UNIVERSITY OF JYVÄSKYLÄ

ISAST Transactions on Computers and Intelligent Systems No.1 Vol.1, 2009

ISAST Transactions on No. 1, Vol. 1, 2009 (ISSN 1798-2448)

Computers and Intelligent Systems

Adaptive Object Tracking in Dynamic Environments with User Interaction

Gabriele Peters and Martin Kluger

***Abstract*—In this article an object tracking system is introduced which is capable to handle difficult situations in a dynamically changing environment. We evaluate the concepts of the proposed system by means of the task of person tracking in crowded seminar rooms. The contributions are threefold. We propose an expansion of the condensation algorithm which results in a more stable tracking in difficult situations such as sudden camera movements. Secondly, a new way of particle diffusion is introduced which allows for the adaption of the tracking module to movements of the camera as, e.g., induced by the user. These two contributions apply not only to person tracking but to object tracking in general. The third contribution, which applies to person tracking only, consists in a flexible way how to represent a moving person. The proposed tracking module meets demands such as real-time tracking of a target person who is allowed to move freely within a group of other persons and thus can be occluded. An interactive selection of new target persons is possible and also the challenge of a dynamically changing background could be coped with.**

***Index Terms*—person tracking, video surveillance, automatic video production, intelligent rooms**

I. Introduction

Tracking objects in a dynamically changing environment belongs to one of the most difficult problems in computer vision. Solutions of this problem are crucial not only for person tracking, which is the application on which we evaluate our proposed tracking method in this article, but also, for example, for dynamic object acquisition where sequences of objects viewed from different view points are analysed. The appearance of objects can change dramatically from frame to frame in a video sequence due to changes in the environment. Intelligent systems should be able to react dynamically to variations in the object's appearance. These variations can have many causes. If, for example, a camera moves around an object the lighting conditions can change or parts of the object can appear or disappear. In addition, the background can change which increases the degree of difficulty to assign corresponding landmarks from frame to frame, i.e., track the object. On the other hand, not only the camera can move but the object can move by itself or, even worse, can change its shape, which holds true for walking persons. The most difficult task is on hand, when also the background varies dynamically. Thus, robust object tracking in dynamic environments should be by itself adaptive to changes in the environment.

In this article we introduce a tracking system which reacts dynamically to recognized changes in the environment (i.e., changes in the camera parameters and interaction by the user). As a result the acquired data depend on the environmental dynamics. These concepts of active, visual object acquisition are applied to moving persons who interact in seminar rooms. The demands of such a person tracking system are manifold and some of them are listed in the following. The person tracking system should be capable of:

1) tracking a selected target person in real-time,
2) tracking a target person within other interacting persons and in front of a dynamic background
3) adapting to changing camera parameters such as orientation and focal length, where changes are induced either by an active camera or are initiated by a user,
4) bridging occlusions of the target object,
5) allowing for an interactive selection of a new target person during runtime.

The tracking system we present in this article is part of a larger project called *Virtual Camera Assistant*. The Virtual Camera Assistant is a prototype of a semi-automatic system, which supports documentary video recordings of indoor events such as seminars. It combines an automatic person tracking system with a controllable pan-tilt-zoom camera. Depending on the output of the tracking module the camera parameters (i.e., orientation and focal length) are determined automatically in such a way that the observer gets a natural impression of the recorded video (e.g., without jerky movements) and the whole movie appears pleasant to the eyes. In addition, (that is the ”semi” in ”semi-automatic”) the user of the Virtual Camera Assistant is able, on the one hand, to control the camera parameters manually as well and, on the other hand, to select the target person to be tracked interactively at any time during recording.

A distinctive feature in comparison to other systems for automatic video production is the possible scenario of person tracking in highly cluttered scenes and within groups of interacting persons.

In this article we will omit a detailed description of the camera control. It is covered in [1] where the whole system is described. Rather, we will concentrate here on the introduction of the tracking module, i.e., on the realization of an automatic tracking of a target person in a crowded room with an active camera. In section II we present the components of the tracking module, in section III we describe our experimental

Manuscript received May 13, 2009; revised August 02, 2009.

This research was funded by the German Research Association (DFG) under Grant PE 887/3-3.

Gabriele Peters is with the University of Applied Sciences and Arts, Computer Science, Visual Computing, Emil-Figge-Strae 42, D-44227 Dortmund, Germany, e-mail: gabriele.peters@fh-dortmund.de (see http://www.inf.fh-dortmund.de/personen/professoren/peters).

Martin Kluger was with the Technical University Dortmund, Computer Science, Computer Graphics, Otto-Hahn-Strae 16, D-44227 Dortmund, Germany.

data, and in section IV the results, including weaknesses and limitations, are summarized, after which we close in section V with the conclusions. But still let us familiarize the reader with systems related to our approach.

A. Related Systems

Direct comparisons of the proposed system are possible with approaches from automatic video production, approaches from intelligent rooms with active cameras, as well as approaches from person tracking in general, e.g., for surveillance.

1) Automatic Video Production: A completely automatic system for the recording of presentation events is proposed in [2]. It is able to track one or more persons on a stage based on two cameras. The first one is static and monitors the whole stage to detect movements. The second camera is flexible and its parameters are controlled by the first one, thus, it is automatically oriented towards events. A similar automatic video production system is presented in [3]. Its tracking component utilizes two cameras as well. Movement on a stage is detected via the difference of two successive frames. In addition, the valid region for movements of the speaker is strongly restricted to a small area at the podium. Summarizing, these systems concentrate on a composition of several video sources and restrict the scenario of person tracking much stronger than the system proposed in this article where also interactions between the target person and other persons are allowed.

2) Intelligent Rooms: Our system can also be compared to the person or head tracking modules of *intelligent* or *smart rooms* (see Fig. 1). However, smart rooms usually are equipped with a number of different cameras, such as static ones, 360 degree omnidirectional cameras [4], or special stereo cameras [5]. With static and omnidirectional cameras it is possible to remove static background of a scene with standard methods [6]. Then further analysis for person tracking can be confined to the foreground. The object position obtained in this way can be used to adjust the synchronized, active cameras with the purpose of recording frames of higher resolution from the best view.

In scenes with dynamic background and without additional information on the parameters of the active camera (as it applies to our system) models can be generated only with explicitly higher effort. In [7] the requirements are reduced by a restriction to pan and tilt movements of the camera only. In addition, during camera movements the segmentation results are improved by a template matching with the foreground recognized in the last step.

A different possibility is a static camera array which simulates a single, virtual, active camera on the overlapping fields of view of all cameras [8]. One advantage is a large static image space with a higher resolution compared with a wide angle camera, which can be segmented with standard background models.

Alternatively, one can dispense with a background model completely when active cameras are used. Instead, potential candidates for the next object position can be predicted model-based and verified afterwards. This is realized, e.g., in [9] and the references cited in subsection I-A3.

3) Person Tracking: As examples for the field of video surveillance [10], [11], [12] and many approaches to the problem of person tracking in general [13], [14], [15] we refer here to the works of Isard [16], Nummiaro et al. [17], and Perez et al. [18] only, as the person tracking of the proposed system is based on their ideas (see also Fig. 2).

By means of several examples of object tracking, mostly based on edge filters, Isard introduces the condensation algorithm. It is a simple but effective particle filter algorithm which works stable also under temporary occlusions and disturbances. Our system is based on this algorithm. Both the other approaches use particle filters as well, but utilize color-based histograms as object features. Whereas in [17] a color histogram is adapted all along the video sequence to compensate for illumination changes, two static color histograms are used in [18] to improve the object hypothesis. Both concepts are applied in our approach as well.

II. Components of the Object Tracking System

In this section we introduce the methods and functioning of the proposed tracking system. In subsection II-A we first give an overview of the object tracking as part of the Virtual Camera Assistant. In addition, we recall the condensation algorithm our system is based on. In subsection II-B we describe our improvement of the condensation algorithm which consists mainly in the reiteration of parts of the original algorithm. Subsections II-C to II-G introduce concepts necessary to understand the object tracking module, such as the definition of the object state, the dynamic model, two different object profiles, the used distance functions, and the adaption of the reference profile of the target object (i.e., the object representation learned so far) to the current measurements. In the last subsection we describe our approach, how the particle diffusion in the condensation algorithm can be dynamically adapted to events in the environment, clarified by means of the example of camera movements.

A. Overview

The Virtual Camera Assistant mentioned in the introduction consists of the modules *Object Tracking* and *Camera Control* (see Fig. 3). The first module extracts the position and size of the currently selected target object from the video stream in real-time. The information obtained by this process is used by the second module to realize the desired camera adjustment. Short-term occlusions of the target object and other disturbances of the footage are mostly recognized and incorporated by the algorithms of both components and bypassed if possible.

Here we will concentrate on the *Object Tracking* module. For the segmentation between foreground (which is the target object) and background one can follow a bottom-up or a top-down approach. In the bottom-up approach regions are constructed, starting from the image pixels, and assigned to foreground or background. In contrast to this, the top-down approach utilizes a priori knowledge on the object, e.g., in the form of an object model, generates hypotheses and verifies them in the image. This is the way we proceed. Fig. 4

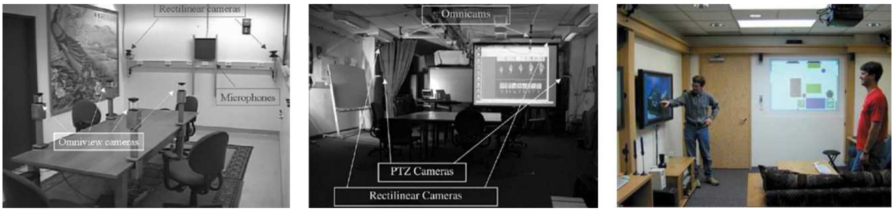


Fig. 1. Examples for smart rooms. Left and middle image taken from [4], right image taken from [5].

Fig. 2. Examples for person tracking with particle filter and different features. Left image taken from [16], middle image taken from [17], right image taken from [18].

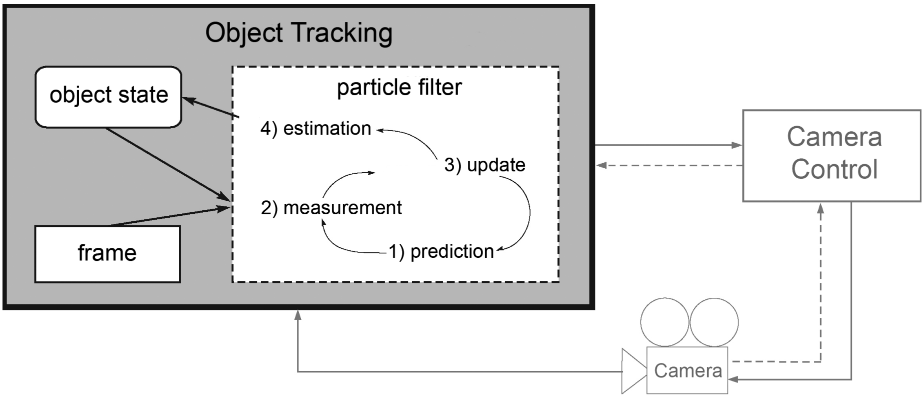


Fig. 4. Object Tracking Module. Object tracking is realized via a particle filter and operates in the four steps depicted here.

gives an idea of the information processing by the tracking module. Object tracking is realized via a particle filter which, roughly speaking, propagates multiple weighted hypotheses (called *particles*) and operates in four steps. First, for each particle the *object state* (see subsection II-C) - here it is the object's position and size in the current frame - is predicted by means of a *dynamic model* (see subsection II-D). Secondly, for each particle values of defined features are measured in the current frame. In the *object profile* (see subsection II-E) we define which features are observed. Thirdly, the hypotheses are compared with a reference profile via a similarity function (see subsections II-E and II-F). The results of the evaluation are used for an update of the particle weights. Fourthly, from all of those newly weighted hypotheses the pose and size estimation of the tracked object is calculated. After this, the reference profile is adapted via a new measurement in the current frame at the estimated position (not shown in this figure) (see subsection II-G). The particle filter is realized by an expansion of the condensation algorithm (see subsection II-B).

In a nutshell, after a manual selection of the target object by the user of the system, a number of features is initialized on the basis of which the hypotheses can be verified. From the current state of the object (e.g., her position and size) at time $t-1$ (i.e., in the current frame) a constant number of hypotheses for the state at time t (i.e., in the next frame) is calculated in a dynamic model. The validity of each hypothesis (i.e., each particle) is assessed in the following way. The features are measured for each particle in the frame at time t. Then they are compared with a reference profile (i.e., the description of the object learned up to this time) resulting in probability values,

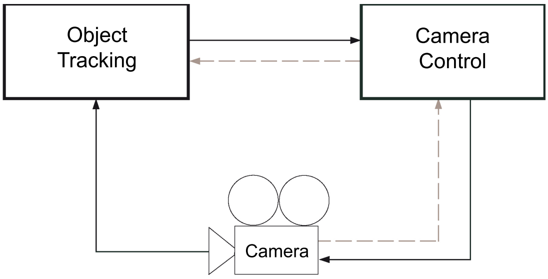


Fig. 3. Virtual Camera Assistant. The *Object Tracking* module recognizes a target object in a video frame, the *Camera Control* module adjusts the camera parameters in such a way that the object is shown in a favoured part of the frame.

which describe the validity of the hypotheses. (Particle weights are initialized by a uniform distriution.) Thus, the set of all hypotheses can be regarded as a discret approximation of the probability distribution for the current state of the target object. The estimated expectation value of this probability distribution, i.e., the weighted mean of all hypotheses, provides the new state of the target object at time t. Finally, the reference profile of the object is adapted by a second measurement at the estimated position of the object in frame t.

1) Condensation Algorithm: More formally spoken, the aim is the approximation of the *filtering distribution* $p(x_t|\mathcal{Z}_t)$ of the internal state x_t at time t on the basis of all previous measurements $\mathcal{Z}_t$. The hypotheses are represented by a constant number N of points $x_t^{(i)}$, $i = 1, \dots, N$, at time t to which probabilities in form of normalized weights $\widetilde{w}_t^{(i)}$ are assigned. The points are called *particles*. The discrete approximation of $p(x_t|\mathcal{Z}_t)$ at time t is given by

$$\widehat{P}(x_t|\mathcal{Z}_t) \approx \sum_{i=1}^{N} \widetilde{w}_t^{(i)} \delta(x_t - x_t^{(i)})$$

with $\delta(.)$ as Dirac impulse. At the transition from $t-1$ to t the particles (and thus the hypotheses) are distributed anew via propagation through the dynamic model. Afterwards they are reassessed by a new measurement. That means, the new particles and weights $(x_t^{(i)}, \widetilde{w}_t^{(i)})$ are calculated from the old particles and weights $(x_{t-1}^{(i)}, \widetilde{w}_{t-1}^{(i)})$. This calculation is described by the condensation algorithm, as proposed by Isard [16], which consist of three steps:

1) *Resampling:* A new sample $\{\widetilde{x}_{t-1}^{(i)}, i = 1, \dots, N\}$ of size N is drawn with replacement from the set of particles $\{x_{t-1}^{(i)}, i = 1, \dots, N\}$ at time $t-1$. The weight of a particle defines its probability to be drawn.
2) *Prediction:* Based on this new sample a prediction is carried out via the dynamic model (described in subsection II-D). Thus, the particles become representatives of a sample $\{x_t^{(i)}, i = 1, \dots, N\}$.
3) *Measurement:* The new particle weights $\widetilde{w}_t^{(i)}$ are calculated as follows. The consistency of the N hypotheses (represented by the N particles) with the current frame is evaluated by means of the similarity function given in formula 1. Finally, these weightings $w_t^{(i)}$ are normalized (resulting in $\widetilde{w}_t^{(i)}$) and assigned to their associated particles.

The weighting $w_t^{(i)}$ for one particle $x_t^{(i)}$ is calculated as follows:

$$w_t^{(i)} = \frac{1}{\sqrt{2\pi\sigma}} \cdot exp\left(-\frac{1}{2\sigma^2} \cdot \left[D_{\widetilde{\mathcal{R}}}\left(x_t^{(i)}\right)\right]^2\right), i = 1, \dots, N. \quad (1)$$

It can be regarded as *confidence value* for the corresponding measurement. How the distance value $D_{\widetilde{\mathcal{R}}}\left(x_t^{(i)}\right)$ regarding the reference profile $\widetilde{\mathcal{R}}$ is calculated is defined in subsection II-E. The Gaussian shape of formula 1 secures that with a growing match of the hypothesis with the measured reference data the assigned weight increases as well. Throughout all experiments we used a value of 0.13 for σ.

From the new particles an estimation of the current position and size of the object is possible, for example, simply by averaging. We estimate the new object state $\hat{x}_t$ by a weighted sum of the particles:

$$\hat{x}_t = \sum_{i=1}^{N} \widetilde{w}_t^{(i)} * x_t^{(i)} \quad (2)$$

The first step (resampling) has the purpose to prevent the particle weights from becoming degenerated [19]. By drawing particles with replacement light weighted particles tend to be excluded from further calculations.

B. Expansion of the Condensation Algorithm

The parameters of the dynamic model (described in subsection II-D) have been determined on the basis of a ”regular”, not overly fast movement of the target person in the context of a seminar scene. In practice, however, it is possible, that the object movement - offset with the camera movement - is not always covered by the model. In these particular cases too few of the propagated hypotheses still cover the ”true” position of the object, normally resulting in a tracking loss. An example for this is shown in the first row of Fig. 5.

For this reason we improve the condensation algorithm by two measures.

- We use a first threshold α_1 for the confidence value of the currently estimated object state $\hat{x}_t$. If this confidence value falls below α_1, a second iteration of steps (1) to (3) of the condensation algorithm is carried out on the same frame with $N_2 \geq N$ particles.
- In case the newly estimated object state in turn provides a confidence value, which lies below a second threshold $\alpha_2 < \alpha_1$, we increase the number of particles to $N_{max} \gg N$ in the resampling step from the next frame on. As soon as the confidence value exceeds threshold α_1 in the further processing, the number of particles is reduced to N again.

The first measure results in a noticeable improvement of the tracking. The second resampling is based on the weights of the first run, and thus the strongest weighted particles (though not strongly weighted) can direct the second run into the right direction, while the object remains in place. This approach

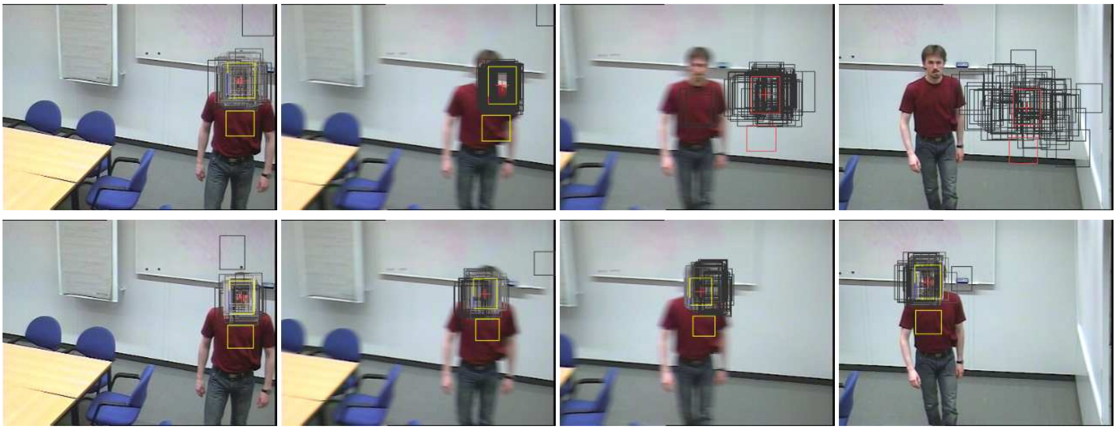


Fig. 5. Condensation algorithm with reiteration. For a description see subsection II-B.

resembles the annealed particle filter proposed by [20], who apply several iterations for one point in time with different weighting functions which mature over the course of iterations. This filter proved itself on person tracking under controlled conditions such as a neutral background.

To evaluate the proposed expansion we used N_2 between 10% and 20% and N_{max} between 50% and 75% larger than N. For α_1 and α_2 the values 0.4 and 0.2 proved to be feasible, respectively.

The first row of Fig. 5 shows frames 15, 20, 23, and 28 of a sequence as an example for a tracking loss due to a sudden camera pan by the user. It results from an application of the condensation algorithm with $N = 75$ particles as described by Isard [16]. The gray rectangles mark the particle distribution, the yellow (or red, depending on the success of the tracking) ones tag the mean of them and thus the estimated position and size of the tracked author. (The reason why there are always two yellow, or red respectively, rectangles becomes clear after reading subsection II-E.) The second row shows the same frames, this time tagged with the results of an application of our expansion of the condensation algorithm with reiteration. The second iteration of steps (1) to (3) of the condensation algorithm with $N = 50$, $N_2 = 75$, and $\alpha_1 = 0.4$ makes for a successful tracking.

C. Object State

As already mentioned above the current state of the object is described by features of the object valid for the current frame, such as its position and size in the frame under consideration. The object state is also the internal state $x_t \in \mathbb{R}^n$ of the system for a discrete point in time $t \in \mathbb{N}_0$ which is propagated in the dynamic model (see subsection II-D). (We omit the time index t in this subsection for simplification.) As we will see in subsection II-E we will compare different object profiles, which differ in the number and choice of the features constituting them. For the sake of a uniform formalism we include all possible features in the definition of the object state and later evaluate only those subsets of them which suit the object profile considered. The complete *object state* is thus defined by

$$x := (p_x, p_y, \dot{p}_x, \dot{p}_y, a, r, \rho)^T \tag{3}$$

with

- $(p_x, p_y)^T$ the position of the target object (i.e., the center of the rectangle selected by the user to mark the person)
- $(\dot{p}_x, \dot{p}_y)^T$ the horizontal and vertical velocity of the target object, respectively
- a width of the target object (in our application the head or face of a person is selected by the user)
- r ratio between width and height of the target object (i.e., $r = height/width$)
- ρ flexibility between two coupled regions of interest (for object profil $\mathcal{P}_2$, see subsection II-E)

With the exception of the last two parameters, units are given as pixels or pixels per time interval for velocities, respectively. The relation between width and height of the rectangle is restricted by a factor between 1.2 and 1.8 to avoid extreme shapes of the rectangle: $r \in [1.2; 1.8]$.

D. Dynamic Model

The dynamics of the system are defined by constant velocities only. The size of the target object is neglected in the dynamic model, because size variations caused by movements of a person in our application are quite small and thus are compensated for by the stochastic term of the model. Similar in design to the approach described in [17] we chose a simple model of first order:

$$x_t = \mathbf{A}x_{t-1} + \phi_t \tag{4}$$

Matrix $\mathbf{A}$ represents the deterministic part of the model, the independent random variable ϕ_t the stochastic part. For the object state x_t, which was specified in the last subsection, $\mathbf{A}$ is defined by

$$\mathbf{A} := \begin{pmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{E_3} \end{pmatrix}, \qquad \mathbf{V} := \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{5}$$

with E_3 the identity of size 3×3.

The elements of the stochastic term ϕ_t are independent and normally distributed with expected values zero, but with different standard deviations. Thus, the extent of the variability of even the static elements of the object state (such as the width a) can be controlled. This enables the user to manipulate the adaptivity of the elements indirectly, without the need of additional variables for periodic changes. The following values have proven to be useful:

$$\Sigma_\phi := \left(\sigma_{p_x}, \sigma_{p_y}, \sigma_{\dot{p}_x}, \sigma_{\dot{p}_y}, \sigma_a, \sigma_r, \sigma_\rho\right)^T =$$

$$\left(3, 2, 2, 1, 1.5, 0.2, 9 \cdot \frac{2\pi}{360}\right)^T \tag{6}$$

1) Start Parameter: The start parameters for the object state depend on the position $(p_x^{t'}, p_y^{t'})$ and the selected object size $(a^{t'}, r^{t'})$ at time t' of the selection of a target object. As the target object is chosen by manual selection and interactively during the tracking of another object, one has to take into account that the target may have moved on already. For this reason the initial distribution of the particles $\{x_0^{(i)}, i = 1 \ldots N\}$ is modeled as normally distributed random variable with a larger standard deviation Σ_0 and with the manually selected values for position and size as expected values: $x_0^{(i)} \in \mathcal{N}(\Theta_0, \Sigma_0)$,

$$\Theta_0 = \left(\mu_{p_x}, \mu_{p_y}, \mu_{\dot{p}_x}, \mu_{\dot{p}_y}, \mu_a, \mu_r, \mu_\rho\right)^T =$$

$$\left(p_x^{t'}, p_y^{t'}, 0, 0, a^{t'}, r^{t'}, 0\right)^T \tag{7}$$

with $\Sigma_0 := \xi \cdot \Sigma_\phi, \xi > 1$. We chose $\xi = 5$.

E. Object Profiles

In this subsection we introduce two different so-called *object profiles* $\mathcal{P}_1$ and $\mathcal{P}_2$ we used for our tracking system. The object profiles define which features $\mathcal{F}(x)$ describe an object state x. Due to these features a person is recognized and tracked from frame to frame. Furthermore, for each object profile we define a distance function $D_{\widetilde{\mathcal{R}}}(x) := D\left(\widetilde{\mathcal{R}}, \mathcal{F}(x)\right)$ which determines the similarity (or better *dissimilarity*) between a *reference profile* $\widetilde{\mathcal{R}} := \mathcal{F}(\hat{x}_0)$ characterizing the target object and the features $\mathcal{F}(x)$ of a current hypothesis x. (The result of the distance function $D_{\widetilde{\mathcal{R}}}$ is finally converted into the particle weight of the considered hypothesis.)

We use color histograms as features. Let $H(p_x, p_y, a, a \cdot r)$ be a color histogram of a rectangular area with center (p_x, p_y), width a, and height $a \cdot r$. Furthermore, let $\Delta(H_1, H_2)$ be a distance function between two histograms H_1 and H_2. (It is defined in formula (15).) The two used object profiles $\mathcal{P}_1$ and $\mathcal{P}_2$ are illustrated in Fig. 6 and are defined now as follows.

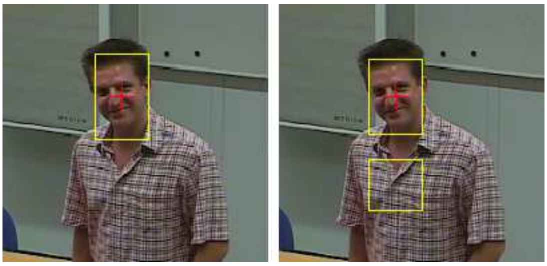


Fig. 6. Two different types of object profiles. Left: profile $\mathcal{P}_1$, right: profile $\mathcal{P}_2$.

1) Profile $\mathcal{P}_1$: Restricted Ratio of Rectangle Width and Height: This profile consists of one histogram H_T only. If the target object is a person H_T should cover the head of the person (see left image of Fig. 6). To prevent the rectangle from degeneration we control the ratio of width and height by restricting r of the state vector $x := (p_x, p_y, \dot{p}_x, \dot{p}_y, a, r, \rho)^T$ as described in subsection II-C ($r \in [1.2; 1.8]$). Features $\mathcal{F}$ and reference profile $\widetilde{\mathcal{R}}$ are then defined as follows:

$$\begin{aligned} \mathcal{F}(x) &:= H_T(x) = H(p_x, p_y, a, a \cdot r) \\ \widetilde{\mathcal{R}} &:= H_T(\hat{x}_0) \end{aligned} \tag{8}$$

For the calculation of the similarity of a hypothesis $x := x_t^{(i)}$ with the reference profile $\widetilde{\mathcal{R}}$ the histogram distance function Δ can be applied directly:

$$D_{\widetilde{\mathcal{R}}}(x) := \Delta\left(H_T(x), \widetilde{\mathcal{R}}\right) \tag{9}$$

2) Profile $\mathcal{P}_2$: Two Rectangles: This object profile is customized especially for the application of the tracking system to moving persons. As a person's front view does not only contain skin colors in the face area, but also, e.g., at the hands, these areas can easily be confused with the face if only one color histogram is evaluated. To prevent confusion we integrate a second rectangular area in profile $\mathcal{P}_2$ the histogram H_B of which is evaluated. This approach is similar to that described in [18]:

$$\begin{aligned} \mathcal{F}(x) &:= \begin{pmatrix} H_T(x) \\ H_B(x) \end{pmatrix} \\ \widetilde{\mathcal{R}} &:= \begin{pmatrix} \widetilde{\mathcal{R}}_T(x) \\ \widetilde{\mathcal{R}}_B(x) \end{pmatrix} \end{aligned} \tag{10}$$

H_B has a square shape with side length a and is (as in [18]) positioned below the region of the first histogram H_T (see right image of Fig. 6). But in contrast to the approach described in [18] we employ a dynamic positioning of the second rectangle on a circular arc with center (p_x, p_y) and radius $b = 1.7a$. These circumstances are depcited in Fig. 7. The extent of the circular arc is determined by the parameter $\rho \in [-\rho_{max}; \rho_{max}]$ which is the last component of the object state vector (see formula (3)). We obtained the best results with $\rho_{max} = 20 \cdot \frac{2\pi}{360}$. This flexible positioning of the second rectangle should improve the tracking in situations where the person displays a laterally crooked posture or when partial

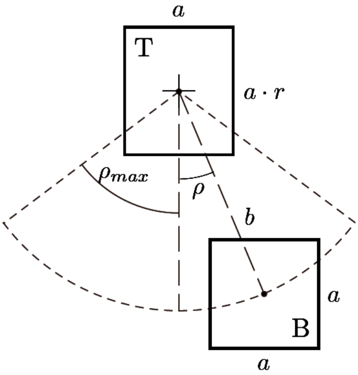


Fig. 7. Object profile $\mathcal{P}_2$. The second rectangle is positioned dynamically on a circular arc.

occlusions below the upper rectangle occur. The histograms are calculated as follows:

$$\begin{aligned} H_T\,(x) &:= H\,(p_x, p_y, a, a\cdot r) \\ H_B\,(x) &:= H\,(p_x + 1.7a\cdot sin(\rho), p_y + \\ &\quad\; +1.7a\cdot cos(\rho), a, a) \end{aligned} \tag{11}$$

For the evaluation of the similarity of a hypothesis x a weighted mean of both histogram distances $D_{\widetilde{\mathcal{R}}_T}(x)$ and $D_{\widetilde{\mathcal{R}}_B}(x)$ is calculated:

$$\begin{aligned} D_{\widetilde{\mathcal{R}}}\,(x) &:= \gamma^+\,(x)\cdot\left[0.6\cdot D_{\widetilde{\mathcal{R}}_T}(x) + 0.4\cdot D_{\widetilde{\mathcal{R}}_B}(x)\right] \\ D_{\widetilde{\mathcal{R}}_T}\,(x) &:= \gamma^-\left(\Delta\left(H_T(x), \widetilde{\mathcal{R}}_T\right)\right) \\ D_{\widetilde{\mathcal{R}}_B}\,(x) &:= \gamma^-\left(\Delta\left(H_B(x), \widetilde{\mathcal{R}}_B\right)\right) \end{aligned} \tag{12}$$

The face region is weighted stronger whereas the rectangle in the bottom can be regarded as an auxiliary region. In addition to this, increases and decreases of the different distance values are carried out depending on heuristic measures governed by the following two rules:

1) Decrease of both single histogram distances by γ^- in case they already fall below a threshold:

$$\gamma^-(s) \quad := \quad \begin{cases} 0.9s & s < 0.3 \\ s & otherwise \end{cases} \tag{13}$$

2) Increase of the total distance by γ^+ in case both single histogram distances still exceed a threshold after the application of rule (1):

$$\gamma^+(x) := \begin{cases} 1.1 & D_{\widetilde{\mathcal{R}}_T}(x) \geq 0.4\; and\; D_{\widetilde{\mathcal{R}}_B}(x) \geq 0.4 \\ 1 & otherwise \end{cases} \tag{14}$$

These rules reinforce a positive as well as a negative bias of the distance values based on experimentally derived thresholds.

F. Histogram Distance

What remains now is the definition of the distance $\Delta(H_1, H_2)$ between two histograms H_1 and H_2. The structure of the histograms we use is based on those utilized in [17] and [18] and is characterized by the color space and the number of the bins. To achieve a larger robustness of the acquired color information against lightness variations we use the HSL color space and partition it into 32 bins. 24 of them are reserved for hue and saturation ($H \times S$), 8 for lightness (L). However, the lightness information is utilized only for those pixels that do not provide reliable color information. We threshold the minimal saturation with $\chi = 16$ out of a maximum of 255, because below this threshold the signal starts to become noisy. The assignment of a pixel to a bin of the histogram is described by the function $\mathcal{B} : H \times S \times L \longrightarrow \{1, \ldots, 32\}$:

$$\mathcal{B}(u_h, u_s, u_l) \quad := \quad \begin{cases} \mathcal{B}_L(u_l) & 0 \leq u_s < \chi \\ \mathcal{B}_{HS}(u_h, u_s) & \chi \leq u_s \leq 255 \end{cases}$$

The case differentiation represents the separation of the bin partitioning and $\mathcal{B}_L$ and $\mathcal{B}_{HS}$ are the corresponding lookup tables. We employ a uniform allocation of the pixels to the bins. In Fig. 8 the described calculation of the histograms is depicted.

Summarizing, we have the following formal description of a normalized histogram H on the rectangle image section denoted by $\Box$ with center (p_x, p_y) and size $a \times a \cdot r$:

$$H(p_x, p_y, a, a\cdot r) \quad := \quad \frac{1}{a\cdot ar}\cdot\begin{pmatrix} h(1,\Box) \\ \vdots \\ h(32,\Box) \end{pmatrix}$$

with $h(i, \Box)$ as function which counts the number of pixels in $\Box$ for bin i:

$$h(i,\Box) \quad := \quad \sum_{u\in\Box} \delta\left(\mathcal{B}(u) - i\right), \qquad i \in \{1, \cdots, 32\}.$$

The normalization by the factor $1/(a \cdot ar)$ allows for the comparison of histograms derived from base areas of different sizes. For the purpose of an efficient calculation of the histograms we enhanced the concept of *integral histograms* ([21], [22]) for particle filters. This is described in detail in [1].

Finally we are able to define the distance between two histograms H_1 and H_2 as follows:

$$\Delta(H_1, H_2) \;:=\; \sqrt{1 - \sum_i \sqrt{H_{1,i}\cdot H_{2,i}}}\;, i \in \{1, \cdots, 32\}. \tag{15}$$

where $H_{1,i}$ and $H_{2,i}$ are the i-th bins in the corresponding histogram vector. Δ is the Bhattacharyya distance, which is used in [17] and [18] as well.

