delay with specified area over all numbers of nodes. In case of dense environments (Maximum = 25 nodes), PFB outperforms STD and PLEB by 13.61% and 10.67%, respectively. In contrast, for sparse environments PFB Outperforms STD by 25.88% and PLEB by 13.15%.

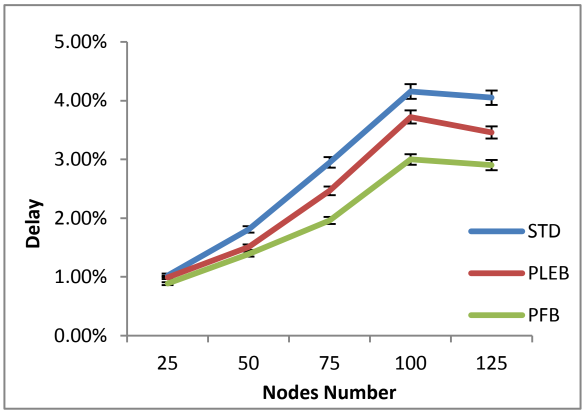


Figure 4: End To End Delay of STD, PLEB and PFB for Sending 10 packets/second, 150 second pause time, and 4 meters/second maximum speed.

Figure 5 shows end to end delay of the STD, PLEB and PFB scenarios over increasing number of nodes until reaching 125 nodes, while each node has a movement speed up to 1 meter per second, and the transmission rate is 20 packets per second. This graph shows that PFB algorithm achieves the lowest delay with specified area over all numbers of nodes. In case of dense environments (Maximum = 25 nodes), PFB outperforms STD and PLEB by 49.44% and 31.22%, respectively. On the other hand, for sparse environments PFB Outperforms STD by 35.99% and PLEB by 31.1%.

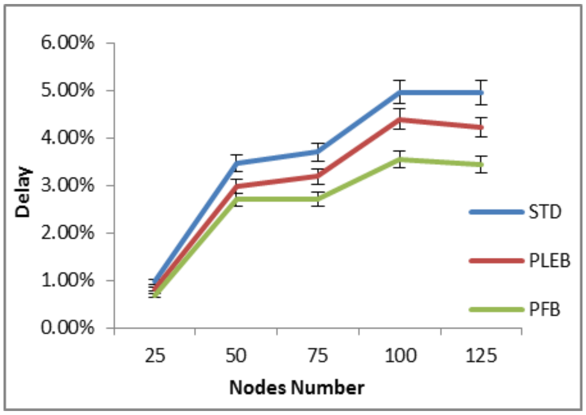


Figure 5: End To End Delay of STD, PLEB and PFB for Sending 20 packets/second, 0 second pause time, and 1 meters/second maximum speed.

Figure 6 shows end to end delay of the STD, PLEB and PFB scenarios over increasing number of nodes until reaching 125 nodes, while each node has a movement speed up to 4 meters per second, and the transmission rate is 20 packets per second. In case of dense environments (Maximum = 25 nodes), PFB outperforms STD and PLEB by 68.26% and 49.33%, respectively. Oppositely, for sparse environments PFB Outperforms STD by 31.80% and PLEB by 21.80%. This graph shows that PFB algorithm achieves the lowest delay with specified area over all numbers of nodes.

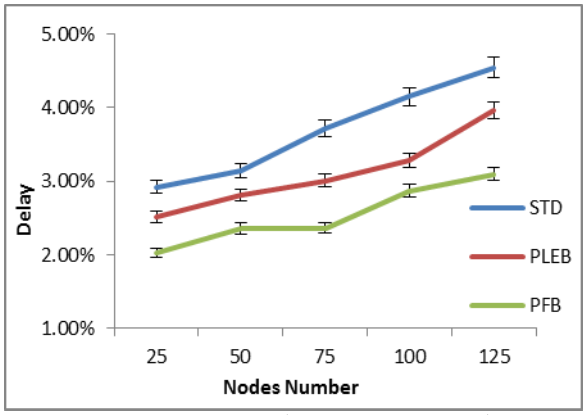


Figure 6: End To End Delay of STD, PLEB and PFB for Sending 20 packets/second, 0 second pause time, and 4 meters/second maximum speed.