G. Adaption of the Reference Profile

For a static, color-based reference profile $\widetilde{\mathcal{R}}$ as introduced in subsection II-E, which is calculated only once, color variations due to changes in lightning or viewpoint pose a serious problem. For this reason we modeled the reference profile, similar to the approach described in [17], dynamically. In case the particle weighting w_t for the currently estimated object state $\hat{x}_t$ (see formulas (1) and (2)) exceeds a minimum value

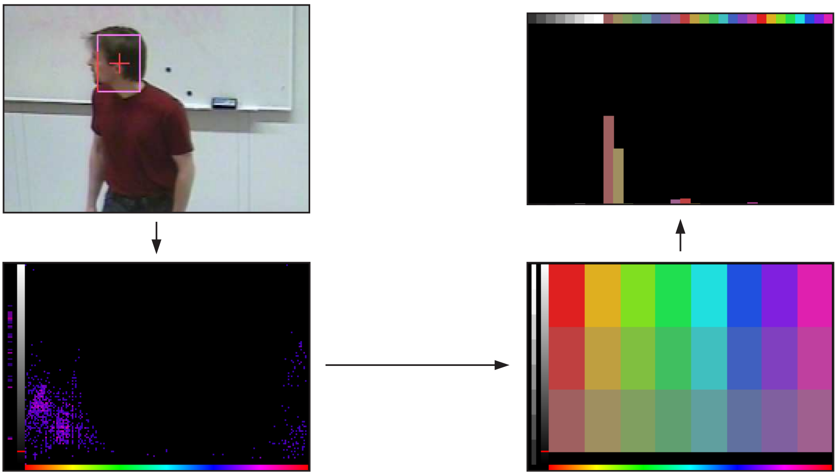


Fig. 8. Calculation of the histograms. The image in the bottom left shows the detailed $(H \times S, L)$ histogram coding the pixel occurrences of the rectangle displayed in the upper left image. In the bottom right image the quantization into $8 \cdot 3$ hue and saturation bins and 8 lightness bins (left hand side of this image) is depicted. This results in the used final histogram form displayed in the upper right image.

w_{min} $(w_t \geq w_{min})$ the data of the estimated object state are integrated into the current reference profil:

$$\begin{aligned} \widetilde{\mathcal{R}}_0 &:= \mathcal{F}(\hat{x}_0) \\ \widetilde{\mathcal{R}}_t &:= \kappa\widetilde{\mathcal{R}}_0 + (1-\kappa)\left[\lambda\widetilde{\mathcal{R}}_{t-1} + (1-\lambda)\mathcal{F}(\hat{x}_t)\right] \end{aligned}$$

$\widetilde{\mathcal{R}}$ is updated componentwise, i.e., for object profile $\mathcal{P}_2$ both histograms are refreshed separately. In contrast to the approach described in [17] the initial object profil $\widetilde{\mathcal{R}}_0$ remains in the current profile with a fixed fraction κ. This prevents the color histograms from beeing totally changed. We chose the following values of the parameters: $w_{min} = 0.25, \kappa = 0.15$, and $\lambda = 0.15$.

H. Competing Reference Profiles

To prevent the reference profile $\widetilde{\mathcal{R}}$ from beeing adapted to wrongly tracked image parts the parameters mentioned in subsection II-G should be chosen carefully. Nevertheless, it is necessary to incorporate different color distributions for one target object as exemplified in Fig. 9. For this reason we utilize, similar to the approach described in [9], several competing reference profiles $\widetilde{\mathcal{R}}^{(j)}$. The similarity of each hypothesis $x_t^{(i)}$ is then calculated as maximum of the similarities for each of the competing reference profiles:

$$D_{\widetilde{\mathcal{R}}}\left(x_t^{(i)}\right) := \max_j \left\{ D_{\widetilde{\mathcal{R}}^{(j)}}\left(x_t^{(i)}\right) \right\}$$

For best results the competing reference profiles should describe clearly distinct representations of the target object, such as front and side views. The competing profiles have been arranged in a circular buffer which means that a new initialization of the profile by the user overwrites an older reference profile in case all positions in the circular buffer are allocated. In our experiments we provided two competing reference profiles for each target object. For all different parameter configurations we analyzed in the experiments the user has initialized for each sequence both reference profiles always in the same frame. An example of a target object successfully tracked because of competing reference profiles is shown in Fig. 15.

I. Adaptive Particle Diffusion

As mentioned in the overview (subsection II-A) the complete Virtual Camera Assistent consists of the two modules *Object Tracking* and *Camera Control*. They interact with each other inter alia by an adaption of the tracking module to the movements of the cameras. In doing so it does not matter whether the camera movements have been induced by the autonomous feedback loop between both modules as indicated in Fig. 3 or the camera - as part of the environment - has been moved by the user. In any case a dynamic adaption of the particle diffusion to the recognized camera movement takes place. This means that the target object is acquired depending on the recognized state of the environment.

As each camera pan causes a shift of the current target object in the image in the opposite direction, the basic idea of the camera movement-controlled particle diffusion consists in a coupling of the particle distribution to the movement of the camera in such a way that the particles are placed smarter before they are weighted by formula 1. The velocity and direction of the camera movement should affect the random particle distribution in the dynamic model. For this purpose we decide for each particle with probability ψ if, instead of ϕ_t (see formula (4)), we use a different distribution ϕ_t^* for the random part of the dynamic model (as in subsection II-D we again omit index i, which indicates a single particle):

$$x_t = \mathbf{A}x_{t-1} + \phi_t^* \qquad (16)$$

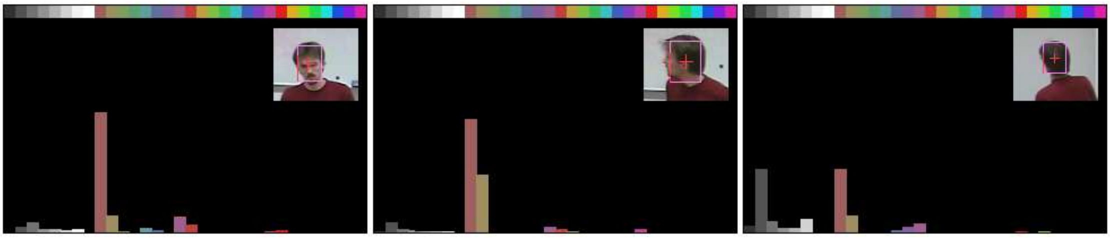


Fig. 9. Necessity for competing reference profiles. Different views of a target person imply different color distributions in the histograms.

As the camera parameters can be separated into horizontal (pan) and vertical (tilt) movement and the variation of the focal length, for each of these parameters an element of the object state vector can be modified. According to this, the random vector ϕ_t^* differs from ϕ_t only in those three entries which describe the random part of the position (p_x, p_y) and width a of the target object:

$$\phi_t^* := \begin{pmatrix} -s_x \cdot |\phi_{p_x}| \\ -s_x \cdot |\phi_{p_x}| \\ \bullet \\ \bullet \\ \phi_a \\ \bullet \\ \bullet \end{pmatrix} \quad with \quad \begin{matrix} \phi_{p_x} \in \mathcal{N}(0, \sigma_{p_x}), \\ \phi_{p_y} \in \mathcal{N}(0, \sigma_{p_y}), \\ \phi_a \in \mathcal{N}(\mu_a, \sigma_a). \end{matrix} \tag{17}$$

The used values for $\sigma_{p_x}, \sigma_{p_y}, \mu_a$, and σ_a are listed at the end of this section. $s_x, s_y \in [-1, 1]$ refer to the current pan and tilt direction, respectively. A change of the focal length usually causes only small changes in the projected width of the target object. This is the reason why changes of a are caused only by noise. A slight shift of the expected value or a slight boost of the standard deviation yields the desired effect. In contrast to this, the shift of the position of the object is quite distintive in terms of the horizontal and vertical movements so that only a wide particle distribution can cover the shift sufficiently. But on the other hand, a strongly shifted expectation value leads to an insufficient covering of the object at the old position, whereas a large standard deviation implies a wider positioning also in the opposite direction. For this reason the absolute values of the random variables in formula (17) ensure the necessary, unambiguous bias in the distribution of the particles. The negative sign secures the desired opposite direction of the distribution. Summarizing, the additive noise of the dynamic model does not display an expectation value of zero anymore, rather it takes a componentwise shift of the mean value into the opposite direction of the camera movement into account.

In Fig. 10 two examples for the adaptive particle diffusion are shown. The combination of the first and second image illustrates one example, the combination of the third and fourth image a second example. First example: The left image is a snapshot from a sequence with a camera pan to the left. The right image displays the final frame of this sequence. To illustrate the effect of the modified distribution of the particles by the adaptive particle diffusion we used a large value for the standard deviation of the horizontal movement, namely $\sigma_{p_x} = 50$. As probability ψ for the choice of this modified distribution we set $\psi = 0.3$ throughout all experiments. The modified particles are marked by a beige colored top bar. (Thus about 30% of the present particles are marked.) In the left image one can recognize that during a camera pan to the left the modified particles have a stronger bias to diffuse to the opposite direction. The unmodified particles realize the movement only after their modification by the measurements as can be seen in the right image. Second example: Here the same holds true only for the opposite direction of the camera pan to the right.

We employ a value of $\psi = 0.3$ for the probability of a modified distribution. Table 1 summarizes the used values of the parameters of formula (17). To allow for a more flexible differentiation of camera velocities we separate the camera parameters (pan, tilt, and focal length) into three velocity categories, namely slow, medium, and fast velocities. For each of these categories we define separate parameters for the underlying distribution. They are summarized in table 1. The division of the velocities into slow, medium, and fast depends on the adjustments the used camera (here camera JVC TK-C655) allows for. For pan and tilt the camera provides 8 levels, for focal length 3. These levels are quantized heuristically into the three categories. Details are described in [1].

parameter	slow velocity	medium velocity	fast velocity
σ_{p_x}	10.0	20.0	50.0
σ_{p_y}	8.0	15.0	30.0
μ_a	0.5	1.0	1.5
σ_a	1.0	2.0	3.0

Tab. 1. Parameter values for different velocities of camera movements.

III. Experiments

We analysed statistically different parameter configurations of the proposed tracking system. In particular, we examined the gain of the expansion of the condensation algorithm on the one hand and the account of the enhanced object profil $\mathcal{P}_2$ on the other hand. These issues represent two of the main contributions of this paper. The results of these experiments are reported in subsections IV-A and IV-B. Furthermore, we carried out several tracking experiments which reflect the

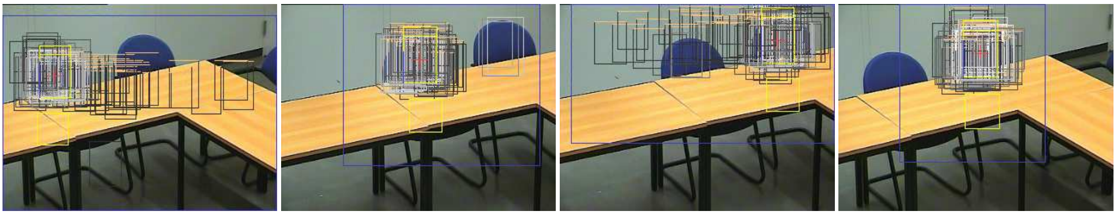


Fig. 10. Adaptive particle diffusion. For an explanation see subsection II-I.

capacities of the person tracking system according to the demands pronounced in the introduction. In the remaining subsections of section IV the results of those experiments are reproduced. For example, the adaption to changing camera parameters (third demand) is addressed in the examples of subsections IV-D and IV-E, results for sequences with occluded target persons are given in subsection IV-F (fourth demand), and the flexibility of the system in terms of user interaction (fifth demand) is the subject of subsection IV-G.

The experiments have been carried out on 3 GHz machines with image resolutions of 352×288 pixels. Although we used 200 particles only, this ensured tracking with even 1500 particles at a frame rate of 25 fps without difficulty.

A. Test Sequences

We carried out our experiments on three sequences with interacting persons in a seminar scene. The target person is selected interactively by the user and tracked autonomously according to the methods described in the last section. The test sequences (TS) are characterized as follows with increasing degree of complexity:

(TS1) Mainly one slightly moving person only at a time, relatively static camera, (typical seminar scene with moving instructor and relatively static audience)

(TS2) Mainly one moving person only at a time, but additionally a zooming and panning camera following the person's movements (and thus a dynamically changing background)

(TS3) Many simultaneously but non-uniformly moving persons with manycrossing roads and mutual oc-
clusions; in addition, we have interaction by the user,who selects different target persons

A subsequence of TS1 is shown in Fig. 14, subsequences of TS2 are displayed in Fig. 15, Fig. 16, and Fig. 17, and Fig. 18, Fig. 19, and Fig. 20 show subsequences of TS3.

B. Error Calculation

Besides the visual analysis of the results from test sequences we examine differences between single applied methods and their variations to evaluate the quality of our tracking approach. As the particle filter is a randomized technique the comparisons reported in the next section are based on 100 independent runs with identical adjustments and predefined start parameters for the target person. For each frame the position and size of the head of a tracked person was manually selected to define the ground truth for the experiments. Deviations from these values by the autonomously tracked positions and sizes are for each frame t of a sequence expressed by a *frame error* ϵ_t and summarized to a *totel error* ϵ of the whole sequence. For frames in which the visible part of the head of the target person is smaller than one third of its real size we evaluate a positive recognition with a maximal error $\epsilon_{max} := 100$. In addition, we penalize a loss of the target person with ϵ_{max} as well. Summarizing, the total error ϵ of a sequence with t_{max} frames is the mean of all single frame errors ϵ_t:

$$\epsilon_t \quad := \quad \begin{cases} \epsilon_{max} & tracking\ failure \\ min\{\epsilon_{max}, \|\hat{x}_t - g_t\|_2\} & otherwise \end{cases}$$

$$\epsilon \quad := \quad \frac{1}{t_{max}} \sum_{t=1}^{t_{max}} \epsilon_t$$

$\| \cdot \|_2$ is the L_2-norm, $\hat{x}_t$ is the estimated object state as defined in formula (2), and g_t denotes the ground truth for frame t.

To recognize differences in terms of ϵ between different configurations of the system we carry out a comparison of means by a significance test (Duncan's test, significance niveau $\alpha = 0.01$). For illustration the 99% confidence intervals of single configurations are displayed in the figures of subsection IV-B.

IV. Results

In this section results from experiments with moving persons in realistic seminar scenes are reported. These results are obtained mainly by visual inspection of tracking examples of subsequences of the test sequences TS1, TS2, and TS3. But we also analyse tracking errors for different parameter configurations of the proposed tracking system. In the image examples the current estimated position and size of the target person are marked by a yellow rectangle with a red cross in its center. As soon as the rectangles are colored red the target person was lost (as, e.g., shown in Fig. 14). If the tracking was successful blue bars in the bottom left of the yellow rectangles visualize the confidence value of the estimation.

A. Expansion of the Condensation Algorithm

The reiteration of the Condensation algorithm as described in subsection II-B improves the tracking accuracy particularly

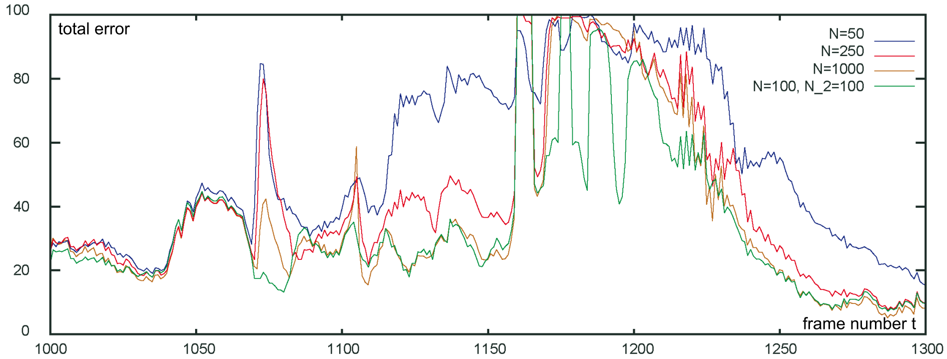


Fig. 11. Gain of the expansion of the condensation algorithm. In this diagram frame errors ϵ_t are plotted against a series of frames t of TS2. Four curves are displayed: three for three different numbers N of particels and one (the green one) for the configuration with reiteration. The chosen subsequence of TS2 contains a strong horizontal camera movement around frame $t = 1075$. As one can see, a sole increase of the number of particles cannot compensate for the camera pan, whereas the proposed expansion of the condensation algorithm can cope with it.

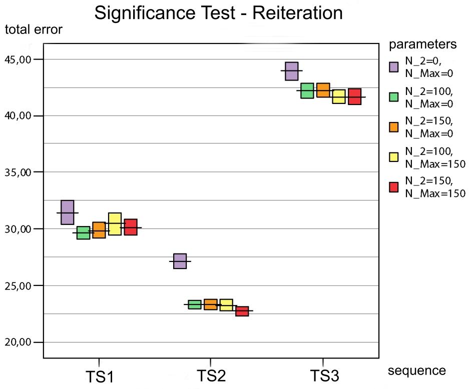


Fig. 12. Significance test for the expansion of the condensation algorithm. Total errors ϵ are plotted for different parameter configuration for the three test sequences TS1, TS2, and TS3. The errors are depicted as 99% confidence intervals.

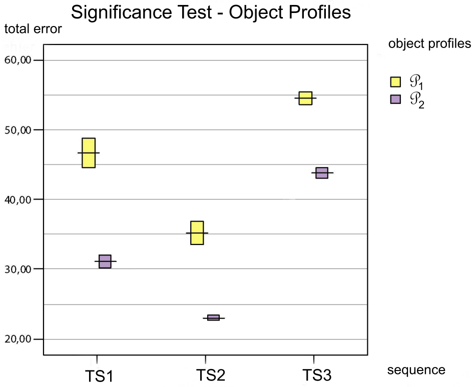


Fig. 13. Significance test for object profiles. Total errors ϵ are plotted for the two object profiles $\mathcal{P}_1$ and $\mathcal{P}_2$ for the three test sequences TS1, TS2, and TS3. The errors are depicted as 99% confidence intervals.

in the case of intense camera pans, which, e.g., occur in TS2. In Fig. 11 a diagram is presented which illustrates this effect.

In Fig. 12 results of significance tests for the three test sequences are displayed. A consideration of the results for different parameters suggests that the second iteration (i.e., the first measure of improvement reported in subsection II-B) is crucial for an improvement of the tracking. Neither the increase of the number of particles for the second iteration nor the increase of N_{max} for the next step indicates a significant improvement. Summarizing, the second iteration of the condensation algorithm is a reasonable alternative to the increase of the number of particles. It can be applied selectively for problems such as camera pans by the adjustment of the threshold α_1 for its activation.

B. Object Profiles in Comparison

In subsection II-E we introduced two different object profiles $\mathcal{P}_1$ and $\mathcal{P}_2$. A significance test for both profiles is displayed in Fig. 13. As expected, the tracking results are better for profil $\mathcal{P}_2$, because the additional color area can resolve ambiguities in a more efficient way than profile $\mathcal{P}_1$. The examples reported in the remains of this section have been obtained with profile $\mathcal{P}_2$.

C. Lost and Found - Functionality of the Particel Filter

Fig. 14 illustrates the functionality of the particel filter, which is able to relocate a totally lost target object by employing multiple hypotheses. In particular the third, fourth, and fifth image show the expansion of the particle cloud in the case of a loss of the target person. But the cloud is

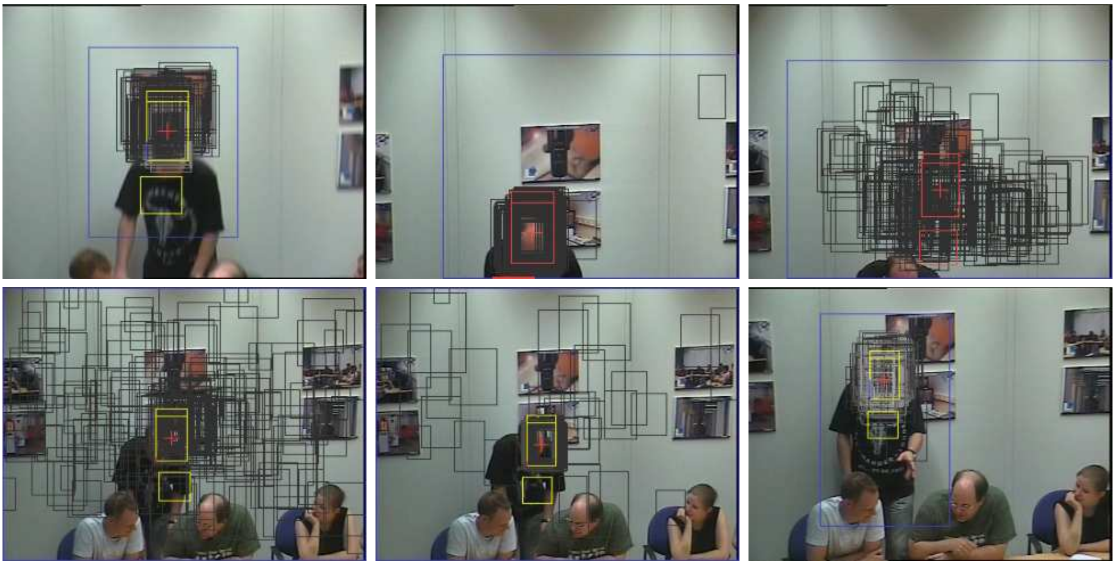


Fig. 14. Lost and found - functionality of the particel filter. Frames of TS1 for $t = 1309, 1324, 1337, 1383, 1384,$ and 1403 are displayed. (The additional partitionings in the top of the upper rectangles in this and the following figures belong to a third object profile not reported in this article.)

attracted again by the correct target person as soon the target is recognized again by at least one hypothesis.

D. Exposure to Changes in Camera Parameters - Pan

Fig. 15 shows a scene from TS2 with a moving person and a continuously panning camera. The person tracking is stable and precise even if the background is partly cluttered with other persons, although the confidence values in images 6 and 7 decrease significantly as visualized by the short confidence bars.

Fig- 16 shows another example of a sequence with a panning camera. Here the target person has a large velocity itself, and in addition, also the focal length of the camera varies. Even if the camera captures the target object only partly at the edges of a frame the estimations remain on the target.

E. Exposure to Changes in Camera Parameters - Focal Length

The other camera parameter that should be variable without affecting the tracking module is the focal length. Fig. 17 shows an example where the camera zooms into the scene. The person is tracked robustly across the different focal lengths and with high confidence values.

In Fig. 18 another example for the exposure of the tracking system to varying focal lengths of the camera is displayed. The camera zooms out of the scene and as soon as the person who disappeared in the previous frames reappears, its position and size are estimated correctly again, although now displayed at a different focal length.

F. Occlusions

Occlusions belong to the most difficult challenges to deal with in tracking systems. The number of occlusions in sequence TS3 is quite high. But the target persons are relocated altogether after a short period of time while they were occluded, frequently only a few frames after they reappeared in the captured scene. Fig. 19 shows three example scenes from TS3 which are effortlessly bridged by the tracking module. In the third image of the bottom row, for example, one can infer from the length of the confidence bar the decreased weighting of the estimation for the tracked woman. Nevertheless, she is tracked on robustly after her total occlusion.

G. Interactive Selection of Another Target Object

The selection of a new target person is done by the user of the system. She has to mark the head of the new target person. In TS3 the user has changed the target person several times. Fig. 20 shows an example where the system immediately takes the control and tracks the newly selected person.

H. Weaknesses of the Proposed Tracking System

Some weaknesses of the system striked during the experiments. One problem consists in the fact that the object profile consists of color information only. Although the expansion of profile $\mathcal{P}_1$ by a second, dynamic color histogram area improves the tracking significantly there still exist cases in which the particle cloud expands after an occlusion of the target object, and rearranges at a wrong image area with similar colors. It can happen that, even if the target object reappears in the scene, the particles remain at the wrong position until the

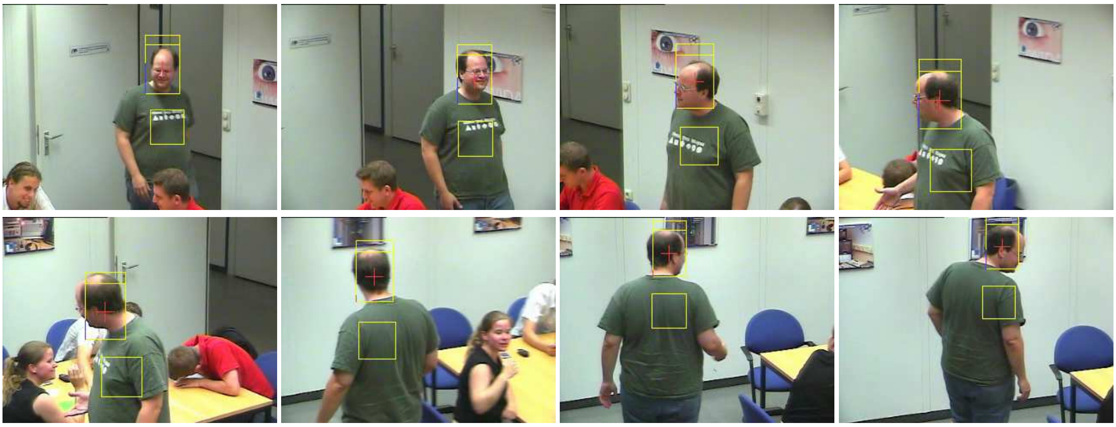


Fig. 15. Exposure to changes in camera parameters - pan. Depicted are frames $t = 475, 500, 545, 580, 615, 655, 675,$ and 700 of TS2 which render snapshots of a smoothly panning camera. Two competing reference profiles used as described in subsection II-H are crucial for the successful tracking of this sequence. The proposed system can keep track with the camera movement.

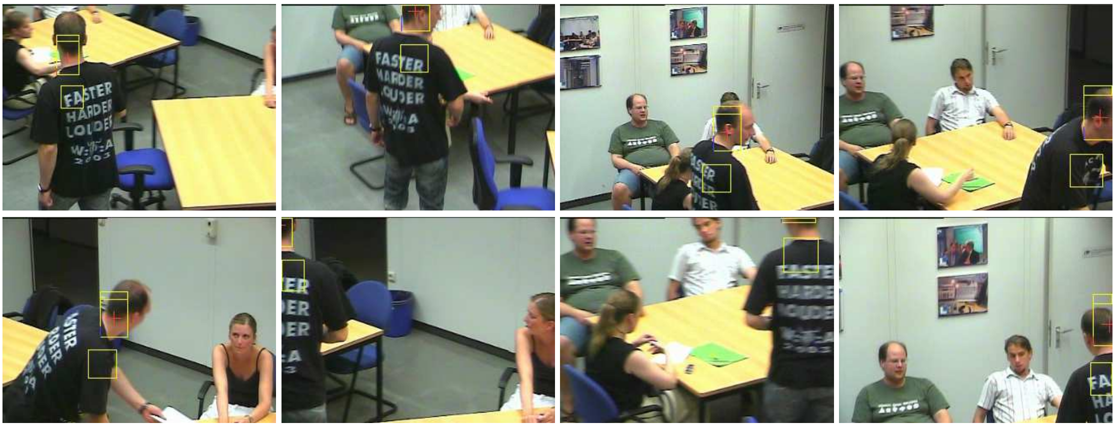


Fig. 16. Exposure to changes in camera parameters - truncations at frame edges. We see frames $t = 1897, 1919, 1935, 1960, 1980, 2000, 2010,$ and 2020 of sequence TS2.

similarity drops below the threshold and some particles detect the target again by chance.

Another weakness is the strong dependence of the tracking success on the initial selection of the target object by the user in form of a rectangle. Especially during camera movements the marking of the target can be too unprecise, resulting in the incorporation of wrong color information in the profil. This holds true as well for target persons who display a laterally crooked posture when profil $\mathcal{P}_2$ is employed. As the initial selection always places the bottom rectangle vertically below the head region, in this case it is positioned partly on the background, resulting in wrong color histograms in the object profil.

V. Conclusions

In this article we introduced an object tracking system which is capable to handle difficult situations in a dynamically changing environment. We evaluated the concepts of the proposed system (e.g., an improved version of the condensation algorithm or particle diffusion adaptive to variations in the environment) by applying it to the task of person tracking in crowded seminar rooms. The demands made on the system comprise robust real-time tracking of a target person who is allowed to move freely within a group of other persons and thus can be occluded. Furthermore, the background may change from frame to frame, and the tracking method should cope with dynamically varying camera parameters as, for example, induced by a user. In addition, the user of the system should be enabled to interactively select a new target person during tracking of another person.

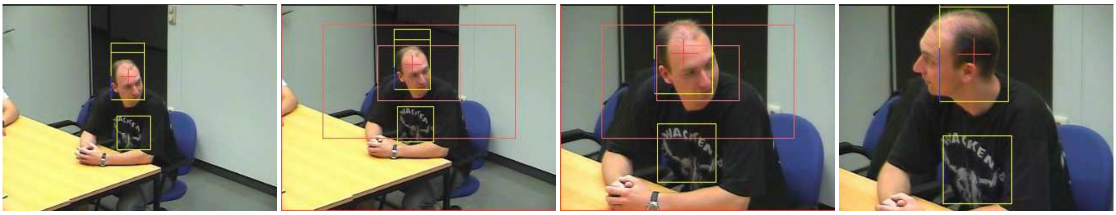


Fig. 17. Exposure to changes in camera parameters - focal length. These are the frames $t = 2620, 2661, 2715$, and 2760 of sequence TS2. (The additional red rectangles belong to the camera control module, which is not subject of this article.)

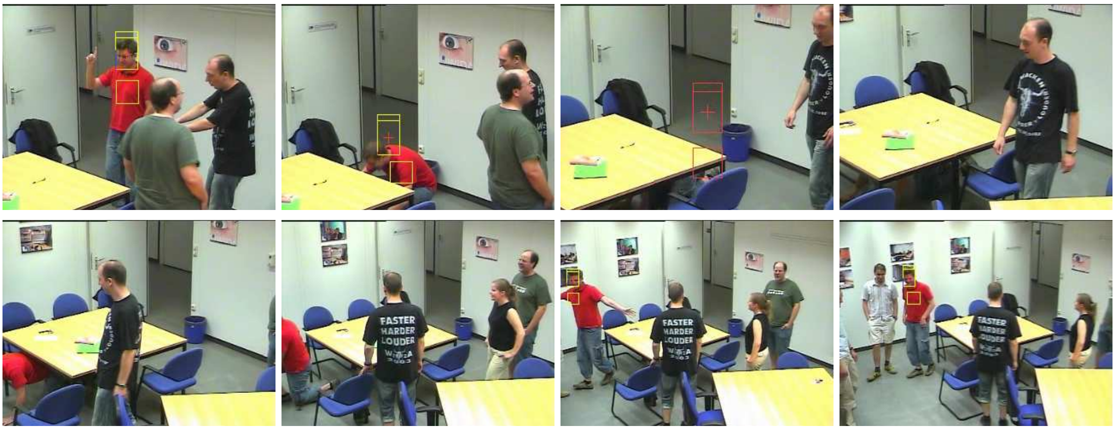


Fig. 18. Exposure to changes in camera parameters - successful relocation after zooming out. Frames $t = 1350, 1375, 1400, 1425, 1450, 1475, 1502$, and 1527 of sequence TS3 are shown.

Our contributions are threefold. First, we proposed an expansion of the condensation algorithm which results in a more stable tracking in difficult situations such as sudden camera movements. Secondly, we introduced a new way of particle diffusion which allows for the adaption of the tracking module to movements of the camera. These two contributions apply not only to person tracking but to object tracking in general. The third contribution consists in a more flexible way how to represent a person than propagated in previous publications. This contribution applies to person tracking only. Summarizing, the proposed tracking module mostly meets the postulated demands. Real-time tracking and the interactive selection of new target persons are possible. The challenges of a dynamically changing background and multiple occlusions could largely be coped with.

Ongoing research concentrates mainly on an expansion of the current object profiles, which are based on color information only. The incorporation of additional features could disambiguate situations where the current system fails because of similar color distributions in target and background. Moreover, it seems to be promising to synthesize a behavioral model which allows for the prediction of future directions of moving persons. This could represent the basis for an even more advanced and intelligent automatization.

References

[1] Kluger, M., Partikelfilterbasierter, virtueller Kameraassistent zur Verfolgung einer Person in einer Gruppe von Menschen, *Diploma Thesis*, University Dortmund, September 2006.

[2] Bianchi, M., Automatic Video Production of Lectures Using an Intelligent and Aware Environment, *Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia*, ACM Press New York, USA, pp. 117-123, 2006.

[3] Rui, Y., Gupta, A., Grudin, J., and He, L., Automating Lecture Capture and Broadcast: Technology and Videography, *ACM Multimedia Systems Journal*, 10(1), pp. 3-15, 2004.

[4] Trivedi, M., Huang, K., and Mikic, I., Dynamic Context Capture and Distributed Video Arrays for Intelligent Spaces, *IEEE Transactions on Systems, Man, and Cybernetics*, 35(A), pp. 145-163, 2005.

[5] Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., and Shafer, S., Multi-Camera Multi-Person Tracking for EasyLiving, *Proceedings of the Third IEEE International Workshop on Visual Surveillance*, Washington, USA, IEEE Computer Society, 2000.

[6] Kim, K., Chalidabhongse, T., Harwood, D., and Davis, L., Background Modeling and Subtraction by Codebook Construction, *IEEE International Conference on Image Processing*, pp. 3061-3064, 2004.

[7] Lin, C., Chang, Y., Wang, C., Chen, Y., and Sun, M., A Standard-Compliant Virtual Meeting System with Active Video Object Tracking, *EURASIP Journal on Applied Signal Processing*, 6, pp. 622-634, 2002.

[8] Nicolescu, M. and Medioni, G., Electronic Pan-Tilt-Zoom: A Solution for Intelligent Room Systems, *IEEE International Conference on Multimedia and Expo*, pp. 1581-1584, 2000.

[9] Nummiaro, K., Koller-Meier, E., Svoboda, T., Roth, D., and Gool, L., Color-Based Object Tracking in Multi-Camera Environments, *Proceedings of the 25th DAGM Symposium on Pattern Recognition*, Lecture Notes in Computer Science 2781, pp. 591-599, 2003.

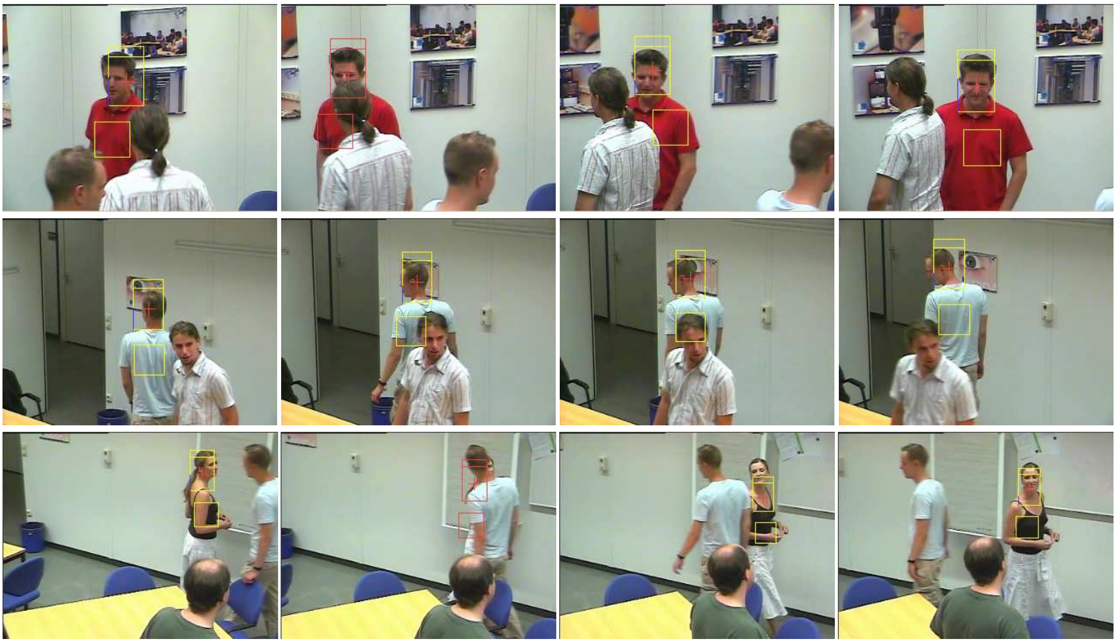


Fig. 19. Occlusions. Three sample scenes for TS3 with occlusions are shown. Upper row: frames $t = 700, 717, 726,$ and 735; middle row: frames $t = 2735, 2745, 2749,$ and 2757; bottom row: $t = 4950, 4958, 4964,$ and 4975.

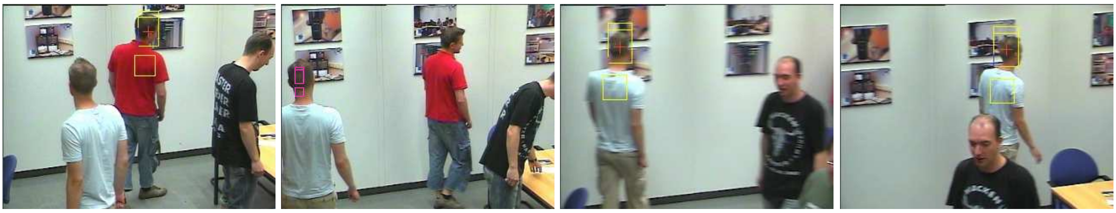


Fig. 20. Interactive selection of another target object. The frames $t = 1815, 1845, 1870,$ and 1923 of TS3 are displayed. In the second image the man with the white T-shirt is selected for tracking instead of the previously tracked man with the red T-shirt. He is successfully taken over by the tracking module.

[10] Hu, W., Tan, T., Wang, L., and Maybank, S., A Survey on Visual Surveillance of Object Motion and Behaviors, *IEEE Transactions on Systems, Man, and Cybernetics*, 34(3), pp. 334-352, 2004.

[11] Martínez-Tomás, R., Rincón, M., Bachiller, M. and Mira, J., On the Correspondence Between Objects and Events for the Diagnosis of Situations in Visual Surveillance Tasks, *Pattern Recogn. Lett.*, 29(8), pp. 1117-1135, 2008.

[12] Park, S. and Trivedi, M. M., Understanding Human Interactions with Track and Body Synergies (TBS) Captured from Multiple Views, *Computer Vision and Image Understanding*, 111(1), pp. 2-20, 2008.

[13] Moeslund, T. and Granum, E., A Survey of Computer Vision-Based Human Motion Capture, *Computer Vision and Image Understanding*, 81(3), pp. 231-268, 2001.

[14] Zhao, T., Nevatia, R., and Wu, B., Segmentation and Tracking of Multiple Humans in Crowded Environments, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7), pp. 1198-1211, 2008.

[15] Ess, A., Leibe, B., Schindler, K., and van Gool, L., Moving Obstacle Detection in Highly Dynamic Scenes, *IEEE Int. Conf. on Robotics and Automation, ICRA 2009*, 2009.

[16] Isard, M., Visual Motion Analysis by Probabilistic Propagation of Conditional Density, *PhD Thesis*, Oxford University, 1998.

[17] Nummiaro, K., Koller-Meier, E., and Gool, L., An Adaptive Color-Based Particle Filter, *Image and Vision Computing*, 21(1), pp. 99-110, 2002.

[18] Perez, P., Hue, C., Vermaak, J., and Gangnet, M., Color-Based Probabilistic Tracking, *Proceedings of the 7th European Conference on Computer Vision*, Lecture Notes In Computer Science 2350, pp. 661 - 675, 2002.

[19] Doucet, A., Godsill, S, and Andrieu, C., On Sequential Monte Carlo Sampling Methods for Bayesian Filtering, *Statistics and Computing*, 10(3), pp. 197-208, 2000.

[20] Deutscher, J., Blake, A., and Reid, I., Articulated Body Motion Capture by Annealed Particle Filtering, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2, pp. 126-133, 2000.

[21] Porikli, F., Integral Histogram: A Fast Way to Extract Histograms in Cartesian Spaces, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1, pp. 829-836, 2005.

[22] Woelk, F., Schiller, I., and Koch, R. An Airborne Bayesian Color Tracking System, *IEEE Intelligent Vehicles Symposium*, pp. 67-72, 2005.

Gabriele Peters received the PhD (Dr. rer. nat.) degree from the Faculty of Technology of the University of Bielefeld, Germany, in 2002. She carried out her PhD studies at the Institute for Neural Computation, Ruhr-University Bochum, Germany. Afterwards she worked as postdoctoral research assistant at the Department of Computer Graphics of the Technical University of Dortmund, Germany, where she focused on machine learning for computer vision and human computer interaction. For several months she worked as research scientist at the Academy of Sciences of the Czech Republic in Prague in the field of image processing and as visiting professor in the Computational Vision Group at the California Institute of Technology in Pasadena, USA, in the field of computational photography. Since 2007 she is professor at the University of Applied Sciences and Arts in Dortmund, Germany, where she heads the Visual Computing Group. She is author of more than 40 peer-reviewed scientific publications. Among other awards, in 2003 Prof. Peters received the Rudoploh Chaudoire award of the Technical Universitiy of Dortmund for her scientific achievements. Since 2004 she is an elected member of the executive committee of the Association for Informatics (GI) and is granted since 2005 by the German Research Association (DFG).

Infrastructure of an Adaptive Multi-agent System for Presentation of Mathematical Expression to Visually-Impaired Users

Ali Awde, Chakib Tadj, Yacine Bellik

Abstract - **The infrastructure of our adaptive system is aimed at correctly presenting a mathematical expression to visually impaired users. Its design is based on a multi-agent system that determines the appropriate presentation format based on the given interaction context (i.e. combined user, environment and system contexts) and the expression complexity as well as the user preferences. The architecture of our infrastructure is layered, thus encapsulating the components of the various layers. The system design is intended to be adaptive, fault tolerant and is capable of self-adaptation under varying conditions (e.g. missing or defective components). In case of failure of media device, our system is capable of replacing the faulty component with a new one (if a replacement is available). If replacement is not possible, the system re-determines the new modality and presentation format apt for the new configuration. The human intervention is greatly reduced in our system as it is capable of self-configuration, and learning. In our work, agent communication simulation has been carried out on the Java Agent Development Framework (JADE[*]) platform. This work is an original contribution to the ongoing research in helping the visually-impaired users to become autonomous in using the computing system. Our aim is to improve the computing productivity of a visually-impaired user.**

Index Terms— **Adaptive system, fault-tolerant system, mathematics for visually-impaired users, multi-agent system.**

I. Introduction

MATHEMATICS is a fundamental foundation of science. Understanding science is impossible without knowing mathematics. Mathematics for visually-impaired users, however, is a challenging task due to the following reasons: First, the visual mathematical representation is bi-dimensional and the interpretation of a mathematical expression is related to one's knowledge of the expression's individual spatial components. Second, the conversion of a multi-dimensional structure to a non-visual representation is a difficult problem. For example, the representation of a mathematical expression in Braille requires supplementary information to denote some components in order for the blind users to read the expressions easily. Also, the conversion of mathematics into an audio format is often ambiguous. Third, the vocabulary terms used by sighted people in a mathematical document are quite large compared with the amount of data accessible by a visually-impaired user. For example, in a standard 6-dot Braille[†], we can encode 64 characters. This number of symbols is, however, insufficient to represent all frequently used mathematical symbols. Also, large number of symbols is a big challenge to blind users. For instance, Braille characters are often embossed into a paper which is a static media; hence, user will not be able to manipulate the data easily. Indeed, some flexible methodologies are needed to allow blind users to access and navigate terms in a mathematical expression easily.

This paper presents the challenges that underlie in designing such infrastructure that provides presentation of mathematical expressions to visually impaired users, and how we address the problem cited above by proposing an intelligent multimodal computing system that interacts with the user, capable of choosing modalities and media devices based on a given interaction context. The format for the presentation of a mathematical expression is selected based on available media devices, user's preferences and context of the mathematical expression. The proposed infrastructure is a multi-agent system. The design of this multi-layered infrastructure enables every layer's calculation and decision making be hidden from other layers. In this way, the possible propagation of ripple effect during any stage of software life cycle becomes limited and restricted only within the boundaries of the concerned layer. Various layers communicate among themselves via parameters passing. In this work, we discuss the design of each agent and its responsibilities. Also, we

Manuscript received February 15, 2009. This work has been made possible through the funding provided by LIMSI laboratory, LATIS laboratory and the Décanat aux études of École de technologie supérieure (ÉTS).

A. Awde is with École de technologie supérieure, Montréal, Canada. Phone: 514-396-8800; fax: 514-396-8684; Email: ali.awde.1@ens.etsmtl.ca.

C. Tadj is with École de technologie supérieure, Canada. Email: ctadj@ele.etsmtl.ca.

Y. Bellik is with Université Paris Sud XI, France. Email: Yacine.bellik@limsi.fr.

[*] JADE: http://jade.tilab.com/

[†] http://6dotbraille.com

present our fault tolerant system design and a JADE simulation of agent communication.

This work is our contribution to make mathematics more accessible to visually-impaired users. It is aimed at providing some autonomy to blind users when dealing with mathematical expressions. The rest of this paper is structured as follows. Section 2 provides a brief review of the state of the art related to our work; Section 3 lists down the technical challenges in this work and essays our approach to address each one of them. Section 4 presents the individual components of our adaptive multimodal computing system. In Section 5, the principles used for the knowledge acquisition of our learning agent are discussed. Examples are provided in Section 6 while Section 7 presents some formal specifications related to the functionalities of the system. The future works and conclusion are presented in Section 8.

II. Review of the State of the Art

To a visually-impaired user, understanding mathematical expression requires repeated passage over the expression, sometimes skipping some secondary information, only to revert back to it again and again until the user fully grasps the expression. A complicated task like this is detailed in [1]. Some tools, however, have been developed to lessen the complexity of performing similar task, among them being the Mathtalk [2], Maths [3], DotsPlus [4], EasyMath [5], and AudioMath [6, 7]. MathTalk and Maths convert a standard expression into audio information. In Maths, the user can read, write and manipulate mathematics using a multimedia interface containing speech, Braille and audio. Raman developed Aster [8], a program that takes a Latex document and reads it loud using several audio dimensions that make up the different components of the expression. VICKIE [9] and BraMaNet [10] are transcription tools that convert mathematical document (written in Latex[‡], MathML[§], HTML, etc.) to Braille representation. DotsPlus is a tactile method of printing documents that incorporates both Braille and graphic symbols (e.g. ∏, ∑, etc.) In EasyMath, regardless of using Braille or overlay keyboard, its main focus is to keep the general structure of mathematical expressions intact.

None of these tools, however, is complete. Studies have been made for evaluating these tools based on users' needs [5, 11]. Results indicate that users are neither independent nor able to do their homework (in case of students) without the help of sighted people. Indeed, each tool has its own set of usage limitations. For example, Aster transforms only a LaTex document into an output suitable for speech while AudioMath transforms only a MathML input document into a speech output. Our approach, therefore, is to get the strength of each tool, integrate each one of them into our work in order to build a system that (1) broadens the limits of utilization, (2) provides the user with opportunities to access as many document types as possible, and (3) presents data output in as many suitable formats as possible after considering user situation and the special symbols within the expression. This work is an essential contribution because we offer all types of data presentation formats yet requires minimum intervention from the user.

HOMERE [12] is a multimodal system that allows visually-impaired users to use haptic/touch and audio modalities to explore and navigate virtual environments. In comparison, our approach is better because there are no pre-defined input-output modalities; the selected modalities are chosen according to their suitability to user's context. To visually-impaired users, multi-modality is even more important as it provides them equal opportunities to use informatics like everybody else. In determining the appropriate modality, the user situation plays an important role. In our work, the notion of user context includes additional handicaps and user preferences on the priority rankings and parameter settings of media devices and presentation formats.

An agent is some software that senses its environment and is capable of reaction, proactivity, and social interaction. A group of agents in a system forms a Multi-Agent System (MAS) [13]. Agents and MAS [14, 15] have been widely used in many applications, from relatively small systems such as email filters up to large, open, complex, mission-critical systems such as air traffic control [16]. Generally, it is preferred over traditional techniques (i.e. functional or object-oriented programming) because the latter is inadequate in developing tools that react on environment events. Significant works on MAS for visually impaired include [17, 18]. For example, Tyflos [17] could help a visually impaired user to be partially independent and able to walk and work in a 3-D dynamic environment. Our work, in contrast, uses agents to detect user context, and other data in order to assist the system to determine the media and modalities that are appropriate for the user.

III. Technical Challenges

In this section, we resume the problems in designing a system that will present a mathematical expression to visually impaired users, pose specific technical challenges that need to be solved and describe our

[‡] L. Lamport, LaTeX: The Macro Package, http://web.mit.edu/texsrc/source/info/latex2e.pdf, 1994

[§] MathML, http://www.w3.org/Math

approach to address those challenges.

In our proposed infrastructure, we envision a system that is rich in media devices and modality selection, data formats, techniques in converting one data format to another, and an adaptive interface that allows user to manipulate mathematical data. To design such a system, a solution must address key requirements, cited below:

Requirement 1: *Provide a multiagent system that coordinates all its components in an orderly manner, providing autonomy to the user and is able to adapt automatically to a given interaction context.* How do we design a multi-layered, multiagent system that satisfies the system requirement? What mechanisms must be adopted to make the system tolerant from faults?

Requirement 2: *Provide a mechanism that allows selecting the appropriate media device supporting the selected modality.* How do we design a system that satisfies the user preferences? How the media device will be detected? Which media device will be activated to support a specific modality?

Requirement 3: *Provide an infrastructure for analysis and conversion of a mathematical expression, embedded within a document (in MathML format) into its corresponding encoded format and then into its presentation format.* How to convert an expression written in MathML format into an expression in encoded format and into an expression using a presentation format such as the Braille, DotsPlus, EasyMath and audio?

Requirement 4: *Provide a mechanism that allows visually impaired users to manipulate terms in a mathematical expression.* How do we design a system component that allows the user to add, modify and delete mathematical terms, and to search the expression for a term in random manner?

The rest of this paper addresses the technical challenges by providing specific solutions to the system requirements cited above.

IV. The Components of an Adaptive Multimodal Computing System

A. Our Adaptive Multimodal Computing System for Mathematical Expression Presentation to visually impaired users

Fig. 1 shows the layered view of an adaptive multimodal computing system that presents mathematical expressions to visually-impaired users. The layers and their roles are as follows: (1) *Physical Layer* – contains all the physical entities of the system, including devices and sensors. The raw data from this layer are sampled and interpreted and forms the current instance of interaction context. (2) The *Context Gathering Layer* – here, the interaction context (of tuple <user, system, environment>) is detected; (3) The *Control and Monitoring Layer* – it controls the system, coordinating the detection of user's interaction context, the mathematical expression, its presentation and/or manipulation; (4) The *Data Analysis Layer* – here, the presentation format of the mathematical expression is selected based on available resources and user's situation; (5) The *Data Access Layer* – in this layer, mathematical expression may be searched or edited by the user; (6) The *Presentation Layer* – here, the mathematical expression is presented through an optimal presentation format.

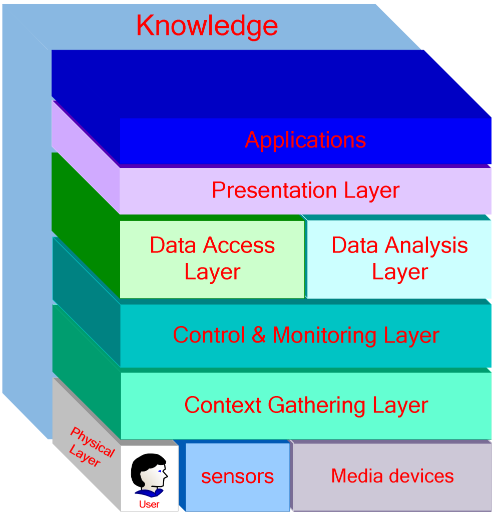


Fig. 1. The architectural abstraction of a generic MM computing system for visually-impaired users.

Fig. 2 shows our model of a multi-agent adaptive multimodal system for visually-impaired users. The agents' functionalities in the system's layers are detailed in the next sections.

B. Modality and Media

In our work, we adopt the concepts of media and modality that are defined by Bellik in [19].

1. *Modality* is defined by the information structure as it is perceived by the user (e.g. text, speech, Braille, etc.).

2. *Media* is defined as a device used to acquire or deliver information or data (e.g. screen, terminal Braille, mouse, keyboard, etc.)

Here, *Vocal* and *Tactile* modalities are possible since we address visually impaired users. Also, in general, interaction is possible if there exists at least one modality for data input and at least one modality for data output. Given a modality set M = $\{V_{in}, T_{in}, V_{out}, T_{out}\}$ wherein V_{in} = *vocal input,* V_{out} = *vocal output,* T_{in} = *tactile input and* T_{out} = *tactile output* then interaction is possible under the following condition:

$$Interaction\ Possible = (V_{in} \vee T_{in}) \wedge (V_{out} \vee T_{out}) \qquad (1)$$

where the symbols $\wedge$ and $\vee$ denote logical AND and OR, respectively. There are usually more than one media that support a specific modality (see section C-3).

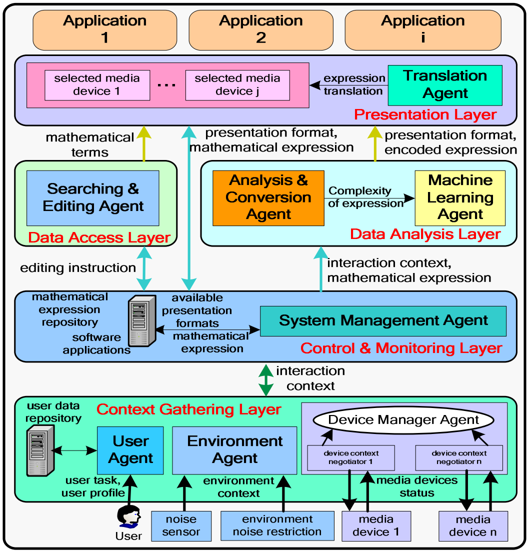


Fig. 2. Architectural layer view of the multimodal computing system

C. *The Context Gathering Layer*

In this layer, the user's interaction context is identified. There are three agents that obtain the context of the *user*, the *environment*, and the *system* which collectively form the interaction context. Here, we present briefly the interaction context, for more details see [20].

1. The User Context

In this work, the user context (UC) is a function of user profile (including any handicap) and preferences. The *user agent* (UA) detects the user's profile and preferences. We have a *user data repository* where user's task (i.e. the mathematical expression) is stored. In such repository, there is a *user profile* (UP) for the user, which contains, among others, the user's username, password, his computing devices and identifications (i.e. IP addresses) and his special needs (i.e. handicap). This information is useful for determining the suitable modalities. For example, being mute prevents the user from using vocal input modality.

2. The Environment Context

The *environment's context* (EC) detected by Environment Agent (*EnvA)*, is the assessment of a user's workplace condition. In this work, EC is based on the following parameters: (1) the workplace's *noise level* – identifies if it is quiet/acceptable or noisy, and (2) the *noise level restriction* – identifies whether a workplace imposes mandatory silence or not. For example: in a library where silence is required, sound producing media (e.g. speaker) needs to be muted or deactivated.

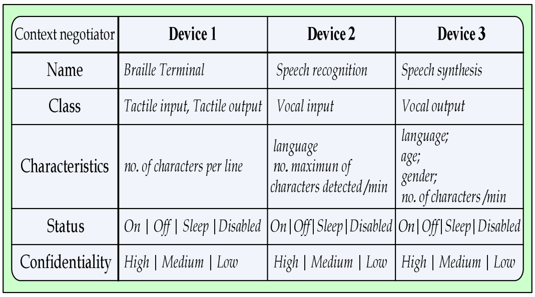
The noise level is interpreted by EnvA from the sampled raw data of a sensor. In our work, *50 dB or less* is considered "*acceptable*" while *51 dB or more* is considered "*noisy*".

For environment noise restriction, we have a database of pre-defined places (e.g. library, park) and their associated noise restrictions (e.g. library: silence required, park: silence optional). User can update and modify some database records.

3. The System Context

In this work, the *system context* (SC) implies the user's computing device and the available media devices. SC is managed by the *Device Manager Agent* (DMA). The computing device (e.g. PC, laptop, PDA, cellular phone) affects the modality selection. For example, using a PDA or cell phone prevents user from using tactile input or output modality. On the other hand, some of the most commonly-used media devices suitable for blind users are: (i) *Keyboard*; (ii) *Microphone*; (iii) *Speech Recognition* (iv) *Speech Synthesis* (v) *Speaker*; (vi) *Headset*; (vii) *Braille Terminal* (viii) *Overlay or Concept Keyboard* (ix) *Tactile Printer or Embosser.*

Every media device has its own *device context negotiator* (DCN). A DCN is an agent that detects the media device's context; it is a link between the actual media device and the DMA. Every DCN has the following attributes (see Fig. 3): its *name, class, characteristics, status*, and *confidentiality. "Name"* identifies the media device it manages (e.g. "Braille context negotiator" detects the context of a Braille terminal). *"Class"* identifies its form of modality. *"Characteristics"* identify the unique features of the device. *"Status"* identifies if the device is on, off, sleep or disabled. *"Confidentiality"* is the perceived reliability of the device and is denoted as high, medium or low.



Context negotiator	**Device 1**	**Device 2**	**Device 3**
Name	*Braille Terminal*	*Speech recognition*	*Speech synthesis*
Class	*Tactile input, Tactile output*	*Vocal input*	*Vocal output*

Characteristics	*no. of characters per line*	*language*
		no. maximum of characters detected/min
			language;
			age;
			gender;
			no. of characters/min

Status	*On | Off | Sleep | Disabled*	*On | Off | Sleep | Disabled*	*On | Off | Sleep | Disabled*
Confidentiality	*High | Medium | Low*	*High | Medium | Low*	*High | Medium | Low*

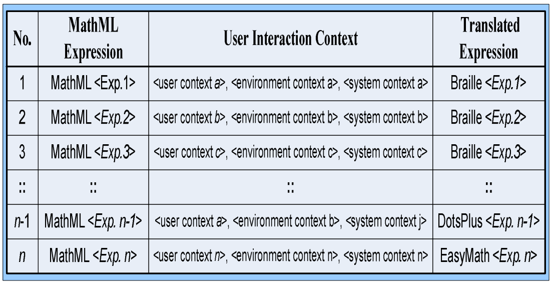
Fig. 3: Attributes of device context negotiator for Braille terminal, speech recognition and synthesis

D. *The Control and Monitoring Layer*

System Management Agent (SMA) monitors and reports on the system configuration and application activity. When a fault occurs, SMA should deal with errors and find solution to keep system working (see section V).

Upon detection of interaction context, the control and monitoring layer via its SMA examines the mathematical expression in the user's task. It determines the available presentation formats by sensing the presence of conversion software in the computer. It also determines if the current mathematical expression and interaction context are already listed in the system's *mathematical expression repository* (MER). If so, then the expression is sent to Presentation Layer for its presentation. Otherwise, the system must learn how to present this expression by sending all the available information to the Data Analysis Layer.

The MER is a *private* database that keeps all mathematical expressions that have already been encountered. All expressions are stored in a tabular form; each table entry contains (1) the original expression in MathML format, (2) the user interaction context, and (3) the translated expression. Fig. 4 shows the MER contents in generic format. Hence, an entry in MER prevents the unnecessary repetition of calculations and analysis that were done when the condition was first encountered. The MER is scalable. Its contents are time-bounded (i.e. similar to electronic mail); its information is periodically updated, and very old records are deleted.



No.	MathML Expression	User Interaction Context	Translated Expression
1	MathML <*Exp.1*>	<user context *a*>, <environment context *a*>, <system context *a*>	Braille <*Exp.1*>
2	MathML <*Exp.2*>	<user context *b*>, <environment context b>, <system context b>	Braille <*Exp.2*>
3	MathML <*Exp.3*>	<user context *c*>, <environment context *c*>, <system context *c*>	Braille <*Exp.3*>
::	::	::	::
n-1	MathML <*Exp. n-1*>	<user context *a*>, <environment context b>, <system context j>	DotsPlus <*Exp. n-1*>
n	MathML <*Exp. n*>	<user context *n*>, <environment context n>, <system context n>	EasyMath <*Exp. n*>

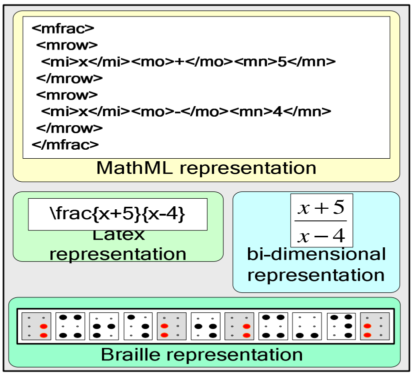
Fig. 4: The mathematical expression repository, in generic format.

E. *The Data Analysis Layer*

Here, we present the analysis and learning methods that are invoked in determining the optimal presentation of a mathematical expression based on a given interaction context.

1. The Visually-Impaired User's View of a Mathematical Expression

To a visually-impaired user, a *simple* mathematical expression (e.g. quadratic equation) becomes complex due to the presence of elements such as the *subscript*, *exponent*, *mathematical symbols* (e.g. , etc.), and the expression's *dimension* (i.e. complex numerator, denominator). In informatics, a mathematical expression is generally written using the syntax of *MathML* or LaTex which are inappropriate format for visually impaired users. For example, a simple fraction in Fig. 5 is shown with its equivalence in MathML, LaTex, Braille and its linear representations.



```
<mfrac>
 <mrow>
  <mi>x</mi><mo>+</mo><mn>5</mn>
 </mrow>
 <mrow>
  <mi>x</mi><mo>-</mo><mn>4</mn>
 </mrow>
</mfrac>
```

\frac{x+5}{x-4}

$$\frac{x+5}{x-4}$$

Fig. 5: A fraction in bi-dimensional form and its corresponding equivalent in LaTex, MathML and Braille.

2. Representation of Mathematical Operation in Different Formats

Presentation formats use different methods to represent a mathematical operation. Using *Braille*, there is a unique symbol for every operation. Using *speech*, an operation is uttered using a specific word (e.g. "+" is "add", "-" is "minus", etc). Using *DotsPlus*, an operation is represented by a unique symbol in a tactile form. Using *EasyMath*, every basic operation (e.g. +, -, ×, ÷, etc.) is represented by a unique symbol similar to its Braille representation. For special operation (e.g. , , log, ∫, ∬, ∭, etc), however, the representation is in tactile form. Note that the representation of special operations in DotsPlus and EasyMath are not the same. For example, the + operation symbol is represented by a Braille symbol whereas in DotsPlus it is represented as "+" in embossed tactile form.

3. The expression complexity

The complexity of the expression affects the choice of the format of presentation. In case of simple expressions (see Fig. 6, expression (b)), the user will choose simple presentation format such as Braille or audio. Note that when the expression is complex, user has to choose more complex presentation format such as DotsPlus's presentation (e.g. expression (a) in Fig. 6). Hence, the complexity of the expression is important for determining the suitable presentation format. In [21], authors proposed a method to determine the complexity of the expression based on: the depth of syntax tree, number of operands and operators.

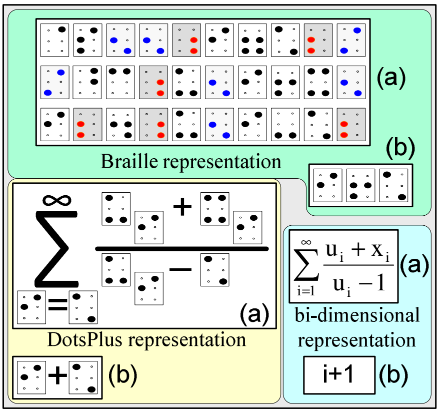


Fig. 6: Two sample expressions in bi-dimensional form and its corresponding equivalent in Braille and DotsPlus format.

4. The Analysis and Conversion Agent

Fig. 7 shows the functionalities of Analysis and Conversion Agent (ACA). It receives a mathematical expression (in MathML format) from SMA. Using *grammar rules and dictionary*, the expression is analyzed lexically. The result yields a list of lexemes. A lexeme is a parameter within an expression which may be an operand or an operator. Given the lexemes, the *parser* analyzes the expression parameters (i.e. operands, operators and syntax tree) then sends parameters to the *Expression Evaluator* to determine its complexity. The *parser* sends then the expression to the *Expression Encoder* to be translated into its encoded format.

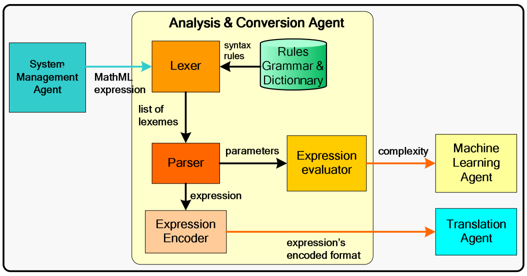
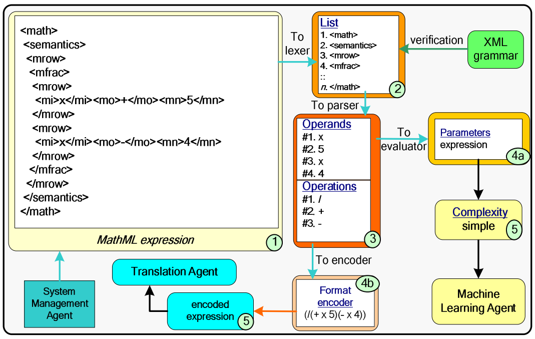


Fig. 7: The Analysis and Conversion Agent.

As an example, Fig. 8 shows the fraction defined in Fig. 5, as a specimen expression. In (step 1), the expression is sent to the Lexer. In (step 2), using the XML grammar, the expression is decomposed into a list of lexemes; the list is then sent to the parser. In (step 3), the operations and operands in the expression are sent to expression evaluator and encoder. Together, in (step 4), the evaluator deduces the complexity of the expression (e.g. simple) while the encoder produces the encoded expression. Finally, in (step 5), the MLA is informed about the complexity of the expression, while the encoded expression is forwarded to the Translation Agent (TA).

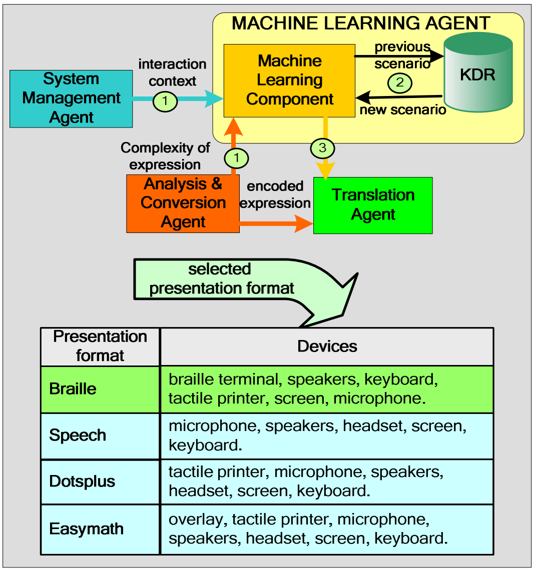


```
<math>
 <semantics>
  <mrow>
   <mfrac>
    <mrow>
     <mi>x</mi><mo>+</mo><mn>5</mn>
    </mrow>
    <mrow>
     <mi>x</mi><mo>-</mo><mn>4</mn>
    </mrow>
   </mfrac>
  </mrow>
 </semantics>
</math>
```

Fig. 8: A sample analysis of a specimen fraction.

5. Determining a Mathematical Expression's Presentation Format

The MLA selects the presentation format based on interaction context and the complexity of the expression. The selection of the appropriate presentation format and the learning process are detailed in [20]. Here, we present briefly the functionalities of MLA that are depicted in Fig. 9.



Presentation format	Devices
Braille	braille terminal, speakers, keyboard, tactile printer, screen, microphone.
Speech	microphone, speakers, headset, screen, keyboard.
Dotsplus	tactile printer, microphone, speakers, headset, screen, keyboard.
Easymath	overlay, tactile printer, microphone, speakers, headset, screen, keyboard.

Fig. 9: The Machine Learning Agent and its interaction with other system components.

In (step 1), MLA receives interaction context from SMA. The ACA informs the MLA of the complexity of the expression as discussed in previous section. The interaction context and the complexity of the expression input to the *machine learning component* (MLC) forms the pre-condition scenario. In (step 2), information about the corresponding post-condition is searched. If it is empty (i.e. it is a new scenario) then the MLA determines the modality that is appropriate for the interaction context. Next, it determines which available media devices support the chosen modality. Using the complexity of the expression, the MLA finally decides the optimal presentation format. The chosen presentation format becomes the post-condition scenario and is recorded in the *scenario repository* (SR). Otherwise, the user condition is not a new scenario; hence the MLA simply retrieves the entry in the post-condition scenario. In (step 3), the chosen presentation format (e.g. Braille) is forwarded to the TA. The suitable media devices also are activated as shown in the bottom of Fig. 9.

F. The Presentation Layer

Here, we present a mathematical expression presentation in a chosen format.

1. The Translation Agent

Using the chosen presentation format sent by MLA, the *Translation Agent* (TA) converts the encoded expression into its final presentation. As shown in Fig. 10, the TA controller forwards the encoded expression to format translator(s) (i.e. *Braille*, *EasyMath*, *DotsPlus*, and *Speech*). See Fig. 11 , for the translation of specimen fraction into each of the 4 formats. The translated expression is then forwarded to the selected media devices and to SMA. The SMA then updates its mathematical expression repository by storing the new information (i.e. MathML expression, user interaction context, and the translated expression), implying that a new scenario (i.e. the current one) is encountered.

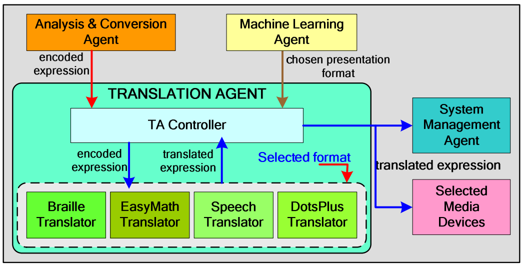
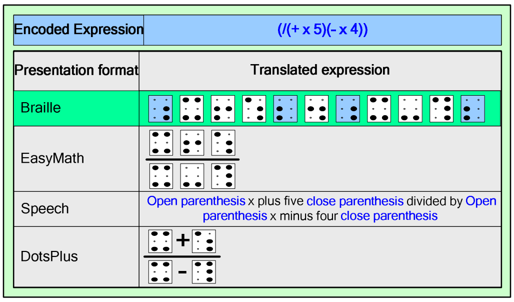


Fig. 10: The Translator Agent and its components.