Figure 7 shows end to end delay of the STD, PLEB and PFB scenarios over increasing number of nodes until reaching 125 nodes, 150 second pause time, while each node has a movement speed up to 15 meters per second, and the transmission rate is 20 packets per second. This graph shows that PFB algorithm achieves the lowest delay with specified area over all numbers of nodes.

In case of dense environments (Maximum = 25 nodes), PFB outperforms STD and PLEB by 20.83% and 12.10%, respectively. Inversely, on the Maximum number of nodes PFB Outperforms STD by 32.26% and PLEB by 21.19%.

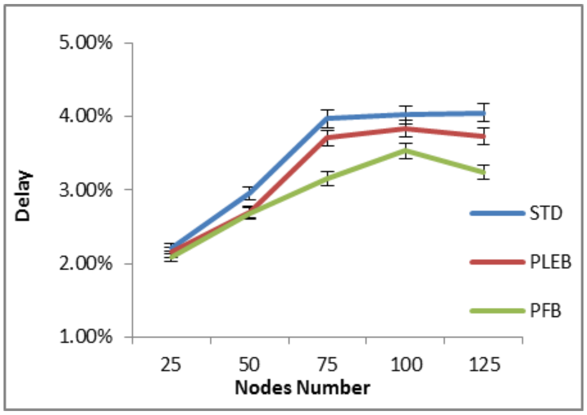


Figure 7: End To End Delay of STD, PLEB and PFB for Sending 20 packets/second, 150 second pause time, and 15 meters/second maximum speed.

Figure 8 shows packet delivery Ratio of the STD, PLEB and PFB scenarios over increasing number of nodes until reaching 125 nodes, while each node has a movement speed up to 1 meter per second, and the transmission rate is 10 packets per second. This graph shows that PFB algorithm delivers more packets than others with the specified area over the total number of nodes. When there are a few number of nodes (maximum = 25 nodes), PFB outperforms STD and PLEB by 00.97% and 00.57%, respectively. Oppositely, on the maximum number of nodes PFB outperforms STD by 00.21% and PLEB by 00.14%.

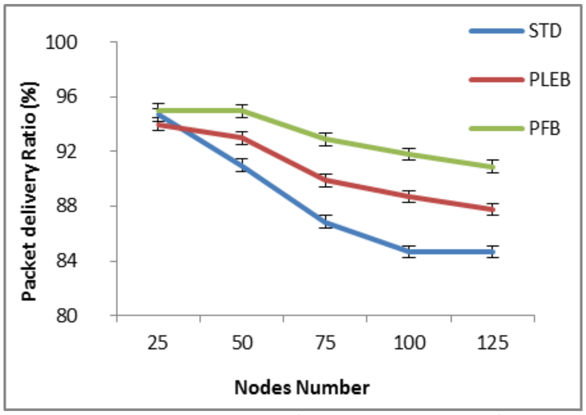


Figure 8: Packet delivery Ratio of STD, PLEB and PFB for Sending 10 packets/second, 0 second pause time, and 1 meters/second maximum speed.

Figure 9 shows the Packet Delivery Ratio of the STD, PLEB and PFB scenarios over increasing number of nodes until reaching 125 nodes, while each node has a movement speed up to 15 meters per second, and the transmission rate is 10 packets per second.

Figure 9 graph depicts that PFB algorithm delivers more packets than others with the specified area over all numbers of nodes. In case of dense environments (Maximum = 25 nodes), PFB outperforms STD and PLEB by 00.05% and 00.02%, respectively. Instead, on the maximum number of nodes PFB Outperforms STD by 00.11% and PLEB by 00.06%.