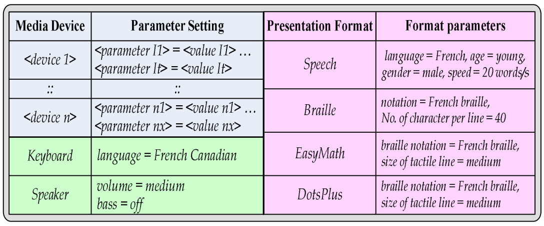


Encoded Expression	(/(+ x 5)(- x 4))
Presentation format	**Translated expression**
Braille	⠰⠭⠖⠱⠆⠲⠰⠭⠤⠹⠆
EasyMath	$\frac{\text{⠭⠖⠱}}{\text{⠭⠤⠹}}$
Speech	Open parenthesis x plus five close parenthesis divided by Open parenthesis x minus four close parenthesis
DotsPlus	$\frac{\text{⠭} + \text{⠱}}{\text{⠭} - \text{⠹}}$

Fig. 11: The translation of a specimen fraction into 4 formats.

2. Parameters Setting of Presentation Formats

TA must know the parameters of every presentation format if it is to produce a correct translation of a mathematical expression. In general, values of these parameters are set by the user himself. For example, for speech presentation, parameters such as language, gender, and speed of audio message are presented as per user specification. In Fig. 12 (Right), the parameters of 4 presentation formats (and their sample values) are shown. The presentation format's parameter settings are part of user preferences as contained in the user profile.



Media Device	Parameter Setting	Presentation Format	Format parameters
<device 1>	*<parameter I1> = <value I1> …* *<parameter It> = <value It>*	*Speech*	*language = French, age = young, gender = male, speed = 20 words/s*
::	::		
<device n>	*<parameter n1> = <value n1> …* *<parameter nx> = <value nx>*	*Braille*	*notation = French braille,* *No. of character per line = 40*
Keyboard	*language = French Canadian*	*EasyMath*	*braille notation = French braille,* *size of tactile line = medium*
Speaker	*volume = medium* *bass = off*	*DotsPlus*	*braille notation = French braille,* *size of tactile line = medium*

Fig. 12: Parameter settings for presentation formats and some media devices.

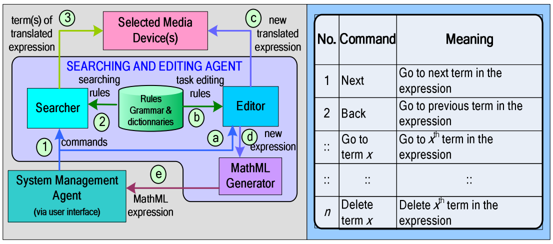
3. Parameters Setting of Selected Media Devices

The selected media devices can also be configured so that their settings suit the user's needs. Fig. 12 (Left) shows the generic format of media devices. As an example, the parameters of keyboard and speaker are set, as shown. During system activation, the parameters of each media device are set according to its media setting record by the device context negotiator. These settings, a part of user preferences, are relayed to the *Control Panel* of the operating system to effect the changes.

G. The Data Access Layer

1. The Searching and Editing Agent

The *Searching and Editing Agent* (SEA) allows navigation in and manipulation of a mathematical expression, controlling how mathematical terms are read and edited. SEA is composed of the following components: (1) the *Searcher*, (2) the *Editor*, and (3) the *MathML Generator*. SEA allows either sequential or random term access. Vocal commands for accessing mathematical terms are adopted because of their proven efficiency (i.e. cases of MathTalk and Meditor [19]). Via vocal commands, the user directly accesses an object, and perceives the modifications on the object via tactile and/or sound feedback. Fig. 13 illustrates the functionalities of SEA. Using user interface, the user can issue a vocal command (e.g. *go to <term 5>*). In (step 1), SMA sends user command to the Searcher. In (step 2), the Searcher verifies the validity of the command using grammar rules. If valid, the command is executed, and the result is sent to selected device(s) for presentation (step 3). Otherwise, no command is executed.



No.	Command	Meaning
1	Next	Go to next term in the expression
2	Back	Go to previous term in the expression
::	Go to term *x*	Go to x^{th} term in the expression
::	::	::
n	Delete term *x*	Delete x^{th} term in the expression

Fig. 13: Searching agent and editor agent cooperates with Device manager agent.

When the user edits an expression, the command (e.g. *Delete* <*term x*>) is sent by SMA to the Editor (step a). Upon command validation (step b), expression editing is executed and result is presented through selected device(s) (step c). This modification produces a new expression (step d). The MathML Generator produces the new MathML code and sent it to SMA for presentation in the user interface (step e). In Fig. 13 (Right), sample SEA commands are shown in tabular form. Note that when the presentation format is DotsPlus, search and modification of mathematical terms are not possible because the data are all embossed on paper (i.e. static media).

V. FAULT TOLERANT SYSTEM

A. *General Principle of a Fault-tolerant system*

In general, a fault-tolerant system is about the ability of a computing system to continue operating properly in the event of the failure of some of its components. It is designed to be able to handle several possible failures that may occur both in software-related faults such as communication between components or hardware-related faults such as input or output device failures.

In a fault-tolerant system, we have to proceed in three steps: 1) determine the set of faults to be tolerated; 2) choice of techniques which provide solution to the identified faults; and 3) test or experiment the efficiency of adopted techniques.

B. *Our Multi-agent Fault-Tolerant System*

The architecture of our system is designed to resist failure. When one or more faults (an agent or device is missing or defective) occur, the system would resist failure by self-reconfiguring. SMA reacts immediately to replace the failed component based on learned knowledge and users preferences (media devices priorities). Here, we can support 2 possible sources of failure: agent failure and defective or missing device.

1. A failed agent

In case of agent, SMA can perform various actions to prevent system crash. These actions are inspired from CONIC [22] and AAA [23]. The state of the agent determines action nature to be executed by the system. An agent can be in 1 of 4 states (see Fig. 14): 1) *Idle* – the agent is ready to execute actions. 2) *Running* – it is active and ready to accept and react with others queries. 3) *Disconnected* – when the agent is still alive but it does not reply correctly. 4) *Stopped* – the agent is missed and can not be repaired.

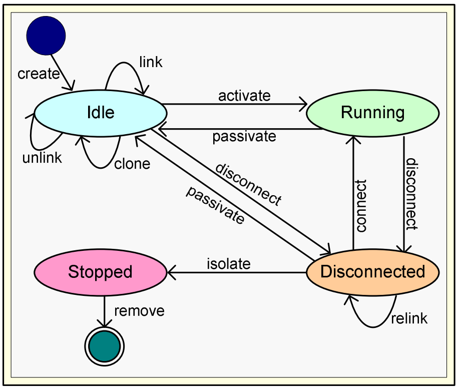


Fig. 14: All possible states of agent and primitives in our system.

In our system, the configuration is realised by using some primitives as shown on Fig. 13: 1) *create* – it is the primitive that allows adding an agent to the configuration, 2) *clone* – it is used to clone an agent in the configuration, 3) *link* – it creates a connection between 2 agents, 4) *unlink* – it removes the connection between 2 agents, 5) *re-link* – it reconstitutes the connection between 2 agents, 6) *isolate* – when an agent is not repairable, it is isolated to be deleted, and 7) *remove* – it destroys the stopped agent.

For example, if the agent A is failed, SMA tries to repair it based on its state. If there is a problem of communication between agents (i.e. agent is disconnected), primitives such as link, unlink and re-link are useful. If problem persists SMA assigns the works to the duplicate of the agent A that is already added to configuration at the beginning (i.e. using clone primitive). So the failed agent is isolated and then it is removed from the configuration.

2. A missing or defective device

Usually, when a media device is malfunctioning or absent (i.e. *failed*), the system searches the device (that is classified in the same group of modality) which is next in priority. If it is found, the replacement device is activated and the search is over. Otherwise, the system keeps searching for a replacement through *priority ranking order*.

Let there be a *media devices priority table* (MDPT) (see Table I) containing media devices grouped according to the modality they support and arranged by priority ranking. When our system implements a modality, it selects the media device(s) that is/are ranked top in priority. It is also through the MDPT that the system searches for a replacement to a failed media device.

TABLE I. A SAMPLE MEDIA DEVICES PRIORITY TABLE (MDPT)

Modality	Media Devices by Priority			
	1	2	3	n
Vocal input (V_{in})	microphone, speech recognition			
Vocal output (V_{out})	speaker, speech synthesis	headset		
Tactile input (T_{in})	keyboard	Braille terminal	overlay	
Tactile output (T_{out})	Braille terminal	tactile printer		

When a *new media device* $\mathbf{d_{new}}$ is *added* or *introduced* to the system for the *first time*, the device is associated to a modality and is given a priority ranking **r** by the user. What happen to the rankings of other devices $\mathbf{d_i}$¸($1 \le i \le n$¸ and n = number of media devices) which are in the same modality as $\mathbf{d_{new}}$ in the MDPT? Two things may happen, depending on the user's selection. The first possibility is that after having the new device's priority **Priority**($\mathbf{d_{new}}$) set to ***r*** then the priority of the other device **i**, ($1 \le i \le n$) denoted **Priority**($\mathbf{d_i}$)**,** remains the same. The second possibility is the priority rankings of all media devices ranked **r** or lower are adjusted such that their new priority rankings are one lower than their previous rankings. Formally, in Z [24], this is specified as: $\boldsymbol{\forall i, \exists r: \mathbb{N}; \forall d_i, \exists d_{new}: Devices \mid (Priority(d_{new}) = r \wedge Priority(d_i) \geq r) \Rightarrow Priority(d_i)' = Priority(d_i) + 1.}$

When a media device fault is detected and replaced by another one, the selection of the appropriate presentation format may be affected. Then ML must search the optimal presentation format based on the new situation. The process of selection of presentation format is discussed in [20].

VI. EXAMPLE OF SIMULATION WITH JADE

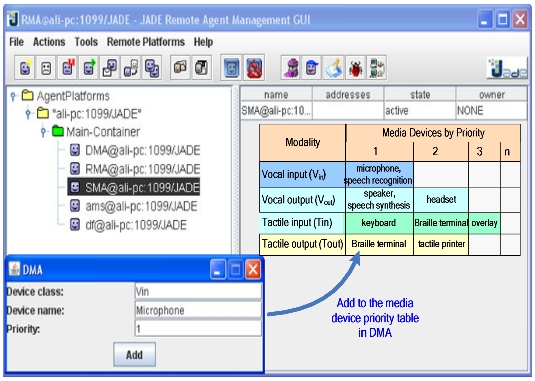
JADE [25] (Java Agent Development framework) is a software framework and middle-ware aimed at developing multi-agent applications conforming to FIPA standards. JADE is an Open Source project and has been coded in Java. Using this framework, a programmer should code his agents in Java. JADE provides an implementation for the following components:

- Agent Management System (AMS). This agent is responsible for controlling access to the platform, authentication and registration of participating agents.
- Directory Facilitator (DF). This agent provides a yellow page service to the agents in the platform.
- Agent Communication Channel (ACC). This agent provides a white page service. It also supports inter-agent communication and inter-operability within and across different platforms.

When a JADE platform is launched, AMS and DF are immediately created and ACC module is set to permit communication between agents by set of messages. The agent platform can be distributed on several hosts. AMS and DF live in the main-container that is an agent and it contains the RMI registry used internally by JADE. The other agents created should be connected to the main-container.

The components of a multi-agent system implemented with JADE communicate with each other using flexible and efficient messaging services. According to the FIPA specification, agents communicate via asynchronous message (i.e. Agent Communication Language ACL messages) and the communication between agents involves an exchange of ACL messages. ACL is conceived in a formal language that avoids any ambiguity.

Fig. 15 shows a Graphic User Interface (GUI) generated by the general management console for a JADE that is called RMA (Remote Management Agent). RMA provides control of all registered agent within platform, acquires information about the platform and executes GUI commands as create new agent, kill agent, etc. Through the RMA, a sniffer agent which is an important tool of JADE for monitoring and checking ACL messages exchanged among agents. When we sniff one agent or more, every message incoming/outgoing to/from agent is tracked and displayed in the Sniffer Agent's GUI in a diagram similar to UML sequence diagrams. There are other useful tools in JADE (Dummy Agent, Introspector) for monitoring and debugging the multi-agent systems. In this example, we present a simple simulation of 2 agents of our system: SMA and DMA.



name	addresses	state	owner
SMA@ali-pc:10...		active	NONE

Modality	Media Devices by Priority			
	1	2	3	n
Vocal input (V_{in})	microphone, speech recognition			
Vocal output (V_{out})	speaker, speech synthesis	headset		
Tactile input (Tin)	keyboard	Braille terminal	overlay	
Tactile output (Tout)	Braille terminal	tactile printer		

Device class: Vin
Device name: Microphone
Priority: 1

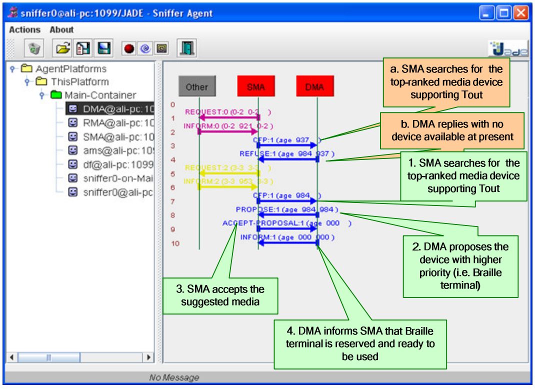
Fig. 15: A simple Jade simulation of SMA and DMA, also a sample MDPT.

In our simulation, we demonstrate the communication and the coordination between SMA and DMA. Usually, DMA communicates with all negotiators that are responsible to determine the status of each device and fills the MDPT. User must only specify its preferences.

However, we use a user interface for DMA in order to fill the MDPT as shown in Fig. 15.

At the beginning, SMA has a modality and searches to select the media device that is top-ranked in priority. To do that, SMA communicates with DMA as shown on Fig. 16.

Messages (a and b) represent the case when there are no device of modality T_{out} available in the MDPT. Messages (1,2,3 and 4) present a typical scenario when DMA has at least one device of the modality searched by SMA (i.e. here it is T_{out}). First, SMA looks to select the best available media device supporting T_{out} (a or 1) to DMA. If it is found, as in our example, it replies with a proposal message (2) that contains name (Braille Terminal) and rank (10) of the found media. Otherwise, it replies with a refuse message (b) to inform SMA that no device is available supporting T_{out}. In case of proposal message, SMA replies with an acceptance (3) and DMA informs (4) SMA that the media device (i.e. Braille Terminal) is ready to be used. Also, these steps are executed in finding replacement to a failed device.



REQUEST:0 (0-2 0-2)
INFORM:0 (0-2 921 0-2)
CFP:1 (age 937)
REFUSE:1 (age 984 937)
REQUEST:2 (3-3 3-3)
INFORM:2 (3-3 953 3-3)
CFP:1 (age 984)
PROPOSE:1 (age 984 984)
ACCEPT-PROPOSAL:1 (age 000)
INFORM:1 (age 000 000)

a. SMA searches for the top-ranked media device supporting Tout

b. DMA replies with no device available at present

1. SMA searches for the top-ranked media device supporting Tout

2. DMA proposes the device with higher priority (i.e. Braille terminal)

3. SMA accepts the suggested media

4. DMA informs SMA that Braille terminal is reserved and ready to be used

Fig. 16 sniffer agent monitors and checks ACL messages exchanged among agents SMA and DMA.

VII. CONCLUSION

Our ongoing research is focused on providing computing infrastructure to visually-impaired user through multimodality. One area of such domain is the infrastructure for mathematical presentation to blind users which this paper addresses. In this work, we presented an infrastructure supporting the presentation of mathematical expressions. Our multi-agent system considers the interaction context (i.e. combined user's, environment's and system's contexts) as well as the nature of the mathematical expression itself and of the user's preferences.

In this paper, we have presented the architecture of our adaptive multi-agent system. Also, we have presented the agents' functionalities in the system's layers. For each layer, we have shown the agents and their behaviour.

The architecture of our infrastructure is layered, thus encapsulating the components of the various layers. It is adaptive that it is capable of determining the best configuration (modality, media, and presentation) for the user. In case of failure of media device, our system is capable of shutting down the faulty component and replacing it with a new one (if a replacement is available). If replacement is not possible, the system re-determines the new modality and presentation format apt for the new configuration. The human intervention is greatly reduced in our system as it is capable of self-configuration, and learning. This system feature promotes autonomy to visually-impaired users, thus enhancing their information processing productivity.
In order to demonstrate the behaviour of the agents, a simulation has been carried out on the Java Agent Development Framework (JADE) platform. This simulation has confirmed the efficiency of our system design.

This work is our continuing contribution to advance research on making informatics more accessible to handicapped users. Our future works involve the prototyping of this infrastructure and simulating its performance using several computing platforms. Such prototype will also be tested on visually-impaired users with other varying interaction contexts.

REFERENCES

1. Stöger, B., K. Miesenberger, and M. Batusic, *Mathematical Working Environment for the Blind Motivation and Basic Ideas*, in *ICCHP*. 2004, Springer. p. 656-663.
2. Edwards, A.D.N. and R.D. Stevens. *A Multimodal Interface for Blind Mathematics Students*. in *INSERM*. 1994. Paris, France.
3. Cahill, H., et al., *Ensuring Usability in MATHS*, in *The European Context for Assistive Technology*. 1995, IOS Press: Amsterdam. p. 66-69.
4. Preddy, M., et al. *Dotsplus: How-to make tactile figures and tactile formatted math.*
5. Podevin, A., *Accès aux formules mathématiques par des personnes non voyantes : étude et définition d'une méthode adaptée*. 2002, Université de CAEN.
6. Ferreira, H. and D. Freitas, *Enhancing the Accessibility of Mathematics for Blind People: The AudioMath Project*, in *9th International Conference on Computer Helping People with Special Needs (ICCHP)*. 2004, Springer Lecture Notes in Computer Science (LNCS): Paris, France. p. 678-685.
7. Ferreira, H. and D. Freitas, *AudioMath: Towards Automatic Readings of Mathematical Expressions*, in *Human-Computer Interaction International (HCII)*. 2005: Las Vegas, Nevada, USA.
8. Raman, T.V., *Audio System for Technical Readings*. Vol. 1410. 1998, Berlin, Germany: Springer-Verlag.
9. Moço, V. and D. Archabault, *VICKIE: A Transcription Tool for Mathematical Braille*, in *7th European Conference for the Advancement of Assistive Technology in Europe (AATE)*. 2003, IOS press: Dublin, Ireland.

10. Schwebel, F. and R. Goiffon. *BraMaNet: Quelques règles simples à connaître pour qu'un aveugle puisse lire vos documents mathématiques et vos pages web*. in *Journées nationales Caen*. 2005. Caen, France.
11. Garlini, P. and F. Fogarolo, *LAMBDA: Linear Access to Mathematics for Braille Device and Audio Synthesis - Analysis of the User's Needs*. 2003, University of Padova: Padova, Italy.
12. Lécuyer, A., et al. *HOMERE: a Multimodal System for Visually Impaired People to Explore Virtual Environments*. in *Proceedings of the IEEE Virtual Reality* 2003. Washington, USA: IEEE Computer Society.
13. Wooldridge, M., *An Introduction to Multiagent Systems*. 2002, Chichester, England: Wiley.
14. Weiss, G., *Multiagent systems.* MIT-Press, 1999.
15. Ferber, J., *Les systemes multi-agents*, ed. V.u.i. collective. 1995, Paris: InterEditions.
16. Jennings, N.R. and M.J. Wooldridge, *Applications of Intelligent Agents*, in *Agent Technology: Foundations, Applications, and Markets*, N.R. Jennings and M.J. Wooldridge, Editors. 1998, Springer-Verlag: Heidelberg, Germany. p. 3-28.
17. Bourbakis, N.G. and D. Kavraki. *An Intelligent Assistant for Navigation of Visually Impaired People*. in *the 2nd IEEE International Symposium on Bioinformatics and Bioengineering Conference*. 2001.
18. Awde, A., C. Tadj, and Y. Bellik, *Un système multi-agent pour la présentation d'expressions mathématiques à des utilisateurs non-voyants*, in *21ième Conférence Canadienne de génie électrique et génie informatique*. 2008, IEEE Canada: Niagara Falls, Ontario, Canada.
19. Bellik, Y., *Interfaces multimodales : concepts, modèles et architectures.*, in *LIMSI*. 1995, Université de Paris-Sud XI Orsay: Paris.
20. Awde, A., et al., *An Adaptive Multimodal Multimedia Computing System for Presentation of Mathematical Expressions to Visually- Impaired Users*, in *Journal of Multimedia (JMM)*. to be published.
21. Awde, A., Y. Bellik, and C. Tadj, *Complexity of Mathematical Expressions in Adaptive Multimodal Multimedia System Ensuring Access to Mathematics for Visually Impaired Users.* International Journal of Computer and Information Science and Engineering, 2008. **2**(2): p. 103-115.
22. Kramer, J. and J. Magee, *Analysing Dynamic Change in Software Architectures: A Case Study*, in *Proceedings of the International Conference on Configurable Distributed Systems* 1998, IEEE Computer Society: Washington, DC, USA.
23. Kumar, S. and P.R. Cohen. *Towards a Fault-Tolerant Multi-Agent System Architecture*. in *The fourth international conference on Autonomous agents* 2000: ACM Press.
24. Lightfoot, D., *Formal Specification Using Z*. 2nd ed. 2001: McMillan Press.
25. Bellifemine, F.L., G. Caire, and D. Greenwood, *Developing Multi-Agent Systems with JADE*. 2007: Wiley. 300.

Ali AWDE is currently a PhD student at the École de technologie supérieure (ÉTS). He obtained his master degree in Computer Science in 2003 from Université de Montréal. His research interests include multimodal multimedia for visually-impaired users, machine learning, and mathematics for the blind. Email: ali.awde.1@ens.etsmtl.ca.

Chakib TADJ is a professor at ETS, Canada. He received his PhD degree from ENST Paris in 1995. His main research interests are automatic speech recognition, human-machine interface, multimodal and neuronal systems. Email: *ctadj@ele.etsmtl.ca*.

Yacine BELLIK is an assistant professor at LIMSI-CNRS (*Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur*). He holds a PhD and HDR in Computer Science. His research interests concern multimodal human-computer interaction, aid for the blind and ambient intelligence. Email: *Yacine.Bellik@limsi.fr*.

Collaborative Testing Based on BPEL and CPP/CPA

Wenli Dong
Institute of Software, The Chinese Academy of Sciences
Beijing, China

Abstract—**This paper proposes a collaborative test framework, and analyzes new requirements arising from collaborative process sped by BPEL and CPP/CPA based on the two phases: collaboration modeling and collaboration execution. Combing the dynamical and real time of collaborative process of service, a collaborative test environment is provided. Under this environment, the collaborative process is transformed to HPN and agents are represented by HPN. With the help of HPN, it is easy to re-compose the Web Service and select the suitable agents to carry out the test task. At last, some techniques applying in collaborative testing including distributed treatment, sharing document organization, and test path selection are presented.**

Keywords-Collaborative Testing; Business Process Execution Language; Collaboration Protocol Profile Collaboration Protocol Agreement

I. Introduction

The interoperability between Web Services has been one of the most important research topics on the SOA field with mounting economic and technical challenges as growing complexity and increased services. Recently, the interoperability between Web Services started to get attention: collaborative process. Business Process Execution Language (BPEL) [1] allows specifying business processes and how they relate to Web services. This includes specifying how a business process makes use of Web services to achieve its goal, as well as specifying Web services that are provided by a business process. And ebXML CPP/CPA (Collaboration Protocol Profile/Collaboration Protocol Agreement) [2] proposed by OASIS removes barriers to entry and eliminates the high cost and complexity of trading partner on-boarding.

Up to now, the researches on Web Service and ebXML focus on implementing business collaboration [3,4]. Some researches attempt to test Web Service by studying the benchmark [5,6], fault analysis [7,8], and model checking [9]. The collaborative testing isn't involved in these researches. Testing collaborative process to gain confidence in its conformance to the desired function with expected QoS is a key problem certainly, because lack of trust will prevent collaborative process from adopting. But, because of its properties, collaborative testing is complex, difficult and time-consuming, which makes collaborative testing the challenges such as easing communication, coordination and sharing test data in a distributed, dynamic, and heterogeneous environment. Driven by increasingly complex of collaborative process, automatic collaborative testing to ensure the quality of collaborative process is required by both the provider and requestor. This paper discusses the relevant issues arising from the design and implementation of the infrastructure to support collaborative testing in a distributed, dynamic, and heterogeneous environment.

The paper is organized as follows. In section 2 new requirements for testing collaborative process are detailed and collaborative testing is analyzed. The collaborative test environment adopted in collaborative testing is discussed in section 3. The architecture of the collaborative test environment is given. Collaborative process is transforming to HPN. All test agents are described and classified in this section. The way for agent binding dynamically is introduced, which is implemented based on HPN. Section 4 gives some techniques applying in collaborative testing including distributed treatment, sharing document organization, and test path selection. Section 5 is the conclusion and future work.

II. New Requirements for Collaborative Testing

Traditionally, various tests of Web Service have remained isolated. However, the predicted collaboration in Web Services is creating new requirements for collaborative testing:

1) dynamic and real time
2) dependent
3) concurrent
4) distributed
5) use of parallel testing increasingly
6) test design and test scheduling analysis
7) test resources reusability

Combining the characteristics of Web Service process, the new requirements are detailed as following.

Collaborative testing can increase our confidence in the collaborative services. Based on our experience, collaborative testing is used to aid popularization and application of collaborative process by iterative applying the technique in the collaborative process life cycle: collaboration modeling, collaboration establishment, collaboration execution, and collaboration termination. Because the collaboration establishment and collaboration termination are mid processes and should not affect the quality of collaborative process for requestor. The collaborative testing just considers the two phases of collaborative process: collaboration modeling and collaboration execution.

Table 1 lists the new collaborative test requirements in collaboration modeling and collaboration execution phases.

TABLE I. New Requirements for Collaborative Testing

Collaborative process	Description	What need do to test
Collaboration modeling	Roles assignment used by many providers is usually computed in real	In a dynamical and competitive environment, collision between services

	time. The collaboration between roles is achieved through the introduction of specifications.	should be tested, and Schema of data type description validation and efficiency of role assignment should be ranked.
Collaboration execution	Tailoring execution to the predicated sequence specification. Process flow is carried out dynamically, distributed, and step by step.	Flexible and composite software environment is required to host and/or integrate a diversity of tools for various platform and language; Software tool can bridge the gap between dynamic testing and static analysis is demanded.

Collaboration modeling phase

In anticipation of the discovery of new business opportunities or threats, the SOA architectural style aims to provide enterprise business solutions that can extend or change on demand. SOA solutions are composed of reusable services, with well-defined, published and standards-compliant interfaces. SOA provides a mechanism for integrating existing legacy applications regardless of their platforms or languages.

The main first-class constructs in an SOA are service components such as operations, services, and business process.

BPEL and WSDL provide standard forms for service interaction and composition. It allows you to create complex processes by creating and wiring together different activities that can, for example, perform Web services invocations, manipulate data, throw faults, or terminate a process. These activities may be nested within structured activities that define how they may be run, such as in sequence, or in parallel, or depending on certain conditions. While BPEL describes the process, WSDL document describes the interface of the process that will be presented. This includes the Schema description of all element, attributes, and message.

What' more, the non functional requirements such as security details should be taken into account in the collaborative process based on CPP/CPA, and XML-based policy can be added to collaborative process to describe it's non functional requirements.

The testing for role assignment is necessary to verifying collision of operators being assigned such as a role to send/receive command. However, to assign an operator, the operation role simply based on its implementation is not sufficient for a dynamic and concurrent collaborative process. In such environment, the state of each role changes quickly as the collaborative process goes on. Collisions against opponents must be avoided. Also, the main objective of the collaborative process is the efficiency e.g. cost; and if a service is in a better position to carry out an operation, it should be given the opportunity. A more efficient schema of role assignment is necessary, and efficient testing should be assigned.

The testing for role assignment algorithm can be implemented using logic. Parameters used as input to the logic for each service are derived from WSDL specification. Logic rule based reasoning is used to decide the efficiency of the collaboration modeling. Some function can be used to analyze the collaboration modeling quantitatively.

Collaboration execution phase

In the collaboration execution, the candidate Web Services have been configured in configured files, and this configure file can be changed if:

1) One or more services in the collaborative process are unavailable

2) One or more services in the collaborative process are inefficient or invalid

The change describes above is called re-configure/re-composition, the re-configure/re-composition is dynamical; to meet the dynamic characteristics of SOA technique, and most tasks will be done on the fly at runtime in a collaborative manner. Without access complete information (i.e. detail development information), the composition can be performed at runtime. If the process invokes another Web service (i.e. if the process contains an invoke activity), then create/obtain the WSDL document(s) that describe the service which is to be invoked. These WSDL documents must have bindings and endpoint information that describe where and how the service may be invoked. The engine supports SOAP, EJB, and direct Java class bindings.

Collaboration deployment is based on the reliability of selected service, after deployment, the CPA between collaboration partner also is formed based on the CPP of partners. If the process that you are deploying invokes external services you will be presented with an interactive interface that enables you to perform re-composition/re-configuration for the process. For each partner that was specified within the process, you will have the opportunity to notify the engineer who will be fulfilling that role by providing it with a WSDL file that describes the partner. For each displayed partner, you need to specify the name of the WSDL file that describes the partner.

If collaborative process was successful deployed, the engineer will remember Process ID (QName), provide you with a link to the WSDL file that describes how to access the process, and the information necessary will be given to use the process in a client if you don't want to read the WSDL document.

Reuse and re-configuration/re-composition are common in collaborative process. The collaborative process use pre-designed replaceable components, which makes collaborative testing challenges. The participant in collaborative process is changed and the relationship between the services is complex, which requires cooperative among multiple domains in the collaborative testing. This requires implementing automatically collaborative testing that predictably satisfied all design requirements quickly and avoids repetitive testing, thus need test sharing and test path selection.

From above discussion, collaborative testing can be analyzed as following:

1) The specifications play important roles in collaborative testing. Collaborative process is based on BPEL, CPA/CPP, and WSDL. While a BPEL business process interoperates with the Web services of its partners,

whether or not these Web services are implemented based on BPEL, CPP/CPA gives a future-proofed, tested and technically proficient standard. And the interface information of services is specified in WSDL. The specification above provides the service-oriented high-lever view of the web service system, which dictated method of test case generation. And test data can be derived based on these specifications [10]. The specification serves as a reference model in analyzing the collaborative process. Here:

a) WSDL offering a verbose, ASCII, standard, and language-agnostic view of services offered to clients. WSDL also provides noninvasive future-proofing for existing applications and services and allows interoperability across the various programming paradigms.

b) BPEL defining an interoperable integration model that should facilitate the expansion of automated process integration in both the intra-corporate and the business-to-business spaces.

c) CPP describing the e-business services system, listing the business flow, information and data exchanging technology. Based on CPP, CPA provides collaboration protocol agreement for collaboration partner. The non functional requirements described by CPP/CPA can be transformed to XML_based policy.

2) High-level Petri nets are a widely used modeling and specification language for information system behavior since it has the ability to model concurrency of the systems, analyze concurrent behavior, and express the dynamically changed software. The related research and tool for Petri net is popular and rather complete. They combine the advantages of a simple graphical notation with mathematical foundation. Moreover, they allow modeling the behavior of an information system and its data structures in one integrated scheme [11]. Petri nets are directly executable by a net interpreter. Thus, Petri net simulation may be used for collaborative testing.

III. COLLABORATIVE TEST ENVIRONMENT

In collaborative testing, test task is accomplished by agents as shown in Fig. 1. To meet these requirements described in section 2, based on semantic Web and HPN (High-level Petri Net), by allowing software agents to communicate and understand the information published on the Internet, a collaborative test environment is proposed. First of all, HPN Model checking tool is employed in collaboration modeling, and in collaboration execution such as generating test case that exacting information from HPN and select the suitable agents to carry out the test task can be implemented with the help of HPN. Secondly, the agent in runtime environment provides a dynamic mechanism for service description, invocation, discovery, and composition dynamically. Finally, the semantic vision is to allow communications from software agent to software agent. Under our test environment, various software agents decompose test task into small subtasks and carry out these tasks. They cooperate with each other to fulfill the whole test task. The collaborative test environment proposed in this paper is composite and extended, and agents can dynamically join and leave the system to achieve the maximum flexibility and extendibility. Combining with the policy, selecting an operation to be invoked is done with the evaluation of "Guards" attached to arcs denoting operation invocations.

As shown in Fig. 1, the test environment consists of a number of agents to fulfill test tasks for collaborative testing. Organizing relative agents into group is convenient for controlling agents [12,13,14]. As shown in Fig. 1, the Get BPEL (GB) agents and the BPEL Analysis (BA) agents are arranged in media agent group for their interaction with BPEL/WSDL specification and test environment. However, these agents can be distributed in different computers. Actually, the distribution of agents is free according to any specific configuration, can move and change their location at runtime.

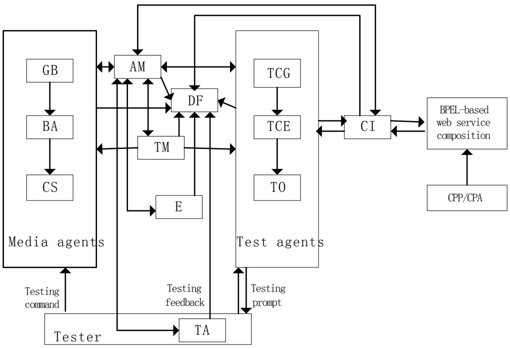


Fig. 1. Agents for testing collaborative process

Get BPEL (GB) agents obtain BPEL/WSDL specification from Test Assistant Agent.

BPEL Analysis (BA) agents analyze the BPEL/WSDL specification, exact useful information, and construct HPNs combining non-functional information from CPP for basic activity on a HPN platform. The structure information of basic activity in HPN form is stored in Knowledge Base (KB).

Composition Structure (CS) agents analyze the structure of BPEL-based collaborative process. Combining non-functional information from CPP generate a HPN presentation to describe the structure.

Test Case Generator (TCG) agents generate test cases to test an activity according to certain test criteria.

Test Case Execution (TCE) agents execute the test cases, and generate execution results.

Coordinate Interface (CI) agents provide flexibility for every kind of Web Service implementation.

Test Oracles (TO) agents verify whether the test results match the given BPEL/WSDL.

Test Assistance (TA) agents provide the interface between tester and computer that guides testers in the process of test.

Test Monitor (TM) agents monitor the test process at runtime and store the monitoring information in KB for later analysis.

Evaluation (E) agents are used to collect monitoring information and log information, employ predefined evaluation model, so that appropriate conclusions about the quality of the collaborative process can be drawn. Agent Management (AM) agents are agents in charge of managing agents, including agent description, agent arrangement. In some degree, it is corresponding to AM (Agent Management System) [15]. An agent management agent has the following capabilities:

1) Creating an agent description. It is based on XML. The description consists of agent name, agent locator, and agent properties.

2) Processing a given query for agents. The query [16] condition can be composite to find suitable agent exactly.

Directory Facilitator (DF) agent is mainly responsible for registering agent to make agent in a test environment visible to AM.

The message mechanism consists of a set of communication primitives such as send, receive message that is passed between agents. Its design objects are applicable, flexible, lightweight, and simple.

As a loose couple distributed communication way, message queue is independent of hardware and operation system, and can ensure data not being lost and copied. It provides an effective communication mechanism for agent in our prototype system.

A. Transforming Collaborative Process to HPN

Fig. 2 illustrates the HPNs for an operation. A part will be presented by a place with token whose type specified by the part type used as the interface of test case generation. An arc is used to link the transition with another arc linking to the input/output message consists of those parts based on the relationship defined in Schema [11]. With multiple input/output, there will be a transition for each input/output message. The physical preconditions described in BPEL are embodied in the HPNs by places. And the cause-effect analysis is adopted in this mapping.

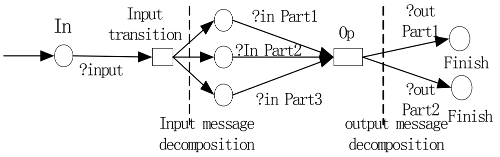


Fig. 2. Atomic operation HPN

The operation cluster is generated by basic activity receive, reply, assign, invoke, empty, terminate, and wait. Operation cluster are generally described in procedural style. Thus we can make a corresponding action or statement to a transition. Places connected to the transition intuitively expresses the states before and after executing the corresponding action or statement. Firing a transition means that the corresponding action is being executed. Operation invocation can be expressed by entering a token in a place which denotes the starting point of the operation. Basic activities translation is shown in Fig. 3.

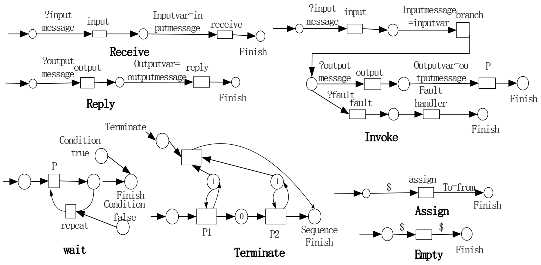


Fig. 3. Basic activity HPN

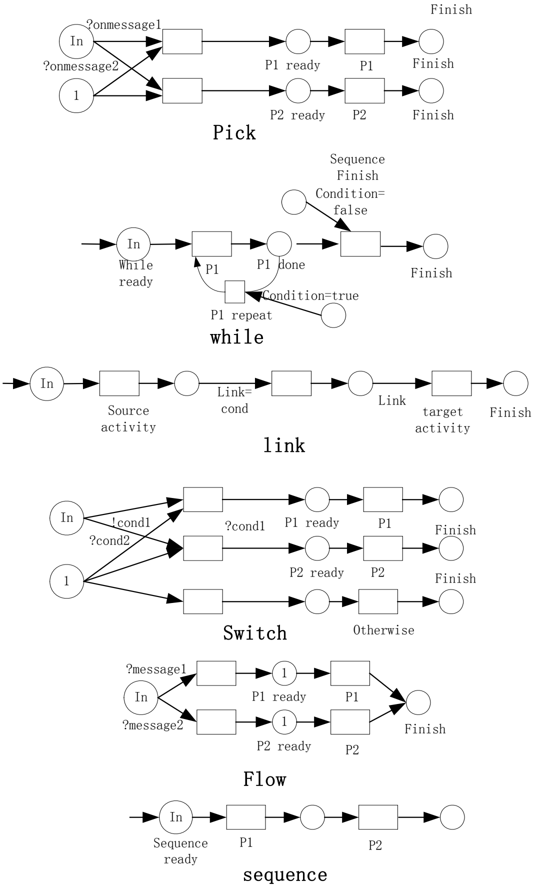


Fig. 4. Structured activity HPN

The operation invocation sequences at service level can be loop, choice, link, parallel, and sequence, presented respectively by structured activities: while, pick/switch, sequence, link, and flow activity in BPEL. The transition of activities at service level implies other problems, e.g., dynamic binding and concurrency. We cannot decide statically which operation occurs. Which operation is really executed is decided during execution because of re-composition/re-configuration of BPEL-based collaborative process. To compute which action will be invoked, we attach the information of the action name to a token and the condition judgment on arc based on global parameters and token value denoting the action selection. Identifying information will be attached to token in places which can be implemented in HPNs. Selecting a operation to be invoked is done with the evaluation of "Guards" attached to arcs denoting operation invocations. Fig. 4 illustrates above translation.

In Web Service composition, an operation in one service may have the same name with another one in another service. E.g., as shown in Fig. 5, the operation "findaddress" in "customer" Web Service is defined to contain two output messages: department and tel, while in "personnel" Web Service, it is defined to contain three output messages: name, city, and street. It means the use of operation "findaddress" in a Web Service composition can cause confusion. In collaborative testing using HPN, this problem can be resolved by presenting the attribute value of a corresponding token with the namespace of the invoked service for identifying the corresponding operation.

```
customer.wsdl:
...
<operation name="findaddress">
<input message="tns:name"/>
<output message="tns:department"/
>
<output message="tns:tel"/>
</operation>
...
personnel.wsdl
...
<operation name="findaddress">
<input message="tns:name"/>
<output message="tns:name"/>
<output message="tns:city"/>
<output message="tns:street"/>
</operation>
    ...
```

Fig. 5. Operations with same name

B. Representing Agents in HPN

Referencing the soft gene definition stated in [17]: a soft gene is an entity consisting of a set of behaviors and attributes. In a HPN, a behavior can be represented by a transition and the attributes can be represented by predicate properties. We define the behavior and attributes:

```
// predicate definition
struct pred { //attribute definition}
 ram <pred> predicate-name
//transition definition
trans name {  //  declarations
               // arcs with the fire rule of transition
              action { // code to evaluate at fire start. }   }
```

Under collaborative test environment, the agent takes part in test or not is decided at runtime. AM agents arrange the work for all available agents, and the state of agent is recorded by AM agents. Under thus open and dynamic environment, agents change state dynamically and interact with other agents. An agent can be defined as an entity with a set of soft gene. At a specified time, the steps of the agent taking are decided by its state and properties at this time. Based on above analysis, we can represent agents in HPN. The agent name is defined in predicate properties that denote the agent to take charge of transition, and bind/discard agent name with a concrete agent is dynamically. AM agents organize the interactions of all available agents.

Assuming P is the set of predicates; A is the set of agents; bind represents the relationship between a predicate and an agent. That is:

$A = \{a_{0,} a_{1,} \dots, a_n\} \quad (n \geq 0)$

$P = \{p_{0,} p_{1,} \dots, P_m\} \quad (m \geq 0)$

bind: A╳P→{0,1}

The value of bind is 1 means at predicate P_x, agent a_y is in the state specified by P_x.

Two types of operations are added in a HPN platform. One is bind (P_x,a_y), that means building relation between P_x and a_y. The other is release (P_x,a_y), that means discarding the relationship between P_x and a_y.

For each agent a_k, there is a set of behaviors represented by transition and pointed by arcs. Every arc will be labeled with the predicate information including agent name, input etc. the behaviors for agent can be represented as $T_s = \{t_{0,} t_{1,} \dots, t_o\}$, and the arcs that fire each transition t_q can be represented as $ARC_{t_q} = \{arc_0, arc_1, \dots, arc_w\}$. A simple sketch is shown in Fig. 6.

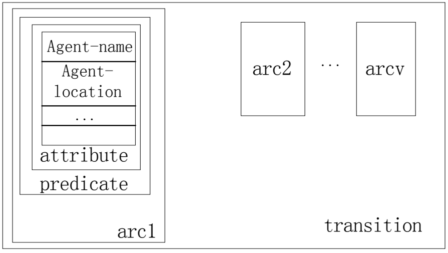


Fig. 6. Binding agents in HPN

IV. Techniques in Collaborative Testing

The techniques adopted in our system mainly include distributed treatment, sharing document organization, and test path selection.

A. Distributed Treatment in Collaborative Test Environment

Distributed testing means test manager invokes test agent to carry out target service.

In distributed testing, the function of test manager (agent AM) is consisted of:

1) At the beginning of the testing, all the agents will be created. Next, the agent description will be recorded by AM.

2) Under the control of AM, based on selection strategy, TA is selected by querying DF, so does GB, BA,

CS, TCG, CI, TO, and TM. And sending test script to test agent

3) According to collaborative process, AM will send query request to DF, DF will find a suitable TCE and reply the request with the agent information

4) The test executive sequence is controlled by AM after it stored the structure information in the form of HPN.

5) When AM finds the testing has been finished, the AM will send query request to find a suitable Evaluation agent to draw a conclusion for the quality of the tested collaborative process.

And test agent (agent GB, BA, CS, TCG, CI, TO, and TM) mainly need:

1) Receiving test script from AM

2) Carrying out remote testing as simultaneous client

3) Collecting test result, evaluating testing, and stored in KB.

Communication between agents and dispatching test task in remote testing based FIPA (Foundations of Intelligent Physical Agents) [15] and semantic information are provided in testing.

B. *Sharing Document Organization*

The file sharing and storing management are important because the interoperability between operations, services, and agents in collaborative testing is frequent. In our collaborative testing:

1) XML_based files are employed because of its flexibility, extensibility etc. And this will benefit the interaction between operations, services, and agents in collaborative testing.

2)Test case directory is designed and built according to service name, port/interface name, and operation name level by level. The different test case names for an operation will be composed of the numbers starting from ‘1’ with the operation name as prefix. Thus, the test case can be easily used and understood without confusing with other operations including those with the same name but different namespaces.

3) Common test data file is shared by all collaborative test case generation and provides the effective value, min value, max value of simple type. So, common test data file is located in root directory that will be found and visited conveniently.

4) After BPEL, WSDL, and CPP/CPA specification are input, the property configure files and policy file are generated combining the initial value defined by common test data files. Various date assigned for input/output message is stored in corresponding files. So the property configure files are build according to service name, port/interface name, and operation name based on the operation level in order to meet test data manage requirements of different services, ports and operations.

5) Excepting the common test data file, the data in a collaborative test is organized in an independent directory called a project, so that the collaborative test can be managed easily as a whole.

C. *Test Path Selection*

In collaborative testing, test path selection is emphasized because of the complexity of teat case combination in collaborative testing [18]. A traditional symbolic executor makes use of execution trees to maintain the information at each decision point. The data structure of the execution tree will store the values of all variables in the branch of execution tree and help the executor keep track of all possible paths. Saving and loading such a tree requires much time and space in a single process. In collaborative testing, tracing each path instead of keeping an overall tree is used. Whenever a decision point is encountered, such as if-then-else, a new child process is forked. The parent process and the child process will explore these two cases respectively. The process that contains a path found to be in conflict with the test case is terminated or suspended.

After obtaining equations in symbolic form, we can solve the equations in order to obtain the test data from solutions. However, since solving the general equations (e.g. datatype) is un-decidable, we will derive approximate solutions, although with limited acceptance. In fact, Different strategies can be used to generate data values such as boundary value, random value, equivalent class, etc. Based on the dependent analysis, various coverage metrics can be used to generate test paths.

V. Conclusion and Future Work

The new requirements for collaborative are analyzed according to the two process phases: collaboration modeling and collaboration execution. Based on the characteristics of collaborative process, a collaborative test environment combined with HPN is proposed to satisfy the distributed, dynamic, and re-composition/re-configuration requirements of collaborative process based on BEPL and CPP/CPA. By representing agent and Web Service in HPN, the test can be executed conveniently at runtime. It is easy to test the collaborative process and select the suitable agents to carry out the test task with the help of HPN. Some techniques applying in collaborative testing including distributed treatment, sharing document organization, and test path selection are analyzed in this paper.

Based on the collaborative testing proposed by this paper, the prototype system have been designed, implemented, and applied in collaborative process testing successfully. Thus, the collaborative testing discussed above has been proved to be practicable.

Our future work includes perfecting the ontology of this collaborative combing the OWL-S and the complex application of this collaborative testing to verify the reliability of the test environment. At the same time, the security of this collaborative testing is also within our further researches because of its distribution.

Acknowledgment

This work is supported by the National High-Tech Research and Development Plan of China under Grant No. 2007AA01Z190.

REFERENCES

[1] Web Services Business Process Execution Language Version 2.0. http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html. 23th August, 2006.

[2] Collaboration-Protocol Profile and Agreement Specification Version 2.0. OASIS ebXML Collaboration Protocol Profile and Agreement Technical Committee, September 23, 2002.

[3] Ja-Hee Kim, Christian Huemer. From an ebXML BPSS Choreography to a BPEL-based Implementation. ACM SIGecom Exchanges. 2004, 5(2), pp.1-11.

[4] Nathaniel Love, Michael Genesereth. Computational Law. Proceedings of the 10th international conference on Artificial intelligence and law. 2005, pp. 205-209.

[5] Michael R. Head, Madhusudhan Govindaraju, Aleksander Slominski, Pu Liu, Nayef Abu-Ghazaleh, Robert van Engelen, Kenneth Chiu, Michael J. Lewis . A Benchmark Suite for SOAP-based Communication in Grid Web Services. Conference on High Performance Networking and Computing. 2005, pp.19-31.

[6] Liming Zhu, Ian Gorton, Yan Liu, Ngoc Bao Bui. Model driven benchmark generation for web services. International Conference on Software Engineering. 2006, pp.33-39.

[7] Liliana Ardissono, Luca Console, Anna Goy, Giovanna Petrone, Claudia Picardi, Marino Segnan, Daniele Theseider Dupré. Advanced fault analysis in web service composition. International World Wide Web Conference. 2005, pp.1090-1091.

[8] Wang, J.Z. Taylor, W. Clemson Univ., Clemson. A Lightweight Fault Tolerance Framework for Web Services. Web Intelligence. 2007, pp.542-548.

[9] Avik Sinha, Amit M. Paradkar. Model-based functional conformance testing of web services operating on persistent data. International Symposium on Software Testing and Analysis. 2006, pp. 17-22

[10] Xiaofei Sun, Shengxian Luo. Research on Technique of Automated Test Case Generation Based on Rule Engine. 2008,1(3),pp.468-471.

[11] DONG Wenli, MENG Luoming. tML Schema Based ICS Proforma and Generation Method. Chinese Journal of Electronics, 2005, 14(4), pp. 681-685.

[12] LIU Ying-qiao, ZHAO Zheng-de et al. Web Service Based Multi-agent Cooperative Platform. Computer Engineering and Design, 2003, 39(21), pp. 1269-1271.

[13] Qingning Huo, Hong Zhu and Sue Greenwood. A Multi-Agent Software Environment for Testing Web-based Applications. COMPSAC, Dallas, USA, November 2003, pp.210-215.

[14] Cecile Aberg, Patrick Lambrix, Nahid Shahmehri. An Agent-based Framework for Integrating Workflows and Web Services. WETICE, Linköping, Sweden, June 2005, pp. 27-32.

[15] FIPA Abstract Architecture Specification. http://www.fipa.org/.

[16] XQuery 1.0: An XML Query Language W3C Candidate Recommendation, November 2005, http://www.w3.org/TR/xquery/.

[17] Q. Yan, X. Mao, L. Shan, Z. Qi, and H. Zhu. Soft Gene, Role, Agent: MABS Learns from Sociology. An IEEE/WIC Conference on Web and Agent Untelligence, October 2003, pp. 450-453.

[18] Wee Kheng Leow, Siau Cheng Khoo, and Yi Sun. Automated Generation of Test Programs From Closed Specifications of Classes and Test Cases. International Conference on Software Engineering. 2004, pp.96-105.

Anonymous signature scheme

Chunbo Ma and Jun Ao

***Abstract*—In order to hide the identity of a signer, an anonymous signature scheme is presented in this paper. In this scheme, a signer located in a specified group produces a signature on behalf of the group. The recipient can verify whether the signature is valid and comes from the specified group, while tracing the signature to its source is impossible. The proposed anonymous signature is similarly to ring signature in some respects, for example, there is no manager, and no revocation mechanism against signer's anonymity. The most different between these two kinds of signatures is that the group in ring signature is adaptively constructed by the signer, while the group in our scheme is fixed.**

***Index Terms*—Anonymity, Signature, Public key, Group**

I. Introduction

DIGITAL signature is one of the crucial primitives in public key cryptography. It has been widely used in providing authenticity, integrity and non-repudiation. However, in many scenarios such as e-voting, e-auction, and many others, we need to protect a signer's identity from being arrested by malicious attackers. Currently, some signatures are designed to conceal the real identity of a signer. Ring-based signature and group signature are of the important two kinds of signatures which provide anonymity for the signer.

The concept of a group signature scheme introduced in 1991 by Chaum and van Heyst [1] is a well studied subject in cryptography. In such a scheme, a trusted group manager distributes specially designed keys to their members. Individual members can then use these keys to anonymously sign messages on behalf of their group. From view of the point of a verifier, the signature produced by different group members look indistinguishable, while the group manager can revoke the anonymity of misbehaving signers.

The concept of ring signatures was formally introduced in [2], and can be considered as a simplified group signature. After that, many proposals of ring signature schemes have been published [3][4][5][6]. In a ring signature scheme, an entity signs a message on behalf of a set of members. The verifier of the signature is convinced that it was produced by some member of the ring, but he does not obtain any information about which member of the ring actually signed. Ring signatures are a useful tool to provide anonymity in the scenarios that a member of a group has to leak some messages on behalf of the group while does not want to open his identity. There are some other works on anonymous signature such as [7][8].

Obviously, the most distinguish different between these two kinds of signatures is that a trusted manager is existed in group signature while it is not in ring signature. The role of the manager is a combiner, and when necessary it can act as an arbiter. The common ground of the two kinds of signatures is that they all provide anonymity for the signer. In some special scenario such as in a fixed group, in which the key of each member is well designed, providing the anonymity for a signer is much easier.

Ma et al. [9] presented a group-based encryption scheme, in which each member of the specified group has ability to decrypt a ciphertext encrypted for the group. In this paper, we present an anonymous signature from this encryption scheme. From view of the point of a verifier, the signature just comes from the specified group, and can't be traced. Similarly to the ring signature, there is no group manager in the group. What the different from ring signature is that the group is predefined and there are some underling relationship among group members.

II. Related Works

The original group signature scheme that first proposed by Chaum and Heyst [1] is linear to the size of the group. Currently, many improved schemes have been proposed to achieve constant signature size, i.e. the signature size is independent of the size of the group. Camenisch and Groth [11] presented an efficient scheme that is secure under strong RSA assumption and the Diffie-Hellman decision assumption. Boneh et al. [12] proposed another efficient group signature scheme under strong Diffie-Hellman and linear assumption. The signature produced in this scheme is under 200 bytes, while provides the same security level as an RSA signature of the same length.

The NS [13] achieves anonymity, but compare to BBS, its computation cost and the size of group signature are larger. Furthermore, although NS claims to be secure in BSZ security model, there are some flaws in the proof. Its security needs in-depth research.

The notion of ring signatures was first introduced in 2001 by Rivets et al. [2], after following lots of related ring signature scheme. In 2002, Abe et al. [14] proposed how to use public-keys of several different signature schemes to generate

This work is supported by the National Natural Science Foundation of China (60862001) and the Foundation of the Education Department of Guangxi Zhuang Autonomous Region of China (200808MS004).

The authors are with the School of Information and Communication, Guilin University of Electronic Technology, Guilin Guangxi, 541004, P. R. China.

1-out-of-n signatures. Another interesting work based on bilinear pairings and identity-based cryptography [15] was presented by Zhang and Kim.

Lee et al. [16] proposed a convertible ring signature the can withdraw the anonymity in 2005. Nguyen [17] designed a dynamic accumulator based on bilinear pairings, and presented an ID-based Ad-hoc anonymous identification scheme. He pointed out that applying the Fiat-Shamir heuristics to the ID-based Ad-hoc anonymous identification scheme results in an ID-based ring signature scheme with constant-size signatures. There are some other researches on ring signature [18] [19]. Chen et al. [20] extended the existing notion of ring signatures, and proposed the concept of identity-based anonymous designated ring signature which can be used in a Peer-to- Peer (P2P) network.

III. Background

A. Bilinear Maps

Let G_1 be a cyclic multiplicative group generated by g, whose order is a prime q and G_2 be a cyclic multiplicative group of the same order q. Assume that the discrete logarithm in both G_1 and G_2 is intractable. A bilinear pairing is a map $e: G_1 \times G_1 \rightarrow G_2$ and satisfies the following properties:

1. *Bilinear:* $e(g^a, p^b) = e(g, p)^{ab}$. For all g, $p \in G_1$ and $a, b \in \mathbb{Z}_q$, the equation holds.
2. *Non-degenerate:* There exists $p \in G_1$, if $e(g, p) = 1$, then $g = \mathrm{O}$.
3. *Computable:* For g, $p \in G_1$, there is an efficient algorithm to compute $e(g, p)$.
4. *Commutativity:* $e(g^a, p^b) = e(g^b, p^a)$. For all g, $p \in G_1$ and $a, b \in \mathbb{Z}_q$, the equation holds.

Typically, the map e will be derived from either the Weil or Tate pairing on an elliptic curve over a finite field. Pairings and other parameters should be selected for efficiency and security.

B. Complexity Assumptions

— Computational Diffie-Hellman Assumption

Given g^a and g^b for some $a, b \in \mathbb{Z}_q^*$, compute $g^{ab} \in G_1$. A (τ, ε)-CDH attacker in G_1 is a probabilistic machine Ω running in time τ such that

$$Succ_{G_1}^{cdh}(\Omega) = \Pr[\Omega(g, g^a, g^b) = g^{ab}] \geq \varepsilon$$

where the probability is taken over the random values a and b. The CDH problem is (τ, ε)-intractable if there is no (τ, ε)-attacker in G_1. The CDH assumption states that it is the case for all polynomial τ and any non-negligible ε.

— k-Strong Diffie-Hellman (k-SDH) Assumption[10]

Given $\{g, g^x, g^{x^2}, \cdots, g^{x^k}\}$ for a random number $x \in \mathbb{Z}_q^*$, the attacker adaptively chooses random $c \in \mathbb{Z}_q^*$ and computes $g^{(c+x)^{-1}}$. A (τ, ε)-k-SDH attacker in G_1 is a probabilistic machine Ω running in time τ such that

$$Succ_{G_1}^{k-sdh}(\Omega) = \Pr[\Omega(g, g^x, g^{x^2}, \cdots, g^{x^k}, c) = g^{(c+x)^{-1}}] \geq \varepsilon$$

We say the k-SDH problem is (τ, ε)-intractable if there is no (τ, ε)-attacker in G_1.

— T- Diffie-Hellman (TDH) Assumption

Given $\{g, g^a, g^{a^2}, \cdots g^{a^t}, g^{ak}, g^{a^2 k}, \cdots, g^{a^t k}\}$ for random numbers $a, k \in \mathbb{Z}_q^*$, compute $g^{a^{t+1}k} \cdot g^r$ and g^{ar}, where $r \in \mathbb{Z}_q^*$. A (τ, ε)-TDH attacker in G_1 is a probabilistic machine Ω running in time τ such that

$$Succ_{G_1}^{tdh}(\Omega) = \Pr[\Omega(g, g^a, g^{a^2}, \cdots g^{a^t}, g^{ak}, g^{a^2 k}, \cdots, g^{a^t k}) = (g^{a^{t+1}k} \cdot g^r, g^{ar})] \geq \varepsilon$$

We say the TDH problem is (τ, ε)-intractable if there is no (τ, ε)-attacker in G_1.

C. Security Notions

The accepted definition of security for signature schemes is *existential unforgeability under adaptive chosen message attack*, which is described in [21][22]. We say that a signature scheme is secure against an existential forgery under adaptive chosen messages attack in random oracle model if no polynomial bounded adversary has a non-negligible advantage in the following game:

1. **Setup:** the *Challenger* runs the **Initialize** algorithm and gives the system parameters to the *Attacker*.
2. **Attack phase:** the *Attacker* performs a polynomial bounded number of requests as follows.
   1). **H** queries: The *Attacker* queries the *Challenger* on a random chosen triple (m_i, R_{1i}, R_{2i}), and the *Challenger* responds with $\mathbf{H}(m_i, R_{1i}, R_{2i})$.
   2). **Sign** queries: The *Attacker* produces a query on m_i. The *Challenger* simulates **Sign** oracle and outputs $(m_i, U_{1i}, U_{2i}, V_{1i}, V_{2i})$ to the *Attacker* as the answer.
3. **Forge phase:** the *Attacker* gives a new signature (m, U_1, U_2, V_1, V_2) and wins the game if the signature can be verified correctly.

We define the advantage of the *Attacker* to be $Adv(Attack) = \Pr[Attack \quad WIN]$. We say that a signature is secure if no polynomial bounded *Attacker* has non-negligible advantage in the game described above.

IV. Signature Scheme

A. Initialize

Let G_1 be a cyclic multiplicative group generated by g, whose order is a prime q and G_2 be a cyclic multiplicative group of the same order q. A bilinear pairing is a map: $e: G_2 \times G_1 \to G_2$ that can be efficiently computed. Define one cryptographic hash function:

$$H: \{0,1\}^* \to Z_q$$

PKG chooses $a \in \mathbb{Z}_q^*$ uniformly at random, and computes $g_1 = g^a$. The master private key is a, and the master public keys are (g_1, g^{a^2}).

B. Key Generation

PKG chooses $k \in \mathbb{Z}_q^*$ uniformly at random as the tag of the group A. Using $PK_A = g^k$ as group A's public key. The member p_i 's private keys can be generated as follows:

1. PKG chooses $r_i \in \mathbb{Z}_q^*$ uniformly at random.
2. Compute and output $d_{i1} = g^{ar_i}$ and $d_{i2} = g^{ak} g^{r_i}$.

The member p_i 's private key is $d_i = \{d_{i1}, d_{i2}\}$.

C. Signature

Signer chooses a random number $t \in \mathbb{Z}_q^*$, and computes following two values

$$U_1 = g^{a^2 \cdot t} \qquad U_2 = g^{a \cdot r_i \cdot t}$$

We have $V_1 = d_{i2}^{(t+h)}$ and $V_2 = d_{i1}^{h}$, and the signature of m is (m, V_1, V_2, U_1, U_2), where $h = H(m, U_1, U_2)$. We say the signature is valid, since

$$e(V_1, g^a) = e((g^{ak} \cdot g^{r_i})^{(t+h)}, g^a)$$
$$= e(g^{ak(t+h)}, g^a) e(g^{r_i(t+h)}, g^a)$$
$$= e(g^{akt}, g^a) e(g^{akh}, g^a) e(g^{r_i t}, g^a) e(g^{r_i h}, g^a)$$
$$= e(g^k, g^{a^2 t}) e(g^{kh}, g^{a^2}) e(g^{ar_i t}, g) e(g^{ar_i h}, g)$$
$$= e(PK_A, U_1) e(PK_A^h, g^{a^2}) e(U_2, g) e(V_2, g)$$

Obviously, any recipient can verify the validity of the signature and accept that the signature comes from the specified group, however, he can't distinguish who signed the message since $t \in \mathbb{Z}_q^*$ is a random number.

V. Security

In this section, we will discuss the security of the proposed anonymous signature scheme. Firstly, we give following lemma.

Lemma. *Suppose the **CDH** assumption holds. Then given $g^b, g^{br_i} \in \mathrm{G}_1$, computing g^{r_i} is intractable.*

Proof. Assume that given $g^b, g^{br_i} \in \mathrm{G}_1$, the attacker Alice has ability to compute g^{r_i}. Then we can design an algorithm to solve ***k-SDH*** problem. In other words, given $g^m, g^{m^2} \in \mathrm{G}_1$, the challenger Bob can compute $g^{m^{-1}}$ by running Alice as a subroutine. Bob Inputs $g^{m^2}, g^m \in \mathrm{G}_1$ to Alice. As we have assumed above, Alice outputs g^{m/m^2} as a feedback. In other words, given $g^m, g^{m^2} \in \mathrm{G}_1$, Bob can solve ***k-SDH*** via Alice.

□

For a person Carol outside the group, she can't forge a valid signature without the help of malicious members. As we have mentioned above, since Carol can't forge a valid d_{c2}, she can't produce a valid V_1. We have following theorem.

Theorem. *If there exists an attacker Alice, who is allowed to request at most q_0 Hash queries and q_{d_s} signature queries, can break the proposed signature scheme with probability ε and within a time bound t, assume that $\varepsilon \ge 10(q_{d_s}+1)(q_{d_s}+q_0)/2^k$, then there exists another attacker Bob, who can solve **TDH** problem by recalling Alice as a subroutine in expected time $t' \le 120686 q_0 t / \varepsilon$.*

Proof. Assume that if the attacker Alice has ability to break the proposed signature scheme with non-negligible probability ε, then we will show how Bob can solve **TDH** problem. In other words, given $g^a, g^{a^{-1}k}, g^{a^2 k}, g^{a^2} \in G_1$, Bob can compute $g^{ak} \cdot g^r$ and g^{ar} with non-negligible probability by running Alice as a subroutine, where $g^r \in G_1$ and random number $r \in \mathbb{Z}_q^*$. The challenger Bob interacts with Alice by simulating **H** and **Sign** oracles.

Bob initializes the system and gives $g^a, g^{a^2} \in G_1$ as the public keys.

H hash queries. In this phase, attacker Alice is allowed to request at most q_0 hash queries. Bob maintains an empty Δ -list. For each query (m_i, R_{i1}, R_{i2}), Bob first checks the list:

1). If there exists an item $(m_i, R_{i1}, R_{i2}, h_i)$ in Δ list, then Bob return h_i to Alice.

2). If there is no such record in Δ list, i.e., the item (m_i, R_{i1}, R_{i2}) has not been queried to H oracle. Challenger Bob chooses a random number $h_i \in \mathbb{Z}_q^*$, and then preserves the item $(m_i, R_{i1}, R_{i2}, h_i)$ in Δ -list. Finally, he returns h_i to Alice as the answer.

Signature queries. In this phase, Alice is allowed to query at most q_{d_s} signature queries. For each query on m_i, Bob performs following step to return an answer.

1). Choose two random numbers $c_i, d_i \in \mathbb{Z}_q^*$, and then computes $U_{i1} = g^{kd_i}$ and $U_{i2} = g^{kc_i d_i - k^2 d_i}$.

2). Choose a random number $h_i \in \mathbb{Z}_q^*$, and then preserves $(m_i, U_{i1}, U_{i2}, h_i)$ in Δ -list.

3). Compute $V_{i1} = g^{a^{-1}kc_i d_i} g^{ac_i h_i}$ and $V_{i2} = g^{a^2 c_i h_i - a^2 k h_i}$, and then Bob returns $(m_i, U_{i1}, U_{i2}, V_{i1}, V_{i2})$ to Alice as the answer.

Actually, challenger Bob sets $g^{r_i} = g^{ac_i - ak}$ and $t = a^{-2}kd_i$ in above process. Then U_{i1} , U_{i2} , V_{i1} and V_{i2} can be expressed as follows.

$$U_{i1} = g^{a^2 t} = g^{a^2 a^{-2} kd_i} = g^{kd_i}$$

$$U_{i2} = g^{ar_i t} = g^{a(ac_i - ak)a^{-2}kd_i} = g^{kc_i d_i - k^2 d_i}$$

$$V_{i1} = (g^{ak} g^{r_i})^{(t+h)} = g^{ac_i(a^{-2}kd_i + h)} = g^{a^{-1}kc_i d_i} g^{ac_i h_i} = g^{a^{-1}kc_i d_i} g^{ac_i h_i}$$

$$V_{i2} = g^{ar_i h_i} = g^{a(ac_i - ak)h_i} = g^{(a^2 c_i - a^2 k)h_i} = g^{(a^2 c_i - a^2 k)h_i}$$

The simulation is perfect in the random oracle. After all the queries, Alice outputs a fresh signature $\sigma_0 = (m^*, U_{j1}, U_{j2}, V_{j1}, V_{j2})$, where warrant m^* has never been queried to the **Sign** oracle. According to the forking lemma [20][21], if $\varepsilon \geq 10(q_{d_s} + 1)(q_{d_s} + q_0)/2^k$, then Bob has ability to produce two valid signatures $\sigma_0 = (m^*, U_{j1}, U_{j2}, V_{j1}, V_{j2})$ and $\sigma_1 = (m^*, U_{j1}, U_{j2}, V'_{j1}, V'_{j2})$ on the same warrant m^* such that $H(m^*, U_{j1}, U_{j2}) \neq H'(m^*, U_{j1}, U_{j2})$. Thus means, Bob can compute d_{j2} and d_{j1} as follows

$$d_{j2} = g^{ak} g^{r_j} = (V'_{j1}/V_{j1})^{(h'_j - h_j)^{-1}} \quad d_{j1} = g^{ar_j} = (V'_{j2}/V_{j2})^{(h' - h)^{-1}}$$

Since we have

$$V'_{j1}/V_{j2} = (g^{ak} g^{r_j})^{(t+h'_j)} / (g^{ak} g^{r_j})^{(t+h_j)}$$

$$= (g^{ak} g^{r_j})^{(h'_j - h_j)}$$

$$V'_{j2}/V_{j1} = (g^{ar_j h'} / g^{ar_i h})$$

$$= g^{ar_i(h' - h)} .$$

According to the forking lemma, Bob can solve the **TDH** problem in expected time $t' \leq 120686 q_0 t/\varepsilon$.

□

For a person Carol who is not a member of the group, without the help of inner members she can't produce valid private keys. Carol chooses a random number $r_c \in \mathbb{Z}_q^*$, then she can compute $d_{c1} = g^{ar_c}$ since g^a is a public value. However, it is impossible for her to draw d_{c2} from public information. Given g^a and g^k , Carol can't compute g^{ak} under the assumption that **CDH** is intractable. This means that Carol can't forge a valid d_{c2} .

For a person David who is a member of the specified group, he can't forge valid private keys without help of other members. Given $d_{i1} = g^{ar_i}$, according to the **Lemma** that has been mentioned above, he can't compute g^{r_i} under **k-SDH** assumption. Thus means drawing g^{ak} from his private keys is intractable. It also indicates that a member can't forge valid private keys via his known information.

VI. Conclusions

The anonymity is crucial in some scenarios where a signer doesn't want to disclose his identity. The signers in both ring signature and group signature achieve anonymity by hiding themselves in a specified group. In this paper, we present another method that hides the signer in a group. From view of the point of verifier, the signature just comes from a specified group, and no one can be traced. The difference between the proposed anonymous signature and the ring signature is that the group used in ring signature is adaptively constructed by the signer, however the group used in our scheme is fixed, i.e., the signer can't choose the group as he want.

References

[1] D. Chaum, V. E. Heyst. Group signature. In Proceedings of EUROCRYPT'91. Lecture Notes in Computer Science, Berlin: Springer-Verlag, 1991, 547:257-265.

[2] R. L. Rivest, A. Shamir, Y. Tauman. How to leak a secret. In Proceedings of ASIACRYPT'01. Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2001, 2248: 552-565.

[3] E. Bresson, J. Stern, M. Szydlo. Threshold ring signatures and applications to ad-hoc groups. In Proceedings of CRYPTO'02. Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2002, 2442: 465-480.

[4] J. Herranz, G. Saez. New identity-based ring signature scheme. In Proceedings of ICICS 2004. Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2004, 3269: 27-39.

[5] J. K. Liu, D. S. Wong. On the security models of (threshold) ring signature schemes. In Proceedings of ICISC 2004. Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2005, 3506: 204-217.

[6] M. H. Au, S. S. Chow, W. Susilo, et al.. Short linkable ring signatures revisited. In Proceedings of EuroPKI 2006. Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2006, 4043: 101-115.

[7] G. Fuchsbauer and D. Pointcheval. Anonymous proxy signatures. In Proceedings of SCN 2008, Lecture Notes in Computer Science, 2008, 5229: 201-217.

[8] X. Boyen and B. Waters. Full-domain subgroup hiding and constant-size group signatures. In Proceedings of PKC 2007, Lecture Notes in Computer Science 2007, 4450: 1-15.

[9] C. B. Ma, J. Ao, and J. H. Li. Broadcast Group-oriented Encryption Secure against Chosen Ciphertext attack. Journal of Systems Engineering and Electronics. 18(4) (2007): 811-817.