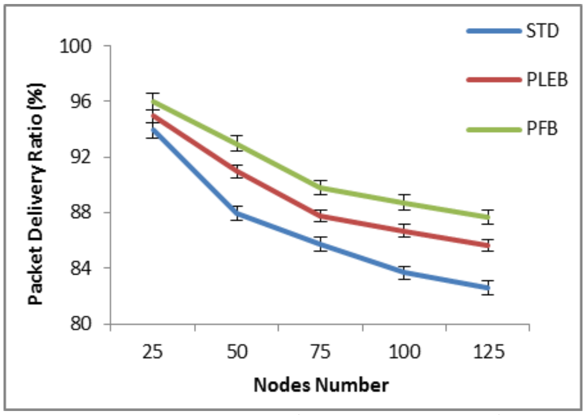


Figure 9: Packet delivery Ratio of STD, PLEB and PFB for Sending 10 packets/second, 0 second pause time, and 15 meters/second maximum speed.

Figure 10 shows packet delivery Ratio of the STD, PLEB and PFB scenarios over increasing number of nodes until reaching 125 nodes, 150 second pause time, while each node has a movement speed up to 15 meters per second, and the transmission rate is 10 packets per second.

This graph shows that the PFB algorithm delivers more packets than others with the specified area over all numbers of nodes. In the case of dense environments (Maximum = 25 nodes), PFB outperforms STD and PLEB by 00.02% and 00.05%, respectively. Oppositely, on the maximum number of nodes PFB Outperforms STD by 00.17% and PLEB by 00.11%.

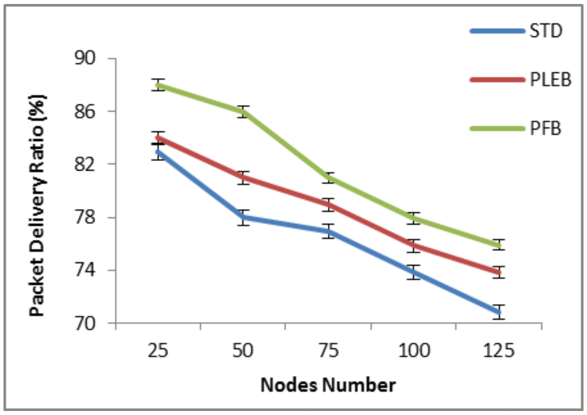


Figure 10: Packet delivery Ratio of STD, PLEB and PFB for Sending 10 packets/second, 150 second pause time, and 15 meters/second maximum speed.

VII. Conclusion and Future Work

This research proposes the PFB algorithm designed which is compared in contrast to the STD and PLEB algorithms. The results show that the new proposed algorithm outperforms the other compared ones, in terms of packet delivery fraction, end-to-end Delay and normalized routing load. Results also show the effect of node speed in all three algorithms.

When the nodes speed is high, the delivery ratio is lower than, when the motion experienced by the node when having slow movements.

Results show the network density variations in terms of the parameterized measurements, where sparse areas have the highest value compared to the dense ones. Results also show that for different approaches, as the node speed increases, the number of successfully delivered packet, increases.

Moreover, this study highlights the importance of developing a new Backoff algorithm that can dynamically adjust the Backoff waiting time, and takes into confederation the increased contention window size that has to be progressively reduced.

Another improving area includes investigating the effect of node's transmission ranges. The PFB algorithm proposed can be used with any Medium Access Control protocol, to achieve uniform distribution using Fibonacci series by reducing the increment factor for large contention window sizes.

The results extracted by conducting simulation experiments show that PFB algorithm achieved better performance than PLEB by a percentage that ranges between 0.01% up to 74% in the cases studied for different network densities and mobility states for the end-to-end delay, the packet delivery Ratio, and the normalized routing load.

In a number of limited cases the PFB algorithm performance was not the optimal, which means that in any of these cases the devices' performance cannot reach peak compared to other established algorithms.

In order to find the most optimal network performance we can test different parameters, increment behaviors and distinct environment conditions.