[10] F. Zhang, R. Safavi-Naini, W. Susilo. An Efficient Signature Scheme from Bilinear Pairings and Its Applications. Practice and Theory in Public Key Cryptography-PKC 2004, Lecture Notes in Computer Science, Springer-Verlag, 2004, 2947: 277-290.

[11] J. Camenisch and J. Groth. Group signatures: Better efficiency and new theoretical aspects. In Security in communication Networks 2004, Berlin: Springer-Verlag, 2005, 3352: 120-133.

[12] D. Boneh, X. Boyen, and H. Shacham. Short Group Signatures. In Proceedings of CRYPTO 2004, Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2004, 3152: 41-55.

[13] L. Nguyen and R. Safavi-Naini. Efficient and Provably Secure Trapdoor-free Group Signature Schemes from Bilinear Pairings. In Proceedings of Asiacrypt'04. Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2004, 3329: 372-386.

[14] M. Abe, M. Ohkubo, K. Suzuki. 1-out-of-n signatures from a variety of keys. In Proceedings of ASIACRYPT'02. Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2002, 2501: 415-432.

[15] F. G. Zhang, K. Kim. ID-based blind signature and ring signature from pairings. In Proceedings of ASIACRYPT'02. Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2002, 2501: 533-547.

[16] K. C. Lee, H. Wei, T. Hwang. Convertible ring signature. IEE Proceedings of Communications, 2005, 152(4): 411-414.

[17] L. Nguyen. Accumulator from bilinear pairings and application to ID-based ring signatures and group membership revocation. In Proceedings of CT-RSA 2005. Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2005, 3376: 275-292.

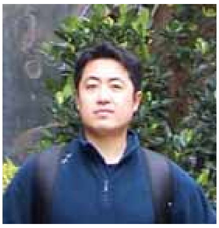
[18] T. Isshiki, K. Tanaka. An (n-t)-out-of-n threshold ring signature scheme. In Proceedings of ACISP 2005. Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2005, 3574: 406-416.

[19] P. P. Tsang, V. K. Wei. Short linkable ring signatures for E-voting, E-cash and attestation. In Proceedings of ISPEC 2005. Lecture Notes in Computer Science, Berlin: Springer-Verlag, 2005, 3439: 48-60.

[20] Y. Q. Chen, W. Susilo, Y. Mu. Identity-based anonymous designated ring signatures. In Proceedings of IWCMC 2006. USA: ACM Press, 2006, 189-194.

[21] D. Proincheval and J. Stern. Security aguments for digital signatures and blind signatures[A]. J. of Cryptology, 2000, 13(3):361-396.

[22] E. Brickell, D. Pointcheval, S. Vaudenay, Yung M. Design validations for discrete logarithm based signature schemes. In PKC'2000, Lecture Notes in Computer Science, Vol. 1751. Springer-Verlag (2000) 276-292.



Chunbo Ma received the PhD in Communication and Information System in 2005, from the Southwest Jiao Tong University in P. R. China. He spent two years of postdoc research at the Shanghai Jiao Tong University in the field of Information Security and Cryptography. Currently, he is a professor for Information Security at the Guilin University of Electronic Technology in P. R. China.

Jun Ao received the M.S. in Guilin University of Electronic Technology in 2003. Currently, she is a PhD. candidate in Xidian University, Shaanxi, P. R. China. Her research interests include Radar Signal Processing, Coding, and Mobile Communication System.

Using Context to Improve Hand-written Character Recognition

Richard Fox and William Hartmann

Abstract—**Automated hand-written character recognition has been solved by a variety of techniques including neural networks, hidden Markov models and Bayesian probabilities. These approaches have led to highly accurate character recognition but these approaches typically do not attempt to bring into the process contextual information. This article describes an approach to hand-written character recognition, using layered abduction, which utilizes domain knowledge to generate top-down guidance in the form of character expectations. Four different domains have been tested, each using different forms of knowledge to drive top-down processing. Character recognition accuracy prior to the use of top-down guidance is poor (50%) but with top-down guidance, accuracy improves to 99% for upper case English letters, digits and some punctuation marks.**

Index Terms—**character recognition, pattern recognition, abduction, knowledge based systems.**

I. Introduction

AUTOMATED hand-written character recognition dates back to the 1960s as the US postal service experimented with the problem [1]. In four decades of research, accuracy has greatly improved, largely through the application of Artificial Intelligence approaches like neural networks, statistics, and genetic algorithms (see for example [2, 3, 4, 5, 6]). When the characters to be recognized are machine-produced, automated recognition systems do extremely well. For human-produced hand-written character recognition, accuracy varies, especially depending on the language (e.g., Arabic characters are harder to recognize than English characters) and whether the characters were printed or written in cursive. Accuracy of 80-90% is often found to be respectable whereas some recent systems have had results in the 97-99.8% range [7]. Consider, the poor accuracy of an HP LaserScanner, shown in fig. 1. The neural network used by the LaserScanner's intelligent scanning feature was trained to recognize machine-produced print, not human hand-writing.

Most automated character recognition systems perform the recognition task without the benefit of domain-specific or pragmatic knowledge. In some cases, this restriction is the consequence brought by the very nature of the problem solver (e.g., neural networks do not have the capacity to utilize higher-level knowledge sources). In other cases the restriction is because such knowledge was just not applied. However, this additional knowledge is applied by humans all the time. Consider for instance a mailman reading a United States address and seeing a character that looks like an "S" in the zip code. The postman will infer that the character in question is a sloppy "5" because US zip codes do not contain letters. Bringing such knowledge into consideration has proved very useful in automated perceptual problem solvers in areas such as speech recognition and natural language processing. However, it has largely been ignored in character recognition, forcing the processing to be merely a bottom-up approach.

Neural networks, which have demonstrated among the highest accuracy for hand-written character recognition, learn the domain through a bottom-up process with top-down learning. However, once training is over, the network remains static and is unable to apply additional knowledge sources. Alternatively, a feature-based pattern matching approach can apply a variety of knowledge sources [8]. Through the use of layered abduction, a problem solver can analyze data in a bottom-up fashion and then apply domain-specific knowledge in a top-down process to permit error correction and search guidance [9]. This paper describes such an approach where layered abduction is applied to printed hand-written character recognition. Four domains have been tested: mathematical equations, postal addresses, bank checks, and English sentences with a restricted syntax. The overall accuracy for character recognition is over 99%. It should be noted that the intention of this research was not to argue that hand-written character recognition should be solved by a knowledge-based approach using features and pattern matching, but rather, the research hopefully demonstrates the value obtained by using abduction and domain-knowledge to provide top-down guidance.

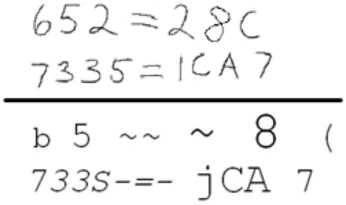


Fig. 1. Human hand-written input is shown in the first two lines of this figure with the HP LaserScanner's inaccurate output shown on the last two lines.

Manuscript received November 10, 2008. This work was supported in part by the Center for Integrated Natural Science and Mathematics at Northern Kentucky University.

Richard Fox is an associate professor in the Department of Computer Science at Northern Kentucky University, Highland Heights, KY 41099 USA (859-572-5334; fax: 859-572-6176; e-mail: foxr@nku.edu).

William Hartmann is a doctoral student in the Department of Computer Science and Engineering at the Ohio State University, Columbus, OH 43210 (e-mail: hartmanw@cse.ohio-state.edu).

This paper is organized as follows. First, abduction and layered abduction are described along with methods for how to implement them and how top-down guidance can be applied. Next, CHREC, the Character RECognizer system is introduced. This is followed by a section that describes the four domains implemented, along with examples and experimental results. A brief conclusion follows.

II. Abduction and Layered Abduction

Abduction is *inference to the best explanation*. The task is one of selecting among plausible hypotheses, the hypothesis or hypotheses that provide the *best* way to explain the given observations or findings. This problem solving task is found in a variety of problems such as diagnosis, theory formation, test interpretation, and perception. Given a set of findings to be accounted for (from this point forward, findings will be referred to as data), hypotheses are proposed to explain the data. A single hypothesis may not be able to account for all of the data, so instead a collection of hypotheses are used to explain the data. This collection, a composite hypothesis, should be consistent and as plausible as possible given the individual hypotheses available.

In fig. 2, an abstract abduction problem is shown in which the goal is to explain the data {D1, D2, D3, D4, D5} using some subset of available hypotheses {H1, H2, H3, H4, H5, H6}. In the figure, the hypotheses have been evaluated by some mechanism that have provided plausibility scores on a scale from 0 to 1 (the higher the value, the more plausible the hypothesis is). Solid lines in the figure denote the explanatory power of a given hypothesis (what a hypothesis can explain) whereas the dotted line indicates that these two hypotheses are mutually exclusive or incompatible with each other (both hypotheses cannot appear in the explanation). In this example, there are several possible explanations for the data such as {H1, H2, H3} and {H1, H5}. However, the best explanation is probably {H4, H6} because it is the shortest complete explanation that has the highest overall plausibility. Note that {H1, H4} can also explain the data, but it is an inconsistent explanation because H1 is incompatible with H4.

Abduction research in Artificial Intelligence dates back to the Internist medical diagnosis system [10] and has been used to tackle such problems as natural language understanding [11], theory formation [12], data interpretation [13], and speech recognition [14]. Different techniques have been implemented for abduction including pattern matching knowledge-based approaches [10, 13], first order predicate calculus [15], neural networks [12], naïve Bayesian probabilities [11, 16], Bayesian networks [17], and hidden Markov Models.

In [13], a single domain- and problem-independent strategy was developed. The strategy comprises subtasks of hypothesis generation, hypothesis instantiation and hypothesis assembly.

Hypothesis generation is the process of obtaining plausible domain hypotheses which can potentially explain findings. Generation may be accomplished by any number of possible methods. In hierarchical classification [18], a search proceeds top-down through a taxonomy of hypotheses organized by specificity. Cuing generates hypotheses using associations with the findings to be explained. Hypotheses may be generated from first principles using some functional or causal model. Heuristic search strategies through some organized hypothesis search-space can be used. A neural network may be able to suggest hypotheses by presenting findings to it so that the network uses spreading activations to generate hypotheses. A series of naïve Bayesian classifiers or a hidden Markov model could also be used to generate hypotheses.

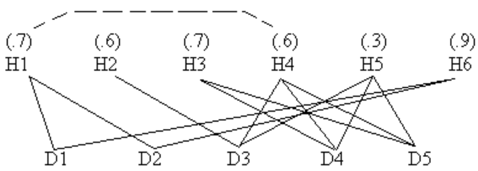


Fig. 2. Abduction is the process of forming a composite explanation for the given data. In this example, six hypotheses have been proposed to explain the five data. Plausibility values have been generated for each of the hypotheses (shown in parentheses as values between 0 and 1), solid lines indicate what a hypothesis can explain and the dashed line denotes mutually exclusive hypotheses.

Hypothesis instantiation is the process of determining each generated hypothesis' relevance for the current case. This subtask consists of at least two separate steps. First, the given hypothesis is evaluated for an initial plausibility. Second, the findings are examined to determine what the given hypothesis might be able to explain. In addition, the hypothesis might be compared to other hypotheses to determine such hypothesis-dependency factors as whether this hypothesis is related to others or is incompatible with others. The evaluation might be performed using pattern matching knowledge by examining the presence and absence of features of interest among the data, or through Bayesian probabilities, statistical pattern matching, strength of neural network activation or some other method. The explanatory coverage and hypothesis interaction knowledge may be given as static information (for instance, in the form of a Bayesian network or hidden Markov Model), derived from first principles, or other approach.

Abductive assembly is the process of building a composite explanation to account for the data using previously generated and instantiated hypotheses. This task might be performed by an attempt to generate all possible combinations of explanations out of the simpler hypotheses, evaluating each combination and selecting the best. A holistic approach might generate the composite explanation all at once, such as by using neural or belief networks. One might attempt an incremental building of a composite hypothesis out of parts to avoid any intractability. The composition process should include some form of criticism so that the composite can be compared with other possible explanations to ensure a final composite which is consistent, parsimonious and superior to any other explanation.

This paper uses the following approach, as shown in fig. 3 (at the top of the next page). A search space of hypotheses is used to generate *plausible* hypotheses. Hypotheses are evaluated using feature-based pattern matching where each individual hypothesis has its own recognizer. The recognizer not only assigns a plausibility value, but also indicates what the given hypothesis can explain from the data. The set of

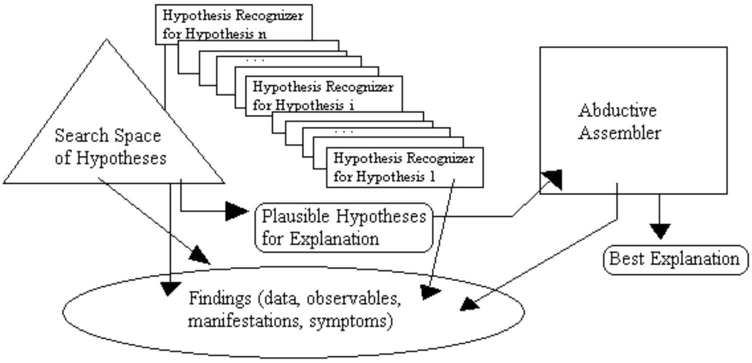

Hypothesis Recognizer for Hypothesis n
. . .
Hypothesis Recognizer for Hypothesis i
. . .
Hypothesis Recognizer for Hypothesis 1
Search Space of Hypotheses
Abductive Assembler
Plausible Hypotheses for Explanation
Best Explanation
Findings (data, observables, manifestations, symptoms)

Fig. 3. The abduction process is shown here. Findings (data, symptoms, etc) are introduced to be explained. A space of hypotheses is searched for plausible hypotheses that can be used to explain some of the findings. Each hypothesis recognizer is specialized to evaluate if its corresponding hypothesis in the search space is plausible. After search, a set of plausible hypothesis is available for the abductive assembler to construct a composite explanation (as shown in figure 2). The output of the process is a composite explanation that can explain all of the findings without having any mutually exclusive or superfluous components and has the overall highest plausibility.

evaluated hypotheses are then provided to an abducer which performs abductive assembly to generate a composite explanation to represent the best explanation for the input data. During abductive assembly, the abducer considers the various findings and selects the best hypothesis to explain them. This process is repeated by selecting other hypotheses until all data are explained. As a hypothesis is selected, the data it can explain are removed from consideration, and any hypothesis-dependent information is applied (for instance, mutually incompatible hypotheses are ruled out). The best explanation is the explanation which is the most plausible, complete, consistent and parsimonious (i.e., no superfluous parts) collection of hypotheses. See [9] for a more complete description of this abduction process and how it performs on a number of different problems.

For perceptual problems, a layered approach is taken where the composite hypothesis accepted at one level of description becomes the data to be explained at a higher level of description [19]. This allows for an explanation to be generated using a variety of different knowledge types that is appropriate to the problem.

For example, in speech recognition, the acoustic signal might be explained by phonemes, but this may not be a sufficient explanation as it does not provide an explanation in terms of words (a composite hypothesis of phonemes does not indicate where one word stops and the next starts) or semantic meaning. The phonemes become data to be explained by word hypotheses, which in turn become data to be explained by grammatical category hypotheses, and so forth until the sentence and its meaning are provided as an explanation. The control strategy for layered abduction includes both bottom-up and top-down elements.

Top-down processing can be used so that accepted hypotheses at a higher level of abstraction provide guidance in further processing at a lower level. In speech recognition, for example, knowing that the verb has already been found in the sentence might assist a lower-level problem solver when it is considering the next word and has as candidates a noun and a verb. The top-down guidance can remove an incompatible hypothesis from consideration. On the other hand, top-down guidance might provide a hypothesis with a boost in plausibility because it was expected. Tough decisions in the assembly process whereby the abducer must select between two or more highly plausible hypotheses might be resolved in this way.

In order to solve hand-written character recognition through abduction, the problem becomes one of hypothesizing the characters responsible for the bitmap input and selecting the best character hypothesis for each character. Hypothesis generation will provide character hypotheses, each of which can explain a section of the input bitmap. Hypothesis instantiation will determine how useful each character hypothesis is in explaining the given character (how plausible it is, what features it can explain, whether it is compatible with other hypotheses). Abductive assembly is the selection of character hypotheses that make up the best explanation, or the recognition of the input. Abductive assembly takes into account consistency by applying domain-specific knowledge and applying top-down guidance when available.

III. THE CHREC SYSTEM

CHREC (Character RECogznier) is a knowledge-based approach to automated character recognition through layered abduction. The system has been implemented to recognize the 26 upper case English letters, the ten digits and a few punctuation marks (period, comma, equal sign, ampersand, division sign). CHREC has been tested in four domains and utilizes domain-specific knowledge for top-down guidance. These domains are mathematical equations, US postal addresses, bank checks and short English sentences.

There are three restrictions placed on input to CHREC. First, characters must be written at a reasonable size. Second, no character is allowed to overlap another horizontally or vertically. Third, characters are expected to be written in a reasonably horizontally straight line.

Fig. 4 illustrates CHREC's architecture, consisting of four components. The *parser* segments the bitmap into regions where each region is either a blank space to separate words or a character to be explained. *Feature detectors* analyze a region and generate features which act as the data to be explained. Features include lines and arcs. There is a set of *character recognizers*, one for each character that the system can recognize. Character recognizers are responsible for hypothesis generation and instantiation. The *abducer* performs two roles. First, it performs abductive assembly to generate an initial explanation. Second, it uses the generated explanation together with domain-specific knowledge to create expectations. These expectations are used as top-down guidance to update individual character hypotheses. This is followed by a second abductive assembly in an attempt to improve CHREC's answer. For the English sentence domain, layered abduction is used. The second abductive assembly allows CHREC to use syntactic knowledge for additional top-down guidance. Grammatical role hypotheses are generated, instantiated and used as an explanation. These components are described in more detail in the following sections.

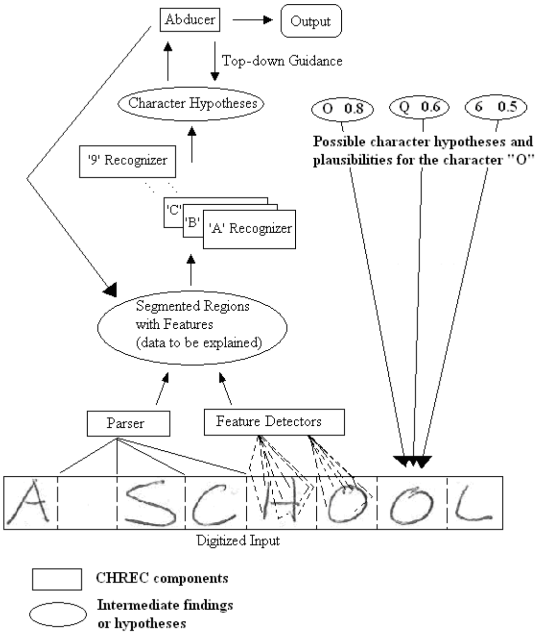


Fig. 4. The CHREC architecture. The parser segments the bitmap into characters to be recognized. The feature detectors generate features to be explained. Character recognizers evaluate how likely each character is at explaining the features. The abducer then generates a best explanation along with top-down guidance to improve its explanation.

A. Parser and Feature Detection

The *parser* examines the bitmap for blank vertical space. There are two types of segmentation that the parser performs. A small amount of blank space (at least one vertical column of white pixels) is expected between characters. When found, the parser denotes this as the separation between two characters. If there are several columns of white pixels, then the parser denotes the region as being between two words. Word separation is found in all but the first domain. It is assumed that words will be separated by a greater amount of blank space than the characters in a word. Once segmented, all features found within a single region are to be explained by a single character hypothesis.

A suite of *feature detectors* examines a given segment of the bitmap to identify features to be explained. These features include lines, curves, and noise (pixels that do not correspond to lines or curves). The feature detectors also establish attributes of each feature such as the direction (downward diagonal, upward diagonal, vertical, horizontal), slope and distance. Curves have degrees of openness, and open up, down, left or right. Feature detectors provide approximate location information of the feature (top, middle, bottom, left, center, right) within the segment. Finally, because there may be some uncertainty in the exact shape, the feature detector provides a rating of how certain the generated feature is. The rating is used by character recognizers to determine whether a given feature truly needs to be explained. Features are the output generated from feature detectors, and are used as the data to be explained by character hypotheses. Table 1 provides the list of features and attributes in CHREC.

TABLE I
CHREC FEATURES

Feature	Attribute / Value
Curve	Angle (line segments < 170 degree angle)
Line	Angle (line segment > 160 degree angle)
Width	Large, Medium, Small
Height	Large, Medium, Small
Slope	Vertical, Horizontal, +/- Diagonal
Curve Opening	N, NE, E, SE, S, SW, W, NW
Connections	None, Partial, Fully
Location	Top, Top Left, Top right, Left, Right, Bottom left, Bottom right
Region	Top, Middle, Bottom, Left, Center, Right

B. Character Recognizers

The current implementation of CHREC can recognize the letters 'A'-'Z', digits '0'-'9', and punctuation for comma, period, equal sign, ampersand, and division sign. For each of these characters, CHREC has a *character recognizer*. For instance, the "A" recognizer will evaluate any set of features in a segment to determine how likely it is that those features make up a letter "A." Character recognizers use feature-based pattern matching. For the given character, features that are expected to be found and features that are not expected are enumerated. Based on the number of features that are found which are expected and those features that are found which are not expected, the character recognizer generates a plausibility score. Plausibility scores are real numbers between 0 (ruled out) and 1 (highly plausible). A score of 0.5 or higher indicates reasonable plausibility. In the first implementation of CHREC, all character recognizers were hand-coded, but this was later replaced by a modest learning algorithm. There almost no difference in CHREC's performance.

As an example, the "7" Character Recognizer looks for a roughly horizontal line near the top and a diagonal line from bottom left to upper right, and no curves or other lines. Table 2 presents the character recognizer for the character "F." The features of interest are listed with a "+" and the features that should not be found are listed with a "-." Omitted from the table are the plausibility values (because of space restrictions).

TABLE II
CHARACTER RECOGNIZER FOR "F"

Feature	Specification	Present or Absent
Line	V, TL, M	+
Line	V, ML, M	+
Line	H, TL, S	+
Line	H, ML, S	+
Line	H, BL, M	-
Line or Curve	-, R, -	-

Specification is abbreviated here. The first specifier is the direction (V for vertical, H for horizontal). The second specifier is the location (T for top, M for middle, B for bottom, L for left, R for right). The final specifier is the size (L for large, M for medium, S for small). + means a the feature is expected, - means the feature is not expected.

Many of the character recognizers contain multiple sets of features because of the varying ways that people might write those characters. "7" for instance is sometimes written with a horizontal line in the middle. The "7" recognizer has one set of features that includes this second horizontal line and one set that does not. The "0" recognizer also has two sets of rules to represent the number without and with a diagonal line going up from left to right. Other character recognizers with multiple sets of features include "4" and "I." Some recognizers recognize very similar characters ("O" and "0," "G" and "6," "M" and "N"). This problem is dealt with by the abducer when possible. Effort must go into fine tuning the character recognizers. In fact, a good deal of the effort to improve CHREC's accuracy has gone into adding more patterns to the character recognizers.

For each segment, CHREC generates a character hypothesis for every character known in its lexicon, rather than just those that are deemed as plausible. This approach is overly cautious in that some character hypotheses would clearly be irrelevant, however since it does not require a significant amount of additional processing to generate hypotheses for all characters, this was the approach taken. CHREC calls upon the appropriate character recognizers to determine how plausible each character hypothesis is.

C. *The Abducer*

Once the character recognizers have executed, CHREC will have hypotheses for each region's set of features. Now, it is up to the abducer to perform abductive assembly to form the best explanation of the digitized input in terms of the most likely characters. See for instance fig. 5 where two character recognizers (E and F) have proposed characters. Assuming that the two hypotheses are equally plausible, the "E" hypothesis can explain more than the "F" hypothesis and so would more likely be selected by the abducer during abductive assembly. It is likely that in fact "E" has a higher plausibility anyway for this example since more of its features are found. However, the benefit of adding explanatory knowledge to an explanation provides a mechanism whereby the abducer can more correctly find the proper hypothesis (see [9] for a more complete analysis of explanatory value).

The abducer examines the findings of each segment and compares them to each reasonably plausible character hypothesis for that segment. Any hypothesis with low plausibility is ignored for now. CHREC currently uses a threshold of 0.3 to determine if a hypothesis is reasonable. The abducer selects one hypothesis to explain the data of that segment. The abducer completes abductive assembly once all of the segments have had a character selected. At this point,

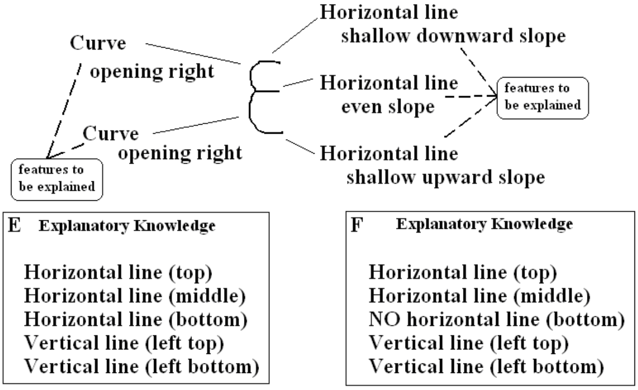


Fig. 5. Two plausible hypotheses have been generated to explain the features found in this segment. The character is an upper case "E", and the two hypotheses are that it is an "E" and an "F". Notice that while both hypotheses may be equally plausible, the "F" hypothesis cannot explain all of the features. Plausibility scores are not shown in this figure.

the abducer "pieces" together the characters to form the explanation. In many cases, the pieced together explanation represents one or more words or numbers.

Now the abducer must examine the explanation for consistency. This step requires the application of domain specific or pragmatic knowledge. One form of pragmatic knowledge is that numeric values will not contain letters, and words will not contain digits. If the abducer decides that the given item should be a number, then any characters that represent letters or punctuation marks must be questioned. Similarly, if the abducer decides that the given item should be a word, then any character that represents a digit or punctuation mark must be questioned. Such knowledge generates expectations for top-down guidance (see the following subsection). Domain specific knowledge is applied in a similar way, once the pragmatic knowledge has been utilized. Domain specific knowledge will be described in more detail in the next subsection and in section 4.1.

D. Top-down Guidance

After the initial abductive assembly is performed, the abducer examines the explanation for consistency. Consistency checking is domain dependent. For instance, in the mathematical equation domain, the two numbers must equate to each other, and in the bank check domain, the amount written in English must match the number. If the abducer finds any inconsistencies, it will generate expectations to provide top-down guidance. An expectation is targeted for a given segment so that this segment can be re-examined. The expectation itself might be an incompatibility such that a given hypothesis was found incompatible with the explanation, or it might be a positive reinforcement of a hypothesis in that the explanation expects a particular hypothesis.

An example of top-down guidance is shown in fig. 6, in which the abducer expects that the state abbreviation should be "MN" based on the recognized city and zip code since the recognized state abbreviation of "MI" is incompatible with the rest of the explanation. Therefore, at least two expectations can be generated, the "I" character hypothesis for the 7th character should be considered less likely and thus its plausibility should be lowered while the "N" character hypothesis should be considered more likely and its plausibility should be raised. It is possible that other expectations, such as for "A" or "T" can also be lowered.

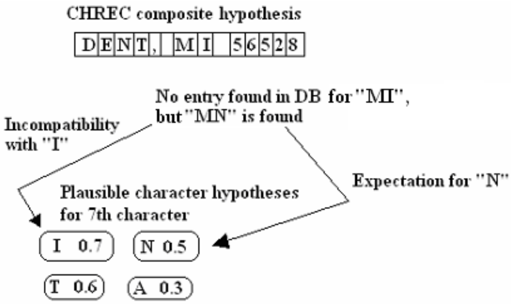

CHREC composite hypothesis
D E N T , M I 5 6 5 2 8
No entry found in DB for "MI", but "MN" is found
Incompatibility with "I"
Expectation for "N"
Plausible character hypotheses for 7th character
I 0.7
N 0.5
T 0.6
A 0.3

Fig. 6. Generating expectations provides top-down guidance for CHREC. Here, CHREC has proposed that the 8th segment can be best explained by the character "I", although "N" is a plausible alternative. However, domain knowledge indicates that there is no Dent, MI 56528 but there is a Dent, MN 56528, so CHREC uses this knowledge to correct itself.

An expectation is applied by altering the target hypothesis' plausibility. For an incompatibility, the plausibility is lowered. For a reinforced expectation, the plausibility is raised. The amount lowered or raised is a pre-set amount, however, it might also be reasonable to adjust the plausibility based on overall plausibility of the remainder of the abducer's explanation. That is, as the abducer becomes more certain in explaining the rest of the data, the greater the expectation should be for the target hypothesis resulting in a greater adjustment to that hypothesis' plausibility.

IV. CHREC IN ACTION

This section describes the four domains for which CHREC has been implemented. A brief example is offered for the first three domains and a more detailed example is offered for the final domain. The section concludes with experimental results of CHREC's accuracy in each of the domains.

A. Domains

In order to test CHREC, the first domain chosen was decimal to hexadecimal equations, such as 106=6A. There were two reasons for this implementation. First, early on CHREC only had recognizers for the ten digits and the equal sign. To complete this implementation, only six letter recognizers were required (for the letters "A" – "F"). Second, it was thought that top-down guidance would be easy to implement in this domain since it only required converting numbers from one base to the other and generating expected characters based on any incorrect characters. This domain turned out to be easy to implement with one exception, when there were errors in both numbers. That is, if the numbers on both sides of the equal sign were incorrect then converting a number from one base to the other would create expectations that were not necessarily useful since there was an error in the original number. CHREC made no assumption regarding which of the two numbers was the decimal value and which was the hexadecimal value. Because of this, it was decided that the abducer would convert both numbers into both bases (if the number were 1234, it would not be possible to determine if this was a decimal or hexadecimal number). The parser would be able to determine the exact number of characters expected for each number. Therefore, if a converted number had too few or too many digits, that number would be discarded and no expectations would be generated from it. Otherwise, all converted numbers would be used to generate expectations. The number with the highest overall plausibility would be used first. Some give and take processing would be required if there were multiple errors. Only in a couple of cases was CHREC so confused that it could not find the correct answer. Once character hypotheses were altered, a second abductive assembly would be performed.

The second domain chosen was US postal addresses. For simplicity, addresses were limited to city, state (two letter abbreviation) and 5-digit zip code. A database of city/state/zip code values was used that contained approximately 27,000 entries. Once the abducer had performed abductive assembly, the city/state/zip were sought from the database. Approximate matching was used to select closely matching entries. Every approximate match would be used to generate expectations (refer back to figure 6), which would be ordered based on how closely each entry matched. Since the state was only two characters in length, an exact match would not necessarily be as confident a match as an exact or close match for city or zip code. Expectations would be used to alter character hypotheses, followed by a second abductive assembly.

The third domain chosen was one of matching bank check dollar amounts as written in English against the numeric values. This is much like the mathematical equation domain in that the dollar and cents amounts must match between the English words and the numeric values. A small database was created of expected words (the various numeric values as English words such as five, fifty, hundred, and the word dollar). Once CHREC's abducer has generated an explanation, the two parts are compared for equality. If a word matched exactly, then expectations could be generated for the numeric value. If a word matched closely, since the words in the database are mostly dissimilar to each other, then the word was corrected and expectations would be generated. Only if a word did not closely match anything in the database would CHREC have difficulty. After expectations were generated and character hypothesis plausibilities' updated, the abducer would perform abductive assembly again. It is assumed that the input English words match the numeric values as one would expect in a legal bank check. For further details on these first three domains, see [20, 21].

The final domain was recognizing words of short English sentences with a lexicon of 500 words. Rather than implementing a full parser, 15 simple syntaxes were derived, five each for four-word sentences, five-word sentences and six-word sentences. This grammatical restriction prohibits CHREC from being a general-purpose English word recognizer, but this was never the goal so it was felt to be a reasonable restriction. For this domain, two levels of abductive assembly would be performed. First, features would be explained by characters which would be grouped into words, and like the postal address and bank check domains, consistency checking would be performed by determining if these words were legal (i.e., found in the database). Top-down guidance would correct words that did not match any given word. Next, the words would be converted into grammatical categories and a second abductive assembly would be performed by matching the grammatical categories to one of the legal syntaxes for the given type of sentence (4-word, 5-word or 6-word). Mismatches would require top-down guidance again by examining the other likely words for that particular portion of the input.

B. Examples

The example for the mathematical domain is the equation 285=11D (see the upper section of fig. 7, on the next page). CHREC originally infers the input to be the equation 285=11B. The problem here is that CHREC had difficulty distinguishing between the features that make "D" differ from "B" with "D" getting a plausibility of 0.8 and "B" getting a plausibility of 0.6. The mistake may have been caused because of the fainter ink on the curve for "D." Top-down guidance detects the problem since 285 should be 11D and 11B should be 283. With this guidance, expectations for "3" in the third position and "D" in the final position are generated. These associated character hypotheses have their plausibility values increased. This results in the score for "3" being raised, but not sufficiently to replace "5" whereas the score for "D" is raised enough to replace "B." Abduction is run again and CHREC concludes that the correct answer is "285=11D."

The second example comes from the postal address domain where the input is the address SAINT JAMES, NY 11780 (see the middle section of fig. 7). CHREC originally recognized the address as SA1NT JAAEC, NV 11780. There are several problems with CHREC's initial attempt at recognition. First, the letter "I" was mistaken for the number "1." Since this portion of the input falls within a group of letters only, CHREC uses pragmatic knowledge that city names should not include numbers, and so decreases the likelihood of any number appearing in that portion of the input. Next, the database fails to find any close match in NV for either SA1NT JAAEC or the zip code 11780, however it does find a close match in NY. Therefore, expectations are generated so that the "Y" hypothesis is increased and the "V" hypothesis is decreased. Finally, no match for a city is found in the database. Both the eighth and tenth characters have problems. The most plausible characters for the eighth position were "A" with a 0.6 score and "M" with 0.5, and the most plausible characters for the tenth position were "C" with 0.56 and "S" with 0.5. Since the abducer found a close match between "SA1NT JAAEC" and "SAINT JAMES," it generates expectations for "I," "M" and "S" for the third, eighth and tenth positions respectively. These plausibility values are all increased. CHREC reruns the abducer and selects "SAINT JAMES, NY 11780" as the best answer, which matches entirely a database entry.

The third example comes from the bank check domain where the input is "FOURTEEN DOLLARS & 09 / 100" and "14.09" (see the bottom section of fig. 7 where the input has been broken into four different parts in order to save space in the article). CHREC originally identifies the input as the very erroneous "F0B2T0EN DOLLAR1 8 O911O0" and "19109." The database of words for this domain includes only words for the various numbers in English along with the words "dollar" and "dollars." The abducer is able to closely match "FOURTEEN" for the first word, "DOLLARS" for the second word and "&" for the next character. These are easily corrected, using the same strategy as that from the postal address domain. More difficult is correcting the cents portion of the English words portion, however since this should equal the cents portion of the numeric value, the abducer examines which is more plausible, the "O0" of the English description side or the "09" of the numeric value. Since there already is an error in "O0" (there should be no letters), and in this case since "0" and "9" were selected with high plausibility, CHREC generates expectations to fix the "O0" instead. The "/ 100" is easily corrected since this is expected to end the

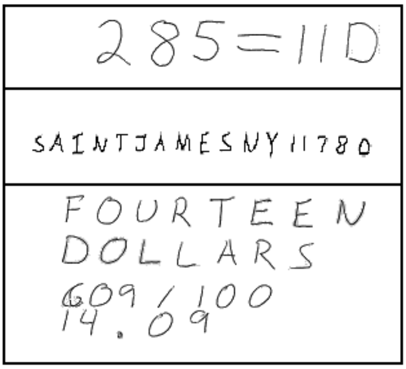


Fig. 7. Sample inputs from three domains, mathematical equations where, in this example, 285 = 11D, post office addresses, in this example, Saint James NY 11780, and bank checks, in this example Fourteen Dollars & 09 / 100 is equivalent to 14.09.

English description portion for all checks. With the expectations generated, the abducer reruns abductive assembly and CHREC is able to obtain the correct answer.

The final domain is the English sentence domain. An example for a four-word sentence is shown in fig. 8 (on the next page), the sentence being "A school was left." Notice that while this domain used syntactic processing, sentences did not necessarily make semantic sense. CHREC begins this example much like the other domains. First, features are detected which are used as data to be explained. Figure 8a shows the features generated for the "school" portion of this example (there is not enough space to show all of the features for the sentence, nor the attributes for the features). Next, CHREC generates plausible character hypotheses for each word. Figure 8b shows the most highly evaluated character hypotheses for the word "school." Notice how some of the characters are numbers.

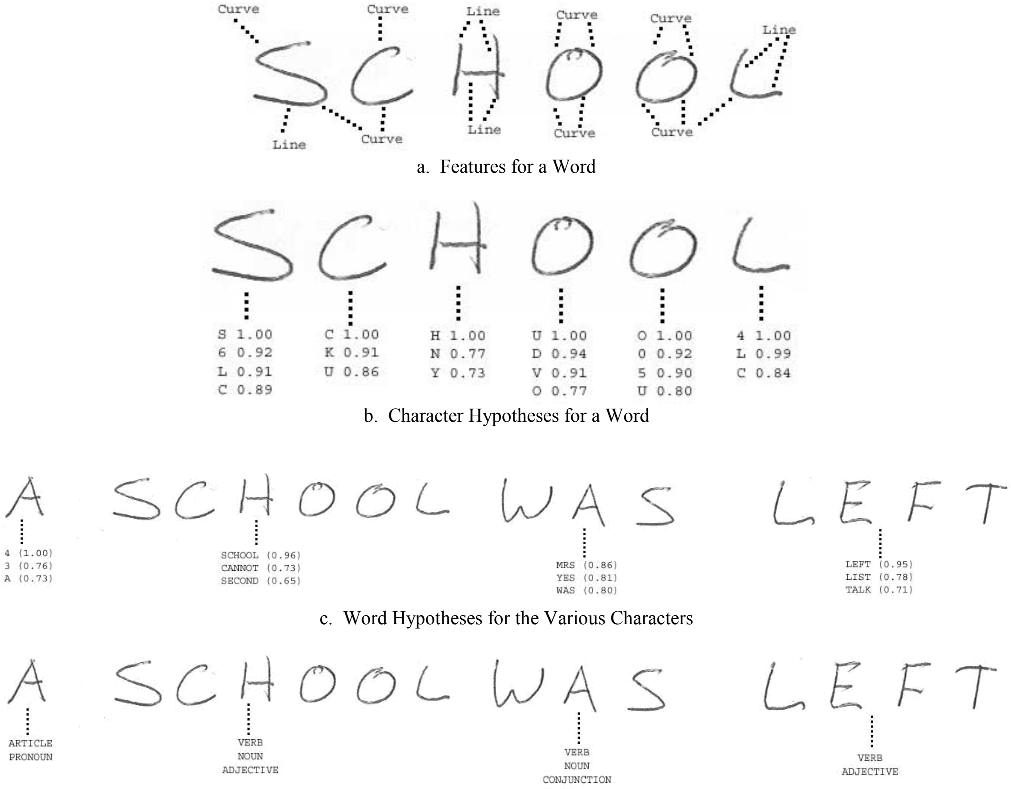
At this point, CHREC has not determined that a given item should fall within a word or a number, and so numeric characters are permissible. With the character hypotheses available, CHREC's abducer performs abductive assembly. Now, CHREC tries to explain the features in terms of characters. The abducer now groups selected characters into words, separating words based on the separations found by the parser. Once a word has been selected for each word in the input, CHREC checks the words for lexical consistency. This is done by consulting the database of words to ensure that every word that CHREC composed was legal. Figure 8c shows the most plausible words as generated by CHREC for the words of the sentence. Plausibility is determined by a simple matching algorithm that determines if plausible characters were found in the location expected. Abductive assembly is performed again, this time where the abducer selects the word that best explains the collection of characters in the region (which in turn best explain the features found in the input). The selection of the word is almost always the most plausible word (exceptions could occur for instance if features were found that a word cannot account for).

The abducer now groups selected characters into words, separating words based on the separations found by the parser. Once a word has been selected for each word in the input, CHREC checks the words for lexical consistency. This is done by consulting the database of words to ensure that every word that CHREC composed was legal. Figure 8c shows the most plausible words as generated by CHREC for the words of the sentence. Plausibility is determined by a simple matching algorithm that determines if plausible characters were found in the location expected.

Abductive assembly is performed again, this time where the abducer selects the word that best explains the collection of characters in the region (which in turn best explain the features found in the input). The selection of the word is almost always the most plausible word (exceptions could occur for instance if features were found that a word cannot account for). In order to determine consistency of the selected words, the abducer now employs syntactic knowledge. This is done by identifying for each selected word its grammatical role in the sentence. Most words in CHREC's lexicon have only one grammatical category, simplifying the task. This step might result in a structure such as ARTICLE – ADJECTIVE – NOUN – VERB. The abducer now checks the validity of this structure; is it one of the five legal syntactic structures for a sentence of this length? If an exact match is found, then the abducer assumes that the recognized sentence is consistent and CHREC terminates. Otherwise, the abducer must find a closely matching syntax. This is done by selecting all of the syntaxes that have at least half of the words' categories match the given syntax.

The abducer now identifies each word that was misrecognized with a grammatical category mismatch. The legal grammatical categories for each word are generated and used as expectations to be applied to the remaining word hypotheses left over from the previous abductive assembly. For instance, if the first word found was not an ARTICLE, and ARTICLE is the only grammatical category that could fit the given sentence, then an "ARTICLE expectation" is generated. All highly rated words that were not selected for this first word are reconsidered. If any are articles, then their plausibility is raised. Similarly, any word that is not an article has its plausibility lowered. If two possible syntactic structures were found in which the first word could be an article or an adjective, then two expectations are generated for the first word of the sentence. Figure 8d demonstrates the grammatical categories, word-by-word, for the given sentence. Once the expectations are applied, the abducer performs a new abductive assembly. This is the final step in this domain and the best explanation here is CHREC's resulting answer. Had there been semantic knowledge, another level of abduction and top-down guidance could be applied.

Consider as another example, the sentence "Think outside the box." CHREC originally recognized the input as "THIVYUVTSFDLTHZR6R." Through top-down guidance with lexical knowledge, the abducer settled for the answer "THINK OUTSIDE THE BOY." In this case, both "boy" and "box" are nouns and so they both satisfy the expected syntactic structure and CHREC was not able to accurately recognize the last character, providing an erroneous sentence.



a. Features for a Word

b. Character Hypotheses for a Word

c. Word Hypotheses for the Various Characters

d. Grammatical Role hypotheses for Recognized Words

Fig. 8. An example of the English sentence domain showing all of the processing levels that CHREC goes through. There are multiple levels of abduction here, to select the proper characters, words and grammatical roles.

With semantic knowledge, CHREC might have been able to find the correct answer.

C. Experimental Results

In the mathematical equation domain, 75 equations of 526 total characters were tested, character accuracy was 96%. In the postal address domain, 100 addresses were tested consisting of 1515 total characters. This domain used a database of over 27,000 city/state/zip code entries. CHREC was nearly perfect in its character recognition accuracy (achieving a 99.6% accuracy). In the bank checking domain, 15 checks were tested comprising 552 characters. The only words used were those that are found in common dollar amounts up through thousands. In the sentence domain, 50 example 4-word, 5-word and 6-word sentences were tested comprising 1042 characters. This domain used a lexicon of 500 words and 5 forms of syntax. CHREC used the learned character recognizers for these last two domains. Character accuracy for these last two domains was 99%.

Table 3 (on the next page) provides the accuracy of CHREC in each domain, along with an overall accuracy. For each domain, the “Before” entry indicates the accuracy as determined by CHREC without top-down guidance. The “After” accuracy is the accuracy after top-down guidance had been applied and abductive assembly performed a second time. For the sentence domain, the abducer would run abductive assembly once if the explanation was syntactically valid, twice otherwise. In the table, the results after the first round of abductive assembly and lexical top-down guidance is denoted as “After 1” and after the second round of abductive assembly and syntactic top-down guidance is denoted as “After 2.”

The table also shows the overall accuracy, the percentage of inputs that were entirely recognized. While the research was intended to demonstrate the improved character accuracy, it was thought that the overall accuracy results might be of interest to the reader. Notice the vast improvement from before top-down guidance to after top-down guidance. As can be seen, in spite of poor initial performance (as low as 41%), CHREC has a 99% character recognition accuracy when top-down guidance is applied.

TABLE III
EXPERIMENTAL RESULTS

Domain	Top-down Guidance	Character Accuracy	Overall Accuracy
Equations	Before	87%	60%
	After	96%	94%
Addresses	Before	68%	0%
	After	100%	98%
Checks	Before	59%	0%
	After	99%	67%
Sentences	Before	41%	0%
	After 1	97%	76%
	After 2	99%	92%
Overall	After	99%	94%

V. CONCLUSIONS

Accurate automated hand-written character recognition solutions vary from neural network to Bayesian probability to stochastic methods. Few of these approaches take advantage of top-down guidance in their processing by utilizing available domain-specific or pragmatic knowledge. By using such knowledge, it is felt that the accuracy can be improved beyond what such systems can already achieve. In this paper, the CHREC system was described. This knowledge-based approach to hand-written character recognition uses layered abduction to explain the bitmap input in terms of characters, words and sentences. By using abduction, domain-specific and pragmatic knowledge can be applied in the form of expectations and incompatibilities. These expectations and incompatibilities supply top-down guidance, which has allowed CHREC to improve its recognition accuracy from as low as 41% up to 99%.

Four different domains were implemented so that CHREC could use domain-specific knowledge. For mathematical equations, the left-hand side value has to equate to the right-hand side value. CHREC used a partially recognized number to generate expectations to help recognize characters that were not already explained. For bank checks, CHREC compared the numerical value to the value written in English. While some words were not well recognized, the ability to compare one fairly accurate value with one inaccurate value allowed CHREC to modify its solution into a correct one. For postal addresses (city names, state abbreviations and zip codes), a database of 27,000 entries was used to determine the closest matching entry to what CHREC had partially recognized. Finally, for English sentences, CHREC used first a database of 500 words to recognize the various words, and then a small syntax to determine if the words selected were accurate by comparing the words against legal grammatical categories. In every domain, prior to the top-down guidance, character accuracy was poor (87% at best, 41% at worst). However, after the application of top-down guidance, character accuracy was uniformly high, 99% at least.

It should be noted that the authors are not necessarily advocating that hand-written character recognition need be solved by a knowledge-based approach or layered abduction. It is merely the claim that domain-specific and pragmatic knowledge can be applied to improve performance. A knowledge-based approach lends itself easily to layered abduction whereas a neural network approach does not. But in any event, this approach is worth noting because of the accuracy achieved.

REFERENCES

[1] H.F. Herbert, *The History of OCR, Optical Character Recognition.* Manchester Center, VT: Recognition Technologies Users Association, 1982.

[2] M. A. Otair and W. A. Salameh, "An improved back-propagation neural network using a modified non-linear function," *Proc. of the IASTED Intl. Conf.*, 2004, pp. 442-447.

[3] G. Mayraz and G. E. Hinton, "Recognizing hand-written digits using hierarchical products of experts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol 24, no. 2, Feb 2002, pp. 189-197.

[4] L. S. Oliveira, R. Sabourin, F. Bortolozzi and C. Y. Suen, "Feature selection using multi-objective genetic algorithms for handwritten digit recognition," *16th Int'l. Conf. on Pattern Recognition*, vol 1. 2002, pp. 572-575.

[5] K. F. Chan and D. Y. Yeung, "An efficient syntactic approach to structural analysis of on-line handwritten mathematical expressions," *Pattern Recognition*, vol 33. no. 3, 2000, pp. 375-384.

[6] E. H. Ratzlaff, "Methods, report and survey for the comparison of diverse isolated character recognition results on the UNIPEN database," *Int'l. Conf. on Document Analysis and Recognition*, vol. 2, 2003, pp. 623-628.

[7] S. Tanner, *Deciding whether Optical Character Recognition is feasible.* London: King's Digital Consultancy Services, 2004.

[8] L. D. Erman and V. R. Lesser, "The Hearsay-II speech recognition system: A tutorial," in W. Lea (editor), *Trends in Speech Recognition*: Englewood Clifrs, NJ: Prentice-Hall, 1980.

[9] J. Josephson and S. Josephson, eds., *Abductive Inference: Computation, Philosophy, Technology*, Cambridge University Press, New York, 1994.

[10] H. Pople, "The formation of composite hypotheses in diagnostic problem solving: An exercise in synthetic reasoning," *Proc. of IJCAI Five*, San Franscisco, 1977, pp. 1030-1037.

[11] V. Dasigi, R. C. Mann and V. A. Protopopescu, "Information fusion for text classification - an experimental comparison," *Pattern Recognition*, vol. 34, issue 12, pp. 2413-2425, 2001.

[12] P. Thagard, "Explanatory Coherence," *Behavioral and Brain Sciences*, vol 12, issue 3, 1989.

[13] J. Josephson, M. Tanner, J. Svirbely and P. Strohm, "Red: Integrating generic tasks to identify red-cell antibodies," *Proc. of the Expert Systems in Government Symposium*, K. N. Karna, Ed, 1985, pp. 524-531, IEEE Computer Society Press.

[14] R. Fox and J. Hartigan, "An algorithm for abductive inference in artificial intelligence," *Encyc. of Library and Information Science*, vol 64, A. Kent, Ed., Marcel Dekker, Inc: New York, 1999, pp. 22-38.

[15] J. R. Hobbs, M. Stickel, D. Appelt and P. Martin, "Interpretation as abduction," *Artificial Intelligence J.*, vol. 63, issue 1-2, pp. 69-142.

[16] J. Reggia, "Diagnostic expert systems based on a set covering model." Internat. *J. Man-Machine Stud.*, vol. 19, 437-460, Nov 1983.

[17] J. Pearl, *Probabilistic reasoning in intelligent systems: Networks of plausible inference*, Morgan Kaufmann, 1988.

[18] T. Bylander and S. Mittal, "CSRL: A language for classificatory problem solving and uncertainty handling," in *AI Magazine*, 7(3):66-77, August 1986.

[19] J. Josephson, "A layered abduction model of perception: Integrating bottom-up and top-down processing in a multi-sense agent" in the *Proc. of the NASA conf. on Space Telerobotics*, 1989, pp. 197-206, Pasadena: JPL publication.

[20] R. Fox and W. Hartmann, "Hand-written Character Recognition Using Layered Abduction," in *The Proceedings of the International Conference on Systems, Computing Sciences and Software Engineering*, 2005, electronic proceedings article #23, IEEE Publication.

[21] R. Fox and W. Hartmann, "An Abductive Approach to Hand-written Character Recognition for Multiple Domains," in *The Proceedings of the 2006 International Conference on Artificial Intelligence*, Volume II, p. 349-355, H. Arabnia editor, 2006, CSREA Press.

A Novel Utility Sentient Approach For Mining Interesting Association Rules

Shankar.S and Dr.T.Purusothaman

Abstract—**Utility-based data mining is a new research area that concentrates on all types of utility factors in data mining processes and is targeted at incorporating utility considerations in both predictive and descriptive data mining tasks. Discovering interesting association rules that are utilized to improve the business utility of an enterprise has long been recognized in data mining community. This necessitates identifying interesting association patterns that are both statistically and semantically important to the business utility. Classical association rule mining techniques are capable of identifying interesting association patterns but they have failed to associate the user's objective and utility in mining. In this paper, we have proposed an approach for mining novel interesting association patterns from transaction data items of an enterprise to improve its business utility. The approach mines novel interesting association patterns by providing importance to significance, utility and subjective interestingness of the users. The novel interesting patterns mined using proposed approach can be used to provide valuable suggestions to the enterprise to improve its business.**

Index Terms— **Data Mining, Frequent Patterns, FP-growth, Association Rules, Economic Utility, Significance, Subjective Interestingness.**

I. INTRODUCTION

DATA Mining is a remarkable field of contemporary research and development with regard to computer science that is alluring greater interest from a huge variety of people. The chief motivation for data mining stems from the decision support problems that troubled a majority of business organizations [1, 3]. Data mining, also known as Knowledge Discovery in Databases, has been defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [2]. Machine learning and various statistical and visualization techniques are utilized by Data Mining in order to ascertain and represent knowledge in an easily interpretable form [4]. Descriptive mining and Predictive mining are the two customary classifications of data mining tasks. The information that is 'mined' is expressed as a model of the semantic structure of the dataset, where in the prediction or classification of the obtained data is facilitated with the aid of the model [5]. Recently, incorporating utility considerations in data mining tasks is gaining popularity in data mining research. One topic that includes all the perspectives of economic utility in Data Mining and that is destined at assimilating utility considerations for predictive mining tasks and descriptive mining tasks as well.

With regard to variety of applications in marketing and retail communities and other areas as well, Association Rule Mining (ARM) is considered to be the most ubiquitous of the descriptive data mining techniques [10]. "Market Basket Analysis", a study on the buying habits of customers, was the initial motivation behind ARM [9]. Extracting remarkable correlations, frequent patterns, association or untailored structures amongst an ensemble of items from the transaction databases or other data repositories is the primary intent of ARM [11]. The objective of the discovery is not pre-determined in the case of the association rule mining and it is capable of identifying all the association rules that exist in the database. This serves as the chief strength of association rule mining. The discovery of association rules can aid in the extension of marketing and placement strategies and in the planning of logistics for inventory management as well.

With the computers and e-commerce gaining recognition far and wide, the availability of transactional databases is on an all time high. Association Rule Mining and determining the correlation of the items present in the transaction records are the chief points of focus with regard to Data Mining on transactional databases. Since qualitative properties such as significance, utility etc., are necessary to completely utilize the attributes in the dataset, the researchers from the data mining community are anxious about these qualitative aspects of attributes in comparison to considering the quantitative ones (e.g. number of appearances in a database etc). Discovering interesting association rules, which are used to improve the business utility of an enterprise, has long been recognized in data mining community. This necessitates identifying interesting association patterns that are both statistically and semantically important to the business utility.

Even though several algorithms are available in the literature for association rule mining, [12-20] a good number of them deal with efficient implementations rather than the production of effective rules [11, 16, 18]. The techniques that aid in the extraction of suitable and genuine association patterns are mostly quantitative in nature [10, 12, 13, 17]. In order to completely utilize the attributes of a dataset the qualitative attributes are necessary. In addition, some methods are available in the literature for mining weighted association rules [22, 23]. Incorporating utility in item-set mining has gained popularity in recent years [7, 8]. In our work, we have focused on user's objective and its utility which is not reported in the literate of classical association rule mining techniques. In addition to user's objective and its utility, we have also chosen another parameter, the subjective interestingness. The main goal of our approach is to identify novel and interesting association patterns from the historical buying patterns of an enterprise to ascend its business further. The proposed approach aims to identify interesting patterns that are both statistically and semantically important to improve the business utility.

The rest of the paper is organized as follows: Section 2 presents a brief review of the algorithms available in the literature for association rule mining. The proposed approach for mining novel interesting association patterns is discussed in detail in Section 3. The experimental results are presented in Section 4 and conclusions are summed up in Section 5.

II. Related Work

A variety of previous works in the field of Association Rule Mining and Utility based Data Mining serve as the inspiration behind the proposed methodology.

Rakesh Agarwal et al. [9] have proposed an efficient algorithm that generates all significant association rules between items in the database. Their algorithm incorporates buffer management and novel estimation and pruning techniques. Ramakrishnan Srikant et al. [10] have introduced the problem of mining association rules in large relational tables containing both quantitative and categorical attributes. They have also introduced a measure of partial completeness which quantifies the information lost due to partitioning.

Charu C. Aggarwal et al. [13] have provided a survey of research on association rule generation. They have discussed a number of variations of the association rule problem which have been proposed in the literature and their practical applications. C.H. Cai et al. [22] have proposed a method of mining weighted association rule. They have provided two different definition of weighted support: without normalization and with normalization and proposed a new algorithm based on the support bounds.

Wei Wang et al. [23] have proposed a two-fold approach, where the frequent item sets were first generated and then the maximum weighted association rules were derived using an "ordered" shrinkage approach. Frans Coenen and Paul Leng [14] have proposed a method for identifying frequent sets, which reduces the task by means of an efficient restructuring of the data accompanied by a partial computation of the totals required.

Liang Dong and Christos Tjortjis [17] have proposed an enhancement with a memory efficient data structure of a quantitative approach to mine association rules from data. They have combined the best features of the three algorithms (the Quantitative Approach, DHP, and Apriori) in their approach. Ferenc Bodon [11] has described an implementation of APRIORI algorithm, which outperformed all implementations available.

Chuan Wang and Christos Tjortjis [18] have proposed an efficient algorithm for mining association rules, which first identifies all large itemsets and then generates association rules. Their approach has reduced large itemset generation time, known to be the most time-consuming step, by scanning the database only once and using logical operations in the process. Verma Keshri et al. [19] have proposed a novel algorithm to find association rule on time dependent data using efficient T-tree and P-tree data structures. The algorithm has elaborated the significant advantage in terms of time and memory while incorporating time dimension.

Yubo Yuan and Tingzhu Huang [20] have proposed an algorithm for efficient generation of large frequent candidate sets, called Matrix Algorithm. The algorithm generated a matrix which entries 1 or 0 by passing over the cruel database only once, and then the frequent candidate sets were obtained from the resulting matrix. Finally association rules were mined from the frequent candidate sets. Girish K. Palshikar et al. [21] have proposed the concept of heavy itemset, which compactly represents an exponential number of rules. They have provided an efficient theoretical characterization of a heavy itemset. They have also presented an efficient greedy algorithm to generate a collection of disjoint heavy itemsets in a given transaction database.

Jieh-Shan Yeh et al. [7] have proposed a novel utility-frequent mining model to identify all itemsets that can generate a user specified utility in transactions, in which the percentage of such transactions in database is not less than a minimum support threshold. They have proposed a bottom-up two-phase algorithm, BU-UFM, for efficiently mining utility-frequent item sets. A top-down two-phase algorithm, TD-UFM, for mining utility-frequent item sets is also presented. Vid Podpe¡can et al. [8] have proposed a novel efficient algorithm FUFM (Fast Utility-Frequent Mining) which finds all utility-frequent itemsets within the given utility and support constraints threshold.

III. Novel Interesting Association Rule mining

In this section, we have presented our approach for association rule mining. The principal objective of our research is to mine novel interesting association patterns from the transaction data items of an enterprise to improve its business. The transaction data items depict the buying patterns of customers. From the historical buying patterns, our approach aims to mine novel interesting association patterns which are not present in transactions, but are used to ascend the enterprise's business more. In our approach, significance, utility and subjective interestingness of the users are taken into consideration along with the frequency of items. The significance of item is taken into consideration because each item in transactions will have different significance of importance. The utility of an item is important because it contains information about subjectively defined utility such as profit in dollars or some other variety of utility.

The quantity of 'interest' that a pattern evokes upon inspection is captured by the Interestingness measures. Since the Interesting patterns offer ample opportunities for express action which in turn accounts for profitable results, they are considered vital in data mining. Basically, the interestingness of an association pattern is prejudiced. Interestingness plays a vital role for the efficacy of an enterprise as well. Generally, the interestingness measures can be classified into objective and subjective measures. In case of the objective measures the interestingness is expressed with the aid of statistical or mathematical criteria, whereas in subjective measures, more practical criteria such as the suddenness or applicability are considered.

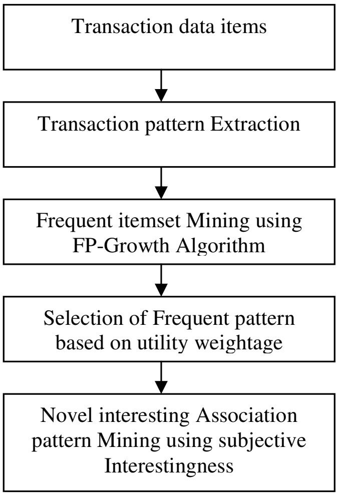

Transaction data items
Transaction pattern Extraction
Frequent itemset Mining using FP-Growth Algorithm
Selection of Frequent pattern based on utility weightage
Novel interesting Association pattern Mining using subjective Interestingness

Fig 1: Block diagram of Proposed Approach

The above figure shows the block diagram of our approach. The major steps involved in our approach are as follows. Initially, the significance weightage of all the items with regard to profit is calculated. The transaction data items are initially scanned for frequent item sets using FP-Growth algorithm. After that, the frequent patterns with utility weightage greater than a threshold value are chosen. As a final point, novel interesting association patterns are mined from the selected patterns based on the subjective interestingness of the users.

A. Association Rule Mining

This sub-section briefly explains the generalized association rule mining. The buying patterns of customers from the basket data can efficaciously be mined by employing Association rules. Support and confidence measures serve as the basis for customary techniques in association rule mining. The task of mining association rules is defined as follows:

Let IS $=\{i_1,i_2,i_3,\ldots,i_m\}$ a set of items and TDI $=\{t_1,t_2,t_3,\ldots,t_n\}$be a set of transaction data items, where $t_i = \{IS_{i1}, IS_{i2,} IS_{i3,\ldots,} IS_{ip}\}$, p<= m and IS_{ij} Є IS, if $X \subseteq I$ with k = |X| is called a k-item set or simply an item set. An expression, where X, Y are item sets and $X \cap Y = \varnothing$ holds is called an association rule X => Y.

The measure of number of transactions T supporting an item set X with respect to TDI is termed as the Support of an item set.

$$\text{Support}(X) = | \{T \in TDI \mid X \subseteq T\} / \; TDI \,|$$

The ratio of the number of transactions that hold X U Y to the number of transactions that hold X is said to be the confidence of an association rule X => Y

$$\text{Conf}\ (X \rightarrow Y) = \text{Supp}\ (X\ U\ Y) / \text{Supp}\ (X)$$

B. Significance Weightage Calculation

The first step of our approach is significance weightage calculation. The significance weightage of each item with respect to profit is calculated using the following equation.

$$SW = P \,/ \textstyle\sum P_i\ ,\ i=1 \text{ to } n$$

Where n is the number of items and P is the profit of an item. Once the weightage is calculated, the data items in the transaction database are represented as a matrix for pattern extraction as follows.

$$T_{ij} = \begin{bmatrix} t_{11}\ \mathrm{t}_{12}\ \mathrm{t}_{13} \ldots t_{1m} \\ t_{21}\ \mathrm{t}_{22}\ \mathrm{t}_{23} \ldots t_{2m} \\ \mathrm{M} \\ t_{n1}\ \mathrm{t}_{n2}\ \mathrm{t}_{n3} \ldots t_{nm} \end{bmatrix}$$

Using the above matrix, the transaction patterns are extracted as a set as follows: where Δ_I represents the item index of all elements in T_{ij} .

$$T_i = \{\Delta_I\,(t_{ij}) : t_{ij} \neq 0\ \}\ ,\ i = 1,\ldots, n \text{ and } j = 1,\ldots, m$$

$$T_n \ll T_i$$

The extracted transaction patterns Tn are then fed as input to FP-Growth algorithm to find frequent patterns.

C. FP- Growth Algorithm

One of the contemporary approaches for frequent item set mining is the FP-growth algorithm [24]. A prefix tree representation of the given database of transactions (called an FP-tree) serves as the basis for the FP-growth algorithm. This can save remarkable amounts of memory for hosting the transactions.

The FP growth algorithm begins with the building of a memory structure called FP-tree. Once the FP-tree is built, the actual FP-growth procedure is recursively applied to it. All the frequent item sets are discovered by this process in a depth-first manner by analyzing projections (conditional FP-trees) of the tree with regard to the frequent prefixes found so far.

D. Frequent Pattern Selection Based On Utility Weightage

In this sub-section, we have presented the selection of frequent patterns based on utility weightage. The frequent patterns extracted using FP-Growth algorithm, are given as input to this stage. The utility weightage of each frequent pattern is calculated using the following equation.

$$Wu = \sum_{i=1}^{n} f * SW_i$$

Where Wu represents the utility weightage, f represents the frequency of pattern and SW_i represent the significance weightage of i^{th} item in the current pattern. Then a set of patterns having utility weightage greater than a predefined threshold value α are chosen as follows.

$$S_P = \{\ F_P : P(F_p)\ \}$$

$$P(F_p) = Wu > \alpha$$

Where S_P is the selected patterns set, α is the predefined threshold value for the selection of patterns and Wu is the utility weightage.

E. INTERESTING ASSOCIATION RULE MINING

The novel interesting association patterns are mined from the selected frequent patterns using the algorithm described in this sub-section. The algorithm implicitly utilized the subjective interestingness of users in interesting pattern mining. The selected frequent patterns from previous sub-section and the output of FP-Growth algorithm are given as input to this algorithm. The algorithm mines novel interesting association patterns, which is not present in transactions, but used to ascend the business. Initially, for every selected pattern, the items are discretely considered and further the patterns with occurrence of each of the item are identified from the frequent patterns. The set difference of the selected item set (pattern) and the identified item set is obtained. The items in the resulting patterns are discretely considered and paired with its reference item and taken into consideration. For every selected pattern, we can get a set of patterns, each with two items. The pseudo code for the above operation is given below.

Assumptions:

Sp_{ij} → Individual item in a selected pattern
Sp_i → A pattern in selected patterns
Fpg → Frequent itemsets from FP-Growth algorithm
RP → Resulting Patterns

Pseudo Code:

```
for each selected pattern
            for each item in selected pattern
                    for each frequent pattern
                        If $Sp_{ij} \subset \mathrm{Fpg}_i$
                        $fp << \mathrm{Fpg}_i \setminus \mathrm{Sp}_i$;
end if
TP1 << $Sp_{ij}$ || fp;
                    end for
TP2 << TP1;
            end for
            RP << TP2;
end for
```

From the identified interesting pattern set of a selected pattern, the patterns with equal second item are considered as a group and its weightage is calculated by adding the weightage of patterns present in it. The weightage of a pattern is calculated by the summation of its item's weightages. By multiplying the frequency with the significance weightage, an item's weightage is calculated. The frequency of an item is the count of that item in transactions having both the items in the pattern. The frequency of item X in pattern X → Y is calculated as follows:

$$C=\sum_{i=1}^{n} tc(X) \quad \text{Where tc(XY)} \neq 0$$

In the above equation tc(X) is the count of item X in current transaction and n is the total number of transactions. The interestingness weightage of a pattern group is calculated by the following equation.

$$Iw=\sum_{\mathrm{j=i}}^{\mathrm{m}}\left(\sum_{\mathrm{i=0}}^{1} C_i W_i\right)$$

Where Iw represents interestingness weightage and m represents the no of patterns in a group. Afterwards, the second item from each group is combined with its corresponding selected pattern and the interestingness weightage of that group is assigned to the newly formed pattern. The newly formed patterns are sorted based on their interestingness weightage. If more than one pattern occurs with same last item, the pattern with highest interestingness weightage is only considered for further process. The patterns with interestingness weightage greater than a threshold value ψ are chosen. The resulting patterns are the novel interesting association patterns mined using our approach.

IV. EXPERIMENTAL RESULTS

In this section, we have presented the analysis of our experimental results. The proposed algorithm is implemented in Java.

Table 1 details the profit and weightage of the sample items taken into consideration. A set of sample transactions are shown in table 2. Table 1 and table 2 are fed as inputs to the various algorithms mentioned in section 3.3, 3.4 and 3.5. The frequencies of occurrence of the frequent patterns are tabulated in table 3. Table 4 shows the utility weightage of each of the patterns. Table 5 contains the patterns with weightage greater than 2.0. The interestingness weightage of the patterns where in the items of the selected pattern occur are tabulated in table 6. From patterns in table 6 having an identical last item the pattern with the greatest interestingness weightage alone is considered and the rest are dropped. Table 8 shows the patterns further filtered from table 7 with the help of a minimum threshold, 10.0.

INPUT TABLE-1
WEIGHTAGEITEMS

ID	ITEM	PROFIT	Significance_weightage
1	A	60	0.285714285714286
2	B	10	4.76190476190476E-02
3	C	30	0.142857142857143
4	D	90	0.428571428571429
5	E	20	9.52380952380952E-02

INPUT TABLE-2
CUSTOMERSTRANSACTIONS

TID	A	B	C	D	E
1	10	1	4	1	0
2	0	1	0	3	0
3	2	0	0	1	0
4	0	0	1	0	0
5	1	2	0	1	3
6	1	1	1	1	1
7	0	2	3	0	1
8	0	0	0	1	2
9	7	0	1	1	0
10	0	1	1	1	1

TABLE-3
FP-GROWTH

Patterns	Frequent
A D	5
D	8
A C D	3
B D	5
A B D	3
C D	4
E D	4
C B D	3
E B D	3
A	5
A C	3
A B	3
C	6
E C B	3
C B	4
E C	3
E B	4
E	5
B	6

TABLE-4
FINDING UTILITY WEIGHTAGE

Patterns	Frequency	utility_weightage
A D	5	3.571428571428571
D	8	3.4285714285714284
A C D	3	2.571428571428571
B D	5	2.380952380952381
A B D	3	2.2857142857142856
C D	4	2.2857142857142856
E D	4	2.095238095238095
C B D	3	1.857142857142857
E B D	3	1.7142857142857142
A	5	1.4285714285714284
A C	3	1.2857142857142856
A B	3	1.0
C	6	0.8571428571428571
E C B	3	0.857142857142857
C B	4	0.7619047619047619
E C	3	0.7142857142857142
E B	4	0.5714285714285714
E	5	0.47619047619047616
B	6	0.2857142857142857

TABLE-5
SELECTED PATTERNS FROM TABLE-4 WHERE UTILITY_WEIGHTAGE>2.0

Patterns	Frequency	utility_weightage
A D	5	3.571428571428571
D	8	3.4285714285714284
A C D	3	2.571428571428571
B D	5	2.380952380952381
A B D	3	2.2857142857142856
C D	4	2.2857142857142856
E D	4	2.095238095238095

TABLE-6
CALCULATING INTERESTINGNESS WEIGHTAGE

group	Interestingness _weightage
C D A	14.142857142857142
B D A	11.761904761904761
A B D C	10.238095238095237
A D C	8.714285714285714
A C D B	8.428571428571427
E D A	8.142857142857142
D A	8.142857142857142
A D B	6.904761904761904
C D B	4.809523809523809
B D C	4.238095238095238
E D B	4.142857142857142
E D C	3.7142857142857144
C D E	3.380952380952381
A C D E	3.380952380952381
D B	3.2857142857142856
A B D E	3.238095238095238
B D E	3.238095238095238
D C	2.7142857142857144
D E	2.380952380952381
A D E	2.380952380952381

TABLE-7
SELECTED PATTERNS FROM TABLE-6

group	Interestingness _weightage
C D A	14.142857142857142
A B D C	10.238095238095237
A C D B	8.428571428571427
C D E	3.380952380952381

TABLE-8
NOVEL INTERESTING PATTERNS

group	Interestingness_weightage
C D A	14.1428571428571
A B D C	10.2380952380952

If items C, D and A tend to go together then it is better to place these items side by side. The items such as A which have close proximity will enhance the chances of a customer who comes to buy only C, D buying A as well. For example a person, who comes to buy a Jeans, T. Shirt, may buy a belt, or perfumes or Toileteries, if they are placed close by. In short, when similar items are placed adjacently the chances of a customer buying the second item as well are much higher even when he/she had no intention of buying it initially.