A future work can merge more Backoff algorithms with different distributions to examine different variations of the pessimistic Fibonacci Backoff algorithm, including more parameters for measuring and validating the efficiency. Finally one of our priorities in future aims of the current research is to evaluate the proposed scheme under real-time parameterizations and conditions using WSN and more specifically the MICA2Dot motes. This will enable in real time the evaluation and efficiency of the proposed scheme under real-time conditions and experimentation.

VIII. References

[1] S. Manaseer. "On Backoff Mechanisms for Wireless Mobile Ad Hoc Networks". Thesis Submitted By For The Degree of Doctor of Philosophy. The Faculty of Information and Mathematical Sciences. University of Glasgow, 2009, PP 1-156.

[2] M. BaniYassein, M. Khaoua, L.M. Mackenzie, S. Papanastasiou. "Performance Analysis of Adjusted Probabilistic Broadcasting in Mobile Ad Hoc Networks". International Journal of Wireless Information Networks, Springer Netherlands, Mar 2006, pp. 1-14.

[3] H. Wu, Y. Pan. " Medium Access Control in Wireless Networks (Wireless Networks and Mobile Computing)". Nova science publishers, 2008.

[4] K. Kalepu, S. Mehra, C. Yu. "Experiment and Evaluation of a Mobile Ad Hoc Network with AODV Routing protocol". MCRL (Mobile Computing Research Lab.), ECE Dept., Cleveland State University, Oct. 2002.

[5] B. Williams, T. Camp. "Comparison of Broadcasting Techniques for Mobile Ad Hoc Networks". ACM international symposium on Mobile ad hoc networking computing (2002), ACM Press, 2002, pp. 194-205.

[6] Y. Tseng, S. Ni, Y. Chen, J. Sheu. "The Broadcast Storm Problem in a Mobile Ad Hoc Network. Wireless Networks". IEEE, presented at Wireless Networks, 2002, pp. 153-167.

[7] M. Masadeh, S. Manaseer. "Pessimistic Backoff for Mobile Ad hoc Networks". IEEE, in proceeding of the 4th international conference on information technology ICIT, Jordan, 2009, pp. 1-10.

[8] J.Al-Karaki, A.Kamal. "Routing Techniques in Wireless Sensor Networks: A Survey". IEEE communications, Volume 11, No. 6, 2004, pp. 6--28.

[9] A. Escolà, J. Cantero. "Development of a wireless sensor network with 6LoWPAN support". Master thesis in Science and Telecommunication Engineering & Management, 2009.

[10] C. Liu, J. Kaiser. "A Survey of Mobile Ad Hoc network Routing Protocols". TR-4, MINEMA, University of Magdeburg, 2005.

[11] M. Kumar, T.V. Kumar, M. Hanumanthappa, E. Geetha. "Secure Mobile Based Voting System". Proceedings of 6th International Conference on E-Governance, IIT Delhi, 2008.

[12] S. Eichler, C. Roman. "Challenges of Secure Routing in MANETs: A Simulative Approach using AODV-SEC". In Proceedings of the 3rd IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS), Vancouver, Canada, 2006, pp. 481 – 484.

[13] S. Manaseer, M. Ould-Khaoua, L. M Mackenzie. "Fibonacci Backoff Algorithm for Mobile Ad Hoc Networks". DCS Technical Report series, Department of Computing Science, University of Glasgow, 2006, pp. 1-6.

[14] N. Song. "Enhancement of IEEE 802.11 distributed coordination function with exponential increase exponential decrease Backoff algorithm". IEEE 57th Semiannual Vehicular Technology Conference, Vol.4, USA, 2003; pp. 2775-2778.

[15] S. Manaseer, M. Ould-Khaoua, L. Mackenzie. "Fibonacci Increment Backoff Algorithm for MAC Protocol in Mobile Ad Hoc Networks". 7th Annual Postgraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting, PGNET, Liverpool, UK, 2006, pp. 103-109.

[16] NS-2 Simulator, at http://www.isi.edu/nsnam/ns/.