V. CONCLUSION

Association rules have been widely used to determine customer buying patterns from market basket data. Incorporating utility considerations in association rule mining has gained popularity in recent years. Discovering interesting association rules, used to improve business utility of an enterprise has long been recognized in data mining community. This necessitated the identification of interesting association patterns that are both statistically and semantically important to the business utility. In this paper, we have presented a novel approach for mining interesting association patterns to improve the business utility. The proposed approach focuses on utility, significance and interestingness in the mining of interesting association patterns. The mined interesting association patterns have been used in providing valuable recommendation to the enterprise in intensifying its business utility.

REFERENCES

[1] S. Tsur, “Data dedging,” IEEE Data Engineering Bulletin, 13(4):58-63, December 1990.

[2] Frawley, W., Piatetsky-Shapiro, G., Matheus, C. (1992) Knowledge Discovery in Databases: An Overview. AI Magazine, Fall 1992, pp. 213-228.

[3] J. T-L. Wang, G-W. Chirn, T. G. Marr, B. Shapiro, D. Shasha, and Ii. Zhang, “Combinatorial pattern discovery for scientific data: some preliminary results,” In Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data, pages 115-125, Minneapolis, MN, May 24-27 1994.

[4] Soundararajan E., Joseph J.V.M., Jayakumar C. and Somasekharan M., "Knowledge Discovery Tools and Techniques," In Proceedings of the Conference on Recent Advances in Information Technology, IGCAR, pp.141 -145, July 14-15, 2005.

[5] Cunningham, S. J. and Holmes, G., "Developing innovative applications in agriculture using data mining," In the Proceedings of the Southeast Asia Regional Computer Confederation Conference, Singapore, 1999.

[6] Weiss G., Zadrozny B., Saar-Tsechansky M.: Utility-based data mining 2006 workshop report. SIGKDD Explorations, volume 8, issue 2.

[7] Jieh-Shan Yeh, Yu-Chiang Li, and Chin-Chen Chang, "Two-Phase Algorithms for a Novel Utility-Frequent Mining Model," PAKDD 2007, Lecture Notes in Computer Science, vol. 4819 ,Springer Berlin / Heidelberg, pp. 433–444, 2007.

[8] Vid Podpecan, Nada Lavrac, and Igor Kononenk, "A fast algorithm for mining utility-frequent itemsets," The 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, September 21, 2007.

[9] Agrawal, R., Imielinski, R., Swami, A., 1993, “Mining Associations between Sets of Items in Large Databases”, Proceedings of the ACM SIGMOD, Washington, DC, pp.207-216, May 1993.

[10] Srikant, R. and Agrawal, R., “Mining Quantitative Association Rules in Large Relational Tables.” In Proc. of ACM SIGMOD Conf. on Management of Data. ACM Press, (1996), 1-12.

[11] Bodon, F.: “A Fast Apriori implementation”, In ICDM Workshop on Frequent Itemset Mining Implementations, vol. 90, Melbourne, Florida, USA (2003).

[12] Sotiris Kotsiantis and Dimitris Kanellopoulos, “Association Rules Mining: A Recent Overview," GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82.

[13] C. C. Aggrawal and P. S. Yu. Mining large itemsets for association rules. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 21(1): 23-31, March 1998.

[14] F. Coenen and P. Leng. Optimising Association Rule Algorithms Using Itemset Ordering. Research and Development in Intelligent Systems XVIII: Proc ES2001 Conference, Springer, pages 53-66, 2001.

[15] F. Coenen and P. Leng. Finding Association Rules with Some Very Frequent Attributes. In Proceedings of PKDD2002, LNAI 2431, pages 99-111, 2002.

[16] Lee, C.-H., Chen, M.-S., Lin, C.-R. , “Progressive Partition Miner: An Efficient Algorithm for Mining General Temporal Association Rules,” IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 4, (2003), 1004 - 1017.

[17] L. Dong and C. Tjortjis, “Experiences of Using a Quantitative Approach for Mining Association Rules,” Lecture Notes in Computer Science Series, Vol. 2690, Springer-Verlag, (2003), 693-700.

[18] Wang, C., Tjortjis, C., “PRICES: An Efficient Algorithm for Mining Association Rules,” Lecture Notes in Computer Science, Volume 3177, Jan 2004, Pages 352 – 358.

[19] Verma, K., Vyas, O.P., Vyas, R., Temporal Approach to Association Rule Mining Using T-Tree and P-Tree, Lecture Notes in Computer Science, Volume 3587, Jul 2005, Pages 651 – 659.

[20] Yuan, Y., Huang, T., A Matrix Algorithm for Mining Association Rules, Lecture Notes in Computer Science, Volume 3644, Sep 2005, Pages 370 – 379.

[21] GK Palshikar, MS Kale, MM Apte, Association rules mining using heavy itemsets, Data and Knowledge Engineering, Vol. 61, No. 1, pp. 93-113, 2007.

[22] Cai, C.H., Fu, A.W-C., Cheng, C. H., Kwong, W.W.: Mining Association Rules with Weighted Items. In: Proceedings of 1998 Intl. Database Engineering and Applications Symposium (IDEAS'98), pages 68--77, Cardiff, Wales, UK, July 1998

[23] Wang, W., Yang, J., Yu, P. S.: Efficient Mining of Weighted Association Rules (WAR). In: Proceedings of the KDD, Boston, MA, August 2000, pp. 270-274

[24] Han, J., Pei, J., and Yin, Y., “Mining frequent patterns without candidate generation,” In Proceedings of 2000 ACMSIGMOD International Conference on Management of Data (SIGMOD’00), pp. 1–12, Dallas, TX, 2000.

Multicast Extensions to the Location-Prediction Based Routing Protocol for Mobile Ad hoc Networks

Natarajan Meghanathan

Abstract— **We propose multicast extensions to the location prediction-based routing protocol (referred to as NR-MLPBR and R-MLPBR) for mobile ad hoc networks to simultaneously reduce the number of tree discoveries and the hop count per path from the source to each of the receivers of the multicast group. Nodes running NR-MLPBR are not aware of the receivers of the multicast group. R-MLPBR assumes that each receiver node also knows the identity of the other receiver nodes of the multicast group. The multicast extensions work as follows: Upon failure of a path to the source, a receiver node attempts to locally construct a global topology using the location and mobility information collected during the latest global broadcast tree discovery. NR-MLPBR attempts to predict a path that has the minimum number of hops to the source and R-MLPBR attempts to predict a path to the source that has the minimum number of non-receiver nodes. If the predicted path exists in reality, the source accommodates the path as part of the multicast tree and continues to send the multicast packets in the modified tree. Otherwise, the source initiates another global broadcast tree discovery. Simulation studies illustrate that NR-MLPBR and R-MLPBR simultaneously minimize the number of global broadcast tree discoveries as well as the hop count per source-receiver path in the multicast trees. In addition, R-MLPBR determines multicast trees with relatively reduced number of links.**

Index Terms— **Multicast Routing, Mobile Ad hoc Networks, Link Efficiency, Hop Count, Simulation**

I. Introduction

A mobile ad hoc network (MANET) is a dynamic distributed system of wireless nodes that move independent of each other in an autonomous fashion. The network bandwidth is limited and the medium is shared. As a result, transmissions are prone to interference and collisions. The battery power of the nodes is constrained and hence nodes operate with a limited transmission range, often leading to multi-hop routes between any pair of nodes in the network. Due to node mobility, routes between any pair of nodes frequently change and need to be reconfigured. As a result, on-demand route discovery (discovering a route only when required) is often preferred over periodic route discovery and maintenance, as the latter strategy will incur significant overhead due to the frequent exchange of control information among the nodes [1]. We hence deal with on-demand routing protocols for the rest of this paper.

In an earlier work [2], we developed a location prediction based routing (LPBR) protocol for unicast routing in MANETs. The specialty of LPBR is that it attempts to simultaneously reduce the number of global broadcast route discoveries as well as the hop count of the paths for a source-destination session. LPBR works as follows: During a regular flooding-based route discovery, LPBR collects the location and mobility information of the nodes in the network and stores the collected information at the destination node of the route search process. When the minimum-hop route discovered through the flooding-based route discovery fails, the destination node attempts to predict the current location of each node using the location and mobility information collected during the latest flooding-based route discovery. A minimum hop path Dijkstra algorithm [3] is run on the locally predicted global topology. If the predicted minimum hop route exists in reality, no expensive flooding-based route discovery is needed and the source continues to send data packets on the discovered route; otherwise, the source initiates another flooding-based route discovery.

Multicasting is the process of sending a stream of data from one source node to multiple recipients by establishing a routing tree, which is an acyclic connected subgraph containing all the nodes in the tree. The set of receiver nodes form the multicast group. While propagating down the tree, data is duplicated only when necessary. This is better than multiple unicast transmissions. Multicasting in ad hoc wireless networks has numerous applications [4]: collaborative and distributing computing like civilian operations, emergency search and rescue, law enforcement, warfare situations and etc.

Several MANET multicast routing protocols have been proposed in the literature [4]. They are mainly classified as: tree-based and mesh-based protocols. In tree-based protocols, only one route exists between a source and a destination and hence these protocols are efficient in terms of the number of link transmissions. The tree-based protocols can be further divided into two types: source tree-based and shared tree-based. In source tree-based multicast protocols, the tree is rooted at the source. In shared tree-based multicast protocols,

This research has been funded by the Army Research Lab through grant number W911NF-08-2-0061.

Natarajan Meghanathan is with the Department of Computer Science, Jackson State University, Jackson, MS 39056 USA (phone: 601-979-3661; fax: 601-979-2478; e-mail: nmeghanathan@jsums.edu).

the tree is rooted at a core node and all communication between the multicast source and the receiver nodes is through the core node. Even though shared tree-based multicast protocols are more scalable with respect to the number of sources, these protocols suffer under a single point of failure, the core node. On the other hand, source tree-based protocols are more efficient in terms of traffic distribution. In mesh-based multicast protocols, multiple routes exist between a source and each of the receivers of the multicast group. A receiver node receives several copies of the data packets, one copy through each of the multiple paths. Mesh-based protocols provide robustness at the expense of a larger number of link transmissions leading to inefficient bandwidth usage. Considering all the pros and cons of these different classes of multicast routing in MANETs, we feel the source tree-based multicast routing protocols are more efficient in terms of traffic distribution and link usage. Hence, all of our work in this research will be in the category of on-demand source tree-based multicast routing.

In this paper, we propose two multicast extensions to LPBR, referred to as NR-MLPBR and R-MLPBR. Both the multicast extensions are aimed at minimizing the number of global broadcast tree discoveries as well as the hop count per source-receiver path of the multicast tree. They use a similar idea of letting the receiver nodes to predict a new path based on the locally constructed global topology obtained from the location and mobility information of the nodes learnt through the latest broadcast tree discovery. Receiver nodes running NR-MLPBR (Non-Receiver aware Multicast extensions of LPBR) are not aware of the receivers of the multicast group, whereas each receiver node running R-MLPBR (Receiver-aware Multicast Extension of LPBR) is aware of the identity of the other receivers of the multicast group. NR-MLPBR attempts to predict a minimum hop path to the source, whereas R-MLPBR attempts to predict a path to the source that has the minimum number of non-receiver nodes. If more than one path has the same minimum number of non-receiver nodes, then R-MLPBR breaks the tie among such paths by choosing the path with the minimum number of hops to the source. Thus, R-MLPBR is also designed to reduce the number of links in the multicast tree, in addition to the average hop count per source-receiver path and the number of global broadcast tree discoveries.

The rest of the paper is organized as follows: Section II provides the detailed design of the two multicast extensions. Section III explains the simulation environment and reviews the MAODV and BEMRP protocols that are studied along with NR-MLPBR and R-MLPBR as part of our simulation studies. In Section IV, we illustrate and explain simulation results for the four multicast routing protocols (MAODV, NR-MLPBR, R-MLPBR and BEMRP) with respect to different performance metrics. Section V concludes the paper.

II. MULTICAST EXTENSIONS TO LPBR

The objective of the multicast extensions to LPBR (referred to as NR-MLPBR and R-MLPBR) is to simultaneously minimize the number of global broadcast tree discoveries as well as the hop count per source-receiver path. In addition, R-MLPBR aims to also reduce the number of links that are part of the multicast tree. The Non-Receiver aware Multicast extension to LPBR (NR-MLPBR) does not assume the knowledge of the receiver nodes of the multicast group at every receiver node. Each receiver node running R-MLPBR learns the identity information of peer receiver nodes through the broadcast tree discovery procedure. Both the multicast extensions assume the periodic exchange of beacons in the neighborhood. This is essential for nodes to learn about the moving away of the downstream nodes in the multicast tree. We assume that a multicast group comprises basically of receiver nodes that wish to receive data packets from an arbitrary source, which is not part of the multicast group.

A. *Broadcast of Multicast Tree Request Messages*

Whenever a source node has data packets to send to a multicast group and is not aware of a multicast tree to the group, the source initiates a broadcast tree discovery procedure by broadcasting a Multicast Tree Request Message (MTRM) to its neighbors. The source maintains a monotonically increasing sequence number for the broadcast tree discoveries it initiates to form the multicast tree. Each node, including the receiver nodes of the multicast group, on receiving the first MTRM of the current broadcast process (i.e., a MTRM with a sequence number greater than those seen before), includes its Location Update Vector, LUV in the MTRM packet. The LUV of a node comprises the following: node ID, X, Y co-ordinate information, Is Receiver flag, Current velocity and Angle of movement with respect to the X-axis. The *Is Receiver* flag in the LUV, if set, indicates that the node is a receiving node of the multicast group. The node ID is also appended on the “Route record” field of the MTRM packet. The structure of the LUV and the MTRM is shown in Figures 1 and 2 respectively.

Node ID	X Co-ordinate	Y Co-ordinate	Node Velocity	Angle of Movement	Is Receiver
4 bytes	8 bytes	8 bytes	8 bytes	8 bytes	1 bit

Figure 1: Location Update Vector (LUV) per Node

Multicast Source	Multicast Group ID	Sequence Number	Route Recorded (List of Node IDs)	Location Update Vectors (LUVs)
4 bytes	4 bytes	4 bytes	Variable Size of 4 bytes	Variable Size of 36 bytes, 1 bit

Figure 2: Structure of the Multicast Tree Request Message

Multicast Source	Originating Receiver	Multicast Group ID	Sequence Number	Route Record from the Receiver to the Source
4 bytes	4 bytes	4 bytes	4 bytes	Variable Size of 4 bytes

Figure 3: Structure of Multicast Tree Establishment Message

Key	Value			
<Source, Multicast Group ID>	<d_a, r_a>	<d_b, r_b>	<,,,, …>	<…, …>

Figure 4: Structure of the Multicast Routing Table at an Intermediate Node

B. Construction of the Multicast Tree through the Multicast Tree Establishment Message

Paths constituting the multicast tree are independently chosen at each receiver node. A receiver node gathers several MTRMs obtained across different paths and selects the minimum hop path among them by looking at the “Route Record” field in these MTRMs. A Multicast Tree Establishment Message (MTEM) is sent on the discovered minimum hop route to the source. The MTEM originating from a receiver node has the list of node IDs corresponding to the nodes that are on the minimum hop path from the receiver node to the source (which is basically the reverse of the route recorded in the MTRM). The structure of the MTEM packet is shown in Figure 3.

An intermediate node upon receiving the MTEM packet checks its multicast routing table whether there exist an entry for the <Multicast Source, Multicast Group ID> in the table. If an entry exists, the intermediate node merely adds the tuple <One-hop sender of the MTEM, Originating Receiver node of the MTEM> to the list of <Downstream node, Receiver node> tuples for the multicast tree entry and does not forward the MTEM further. The set of downstream nodes are part of the multicast tree rooted at the source node for the multicast group. If a <Multicast Source, Multicast Group ID> entry does not exist in the multicast routing table, the intermediate node creates an entry and initializes it with the <One-hop sender of the MTEM, Originating Receiver node of the MTEM> tuple. Note that the one-hop sender of the MTEM is learnt through the MAC (Medium Access Control) layer header and verified using the Route Record field in the MTEM. The intermediate node then forwards the MTEM to the next downstream node on the path towards the source. The structure of the multicast routing table at a node is illustrated in Figure 4. Note that the tuples <d_a, r_a>, <d_b, r_b>, <…, …> indicate the downstream node d_a for receiver node r_a, downstream node d_b for receiver node r_b and so on. A node could be the downstream node for more than one receiver node. The source node maintains a multicast routing table that has the list of <Downstream node, Receiver node> tuples for each of the multicast groups to which the source is currently communicating through a multicast session. For each MTEM received, the source adds the neighbor node that sent the MTEM and the corresponding Originating Receiver node to the list of <Downstream node, Receiver node> tuples for the multicast group.

C. Multicast Tree Acquisition and Data Transmission

After receiving the MTEMs from all receiver nodes within a certain time called Tree Acquisition Time (*TAT*), the source starts sending the data packets on the multicast tree. The *TAT* is based on the maximum possible diameter of the network (an input parameter in our simulations). The diameter of a network is the maximum of the hop count of the minimum hop paths between any two nodes in the network. The *TAT* is dynamically set at a node based on the time it took to receive the first MTEM for a broadcast tree discovery procedure.

The structure of the header of the multicast data packet is shown in Figure 5. The source and destination fields in the header include the identification for the source node and the multicast group ID respectively. The sequence number field in the header can be used by the receivers to accumulate and reorder the multicast data packets, incase if they are received out of order. In addition to these regular fields, the header of the multicast data packet includes three specialized fields: the ‘More Packets’ (*MP*) field, the ‘Current Dispatch Time’ (*CDT*) field and the ‘Time Left for Next Dispatch’ (*TNLD*) field. The *CDT* field stores the time as the number of milliseconds lapsed since Jan 1, 1970, 12 AM. These additional overhead (relative to that of the other ad hoc multicast routing protocols) associated with the header of each data packet amounts to only 12 more bytes per data packet.

Multicast Source	Multicast Group ID	Sequence Number	More Packets	Current Dispatch Time	Time Left for Next Dispatch
4 bytes	4 bytes	4 bytes	1 bit	8 bytes	4 bytes

Figure 5: Structure of the Header of the Multicast Data Packet

The source sets the *CDT* field in all the data packets sent. In addition, if the source has any more data to send, it sets the *MP* flag to 1 and sets the appropriate value for the *TLND* field, which indicates the number of milliseconds since the *CDT*. If the source does not have any more data to send, it will set the *MP* flag to 0 and leaves the *TLND* field blank. As we assume the clocks across all nodes are synchronized, a receiver node will be able to calculate the end-to-end delay for the data packet based on the time the data packet reaches the node and the *CDT* field in the header of the data packet. Several clock synchronization algorithms (example [5][6]) have been proposed for wireless ad hoc networks. The receiver node computes and maintains the average end-to-and delay per data packet for the current path to the source by recording the sum of the end-to-end delays of all the data packets received so far on the path and the number of data packets received on the path. Accordingly, the average end-to-end delay per data packet for the current path is updated every time after receiving a new data packet on the path. If the source node has set the MP flag, the receiver node computes the ‘Next Expected Packet Arrival Time’ (*NEPAT*), which is *CDT* field + *TLND* field + 2*Average end-to-end delay per data packet. A timer is started for the *NEPAT* value. Since, we are using only the average end-to-end delay per data packet to measure the *NEPAT* value, the variations in the end-to-end delay of particular data packets will not very much affect the *NEPAT* value. So, the source and receiver nodes need not be perfectly synchronized. The clocks across the nodes can have small drifts and this would not very much affect the performance of the multicast extensions of LPBR.

D. Multicast Tree Maintenance

We assume that each node periodically exchanges beacon messages with its neighbors, located within its default maximum transmission range. If an intermediate node notices that its link with a downstream node has failed (i.e., the two nodes have moved away and are no longer neighbors), the intermediate node generates and sends a Multicast Path Error Message (MPEM) to the source node of the multicast group entry. The MPEM has information about the receiver nodes affected (obtained from the multicast routing table) because of the link failure with the downstream node. Figure 6 shows the structure of an MPEM. The intermediate node removes the tuple(s) corresponding to downstream node(s) and the affected receiver node(s). After these deletions, if no more <Downstream node, Receiver node> tuple exists for a <Source node, Multicast group ID> entry, the intermediate node removes the entire row for this entry from the routing table.

Multicast Source	Originating Intermediate Node	Multicast Group ID	IDs of Affected Receivers
4 bytes	4 bytes	4 bytes	Variable Size of 4 bytes

Figure 6: Structure of a MPEM Message

The source, upon receiving the MPEM, will wait to receive a Multicast Predicted Path Message (MPPM) from each of the affected receivers, within a MPPM-timer maintained for each receiver. The source estimates a Tree-Repair Time (*TRT*) for each receiver as the time that lapsed between the reception of the MPEM from an intermediate node and the MPPM from the affected receiver. An average value for the TRT per receiver is maintained at the source as it undergoes several path failures and repairs before the next global broadcast based tree discovery. The MPPM-timer (initially set to the time it took for the source to receive the MTEM from the receiver) for a receiver will be then set to 1.5* Average *TRT* value, so that we give sufficient time for the destination to learn about the route failure and generate a new MPPM. Nevertheless, this timer will be still far less than the tree acquisition time that would be incurred if the source were to launch a global broadcast tree discovery. Hence, our approach will only increase the network throughput and does not decrease it.

E. Prediction of Node Location using the LUVs

If a multicast receiver does not receive the data packet within the *NEPAT* time, it will attempt to locally construct the global topology using the location and mobility information of the nodes learnt from the latest broadcast tree discovery. Each node is assumed to be moving in the same direction with the same speed as mentioned in its latest LUV. Based on this assumption and information from the latest LUVs, the location of each node at the *NEPAT* time is predicted.

We now explain how to predict the location of a node (say node *u*) at a time instant *CTIME* based on the LUV gathered from node *u* at time *STIME*. Let $(X_u^{STIME}, Y_u^{STIME})$ be the X and Y co-ordinates of *u* at time *STIME*. Let $Angle_u^{STIME}$ and $Velocity_u^{STIME}$ represent the angle of movement with respect to the X-axis and the velocity at which *u* is moving. The distance traveled by node *u* from time *STIME* to *CTIME* would be: $Distance_u^{STIME\text{-}CTIME} = (CTIME - STIME + 1) * Velocity_u^{STIME}$.

Let $(X_u^{CTIME}, Y_u^{CTIME})$ be the predicted location of node *u* at time *CTIME*. The value of X_u^{CTIME} and Y_u^{CTIME} are given by $X_u^{STIME}+Offset\text{-}X_u^{CTIME}$ and $Y_u^{STIME}+Offset\text{-}Y_u^{CTIME}$ respectively. The offsets in the X and Y-axes, depend on angle of movement and the distance traveled, and are calculated as follows:

$$Offset\text{-}X_u^{CTIME} = Distance_u^{STIME\text{-}CTIME} * \cos(Angle_u^{STIME})$$
$$Offset\text{-}Y_u^{CTIME} = Distance_u^{STIME\text{-}CTIME} * \sin(Angle_u^{STIME})$$

$$X_u^{CTIME} = X_u^{STIME} + Offset\text{-}X_u^{CTIME}$$
$$Y_u^{CTIME} = Y_u^{STIME} + Offset\text{-}Y_u^{CTIME}$$

We assume each node is initially configured with information regarding the network boundaries, given by [0, 0], $[X_{max}, 0]$, $[X_{max}, Y_{max}]$ and $[0, Y_{max}]$. When the predicted X and/or Y co-ordinate is beyond the network boundaries, we set their values to the boundary conditions as stated below.

If $(X_u^{CTIME} < 0)$, then $X_u^{CTIME} = 0$;
If $(X_u^{CTIME} > X_{max})$, then $X_u^{CTIME} = X_{max}$
If $(Y_u^{CTIME} < 0)$, then $Y_u^{CTIME} = 0$;
If $(Y_u^{CTIME} > Y_{max})$, then $Y_u^{CTIME} = Y_{max}$

Based on the predicted locations of each node in the network at time *CTIME*, the receiver node locally constructs the global topology. Note that there exists an edge between two nodes in the locally constructed global topology, if the predicted distance between the two nodes (with the location information obtained from the *LUV*) is less than or equal to the transmission range of the nodes. The two multicast extensions NR-MLPBR and R-MLPBR differ from each other on the nature of the paths predicted at the receiver node.

Multicast Source	Originating Receiver Node	Multicast Group ID	Predicted Path to the Multicast Source (List of Node IDs)
4 bytes	4 bytes	4 bytes	Variable Size of 4 bytes

Figure 7: Structure of the Multicast Predicted Path Message

F. NR-MLPBR: Multicast Path Prediction

The receiver node locally runs the Dijkstra's minimum hop path algorithm [3] on the predicted global topology. If at least one path exists from the source node to the receiver node in the generated topology, the algorithm returns the minimum hop path among them. The receiver node then sends a MPPM (structure shown in Figure 7) on the discovered path with the route information included in the message.

G. R-MLPBR: Multicast Path Prediction

The receiver node uses the LUV obtained from each of the intermediate nodes during the latest global tree broadcast discovery to learn about the identification of its peer receiver nodes that are part of the multicast group. If there existed a direct path to the source on the predicted topology, the receiver chooses that path as the predicted path towards the source. Otherwise, the receiver determines a set of node-disjoint paths on the predicted global topology. The node-disjoint paths to the source are ranked depending on the number of non-receiver nodes that act as intermediate nodes on the path. The path that has the least number of non-receiver nodes as intermediate nodes is preferred. The reason is a path that has the least number of non-receiver nodes is more likely to be a minimum hop path and if a receiver node lies on that path, the number of newly added links to the tree would also be reduced. R-MLPBR thus aims to discover paths with the minimum hop count and at the same time attempts to conserve bandwidth by reducing the number of links that get newly added to the tree as a result of using the predicted path. The

MPPM is hence sent on the predicted path that has minimum number of non-receiver nodes. If two or more paths has the same minimum number of non-receiver nodes, R-MLPBR breaks the tie by choosing the path with the minimum hop count to the source. Figure 8 illustrates the algorithm used by R-MLPBR at a receiver node to select the best predicted path to the source.

```
Input: Graph G (V, E), Set of Multicast receivers $M_R$, source s and receiver d
Output: s-d path
Auxiliary Variables: Graph G'' (V'', E''), Set of Node-disjoint paths $P_N$
Initialization: G'' (V'', E'') ← G (V, E), $P_N$ ← φ.

Begin
while ( ∃ at least one s-d path in G'')
    p ← Minimum hop s-d path in G''.
    if (hop count of p = 1)
        return p
    end if
    $P_N$ ← $P_N$ U {p}
    ∀ vertex, v∈p, v≠s,d  edge, e∈Adj−list(v)   G''(V'',E'')←G''(V''-{v},E''-{e})
end while
minNonReceivers ← ∞ // the count for the minimum number of non-receivers is initialized to ∞.
bestPath ←NULL // the best path is initialized to NULL
minHops ← ∞ // the minimum hop count of the best path initialized to ∞ (a very large value).
for (∀ path p∈ $P_N$)
    countPathNonReceivers ← 0 // keeps track of the number of non-receiver nodes in path p
    for (∀ intermediate node n∈ p)
        if (n∉ $M_R$)
            countPathNonReceivers←countPathNonReceivers + 1
        end if
    end for
    if (minNonReceivers ≥ countPathReceivers)
        if (minNonReceivers = countPathReceivers AND minHops > hop count of p)
            bestPath ← p
            minHops ← hop count of p
        end if
        if (minNonReceivers > countPathReceivers)
            minNonReceivers ← countPathReceivers
            bestPath ← p
            minHops ← hop count of p
        end if
    end if
end for
return bestPath
End
```

Figure 8: R-MLPBR Predicted Path Selection Algorithm

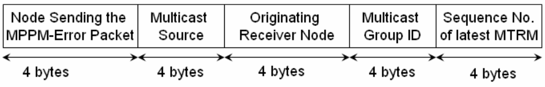
Note that we designed R-MLPBR to choose the path with the minimum number of non-receiver nodes, rather than the path with the maximum number of receiver nodes, as the latter design has the possibility of yielding paths with significantly larger hop count from the source to the receiver node without any guarantee on the possible reduction in the number of links. Our design of choosing the path with the minimum number of non-receiver nodes helps to maintain the hop count per source-receiver path close to that of the minimum hop count and at the same time does helps to reduce the number of links in the tree to a certain extent.

H. Propagation of the Multicast Predicted Path Message towards the Source

An intermediate node on receiving the MPPM, checks its multicast routing table if there already exists an entry for the source node and the multicast group to which the MPPM belongs to. If an entry exists, the intermediate node merely adds the tuple <One-hop sender of the MPPM, Originating Receiver node of the MPPM> to the list of <Downstream node, Receiver node> tuples for the multicast tree entry. If the <Multicast Source, Multicast Group ID> entry does not exist in the multicast routing table, the intermediate node creates an entry and initializes it with the <One-hop sender of the MPPM, Originating Receiver node of the MPPM> tuple. In either case, the MPPM is then forwarded to the next downstream node on the path towards the source. If the source node receives the MPPM from the appropriate receiver node before the MPPM-timer expires, it indicates that the predicted path does exist in reality. A costly global broadcast tree discovery has been thus avoided. The source continues to send the data packets down the multicast tree. The source node estimates the Tree Repair Time (TRT) as the time lapsed between the reception of the MPEM from an intermediate node and the MPPM from the appropriate receiver node. An average value of the TRT for each receiver node is thus maintained at the source as it undergoes several route failures and repairs before the next global broadcast-based tree discovery.

I. Handling Prediction Failure

If an intermediate node attempting to forward the MPPM of a receiver node could not successfully forward the packet to the next node on the path towards the source, the intermediate node informs the absence of the route through a MPPM-Error packet (structure shown in Figure 9) sent back to the receiver node. The receiver node on receiving the MPPM-Error packet discards all the LUVs and does not generate any new MPPM. The receiver will wait for the multicast source to initiate a global broadcast-based tree discovery. After the MPPM-timer expires, the multicast source initiates a new global broadcast-based tree discovery procedure.



Node Sending the MPPM-Error Packet	Multicast Source	Originating Receiver Node	Multicast Group ID	Sequence No. of latest MTRM
4 bytes	4 bytes	4 bytes	4 bytes	4 bytes

Figure 9: Structure of the MPPM-Error Packet

III. SIMULATION ENVIRONMENT AND PROTOCOL REVIEW

The network dimension used is a 1000m x 1000m square network. The transmission range of each node is assumed to be 250m. The number of nodes used in the network is 25 and 75 nodes representing networks of low and high density with an average distribution of 5 and 15 neighbors per node respectively. Initially, nodes are uniformly randomly distributed in the network. We compare the performance of NR-MLPBR and R-MLPBR with that of the minimum-hop based MAODV and the link-efficient BEMRP protocols. We implemented all of these four multicast routing protocols in a discrete-event simulator developed in Java. The broadcast tree discovery strategy employed is the default flooding approach. The simulation parameters are summarized in Table 1.

Table 1: Simulation Conditions

Network Size	1000m x 1000m	
Number of nodes	25 (low density) and 75 (high density)	
Transmission Range	250 m	
Physical Layer	Signal Propagation Model	Two-ray ground reflection model [7]
MAC Layer	IEEE 802.11 [8]	
	Link Bandwidth	2 Mbps
	Interface Queue	FIFO-based, size 100
Routing Protocols	BEMRP [9], MAODV [10], NR-MLPBR and R-MLPBR	
Broadcast Strategy	Flooding	
Mobility Model	Random Way Point Model [11]	
	Minimum Node Speed, m/s	0 m/s

	Maximum Node Speed, m/s	Low-10;
		Medium-30;
		High-50

	Pause Time	0 second
Traffic Model	Constant Bit Rate (CBR), UDP	

	Multicast Group Size (# Receivers)	Small: 2;
		Medium: 4, 8;
		High: 12, 24

	Data Packet Size	512 bytes
	Packet Sending Rate	4 Packets/ second

Simulations are conducted with a multicast group size of 2, 4 (small size), 8, 12 (moderate size) and 24 (larger size) receiver nodes. For each group size, we generated 5 lists of receiver nodes and simulations were conducted with each of them. Traffic sources are constant bit rate (CBR). Data packets are 512 bytes in size and the packet sending rate is 4 data packets/second. The multicast session continues until the end of the simulation time, which is 1000 seconds. The node mobility model used is the Random Waypoint model [11]. The transmission energy and reception energy per hop is set at 1.4 W and 1 W respectively. Initial energy at each node is 1000 Joules. Each node periodically broadcasts a beacon message within its neighborhood to make its presence felt to the other nodes in the neighborhood.

A. *Multicast Extension of Ad hoc On-demand Distance Vector (MAODV) Routing Protocol*

MAODV [10] is the multicast extension of the well-known Ad hoc On-demand Distance Vector (AODV) unicast routing protocol [12]. Here, a receiver node joins the multicast tree through a member node that lies on the minimum-hop path to the source. A potential receiver wishing to join the multicast group broadcasts a *Route-Request* (RREQ) message. If a node receives the RREQ message and is not part of the multicast tree, the node broadcasts the message in its neighborhood and also establishes the reverse path by storing the state information consisting of the group address, requesting node id and the sender node id in a temporary cache. If a node receiving the RREQ message is a member of the multicast tree and has not seen the RREQ message earlier, the node waits to receive several RREQ messages and sends back a *Route-Reply* (RREP) message on the shortest path to the receiver. The member node also informs in the RREP message, the number of hops from itself to the source. The potential receiver receives several RREP messages and selects the member node which lies on the shortest path to the source. The receiver node sends a *Multicast Activation* (MACT) message to the selected member node along the chosen route. The route from the source to receiver is set up when the member node and all the intermediate nodes in the chosen path update their multicast table with state information from the temporary cache. A similar approach can be used in NR-MLPBR and R-MLPBR when a new receiver node wishes to join the multicast group.

B. *Bandwidth-Efficient Multicast Routing Protocol (BEMRP)*

According to BEMRP [9], a newly joining node to the multicast group opts for the nearest forwarding node in the existing tree, rather than choosing a minimum-hop path from the source of the multicast group. As a result, the number of links in the multicast tree is reduced leading to savings in the network bandwidth. Multicast tree construction is receiver-initiated. When a node wishes to join the multicast group as a receiver, it initiates the flooding of *Join control* packets targeted towards the nodes that are currently members of the multicast tree. On receiving the first *Join control* packet, the member node waits for a certain time before sending a *Reply* packet. The member node sends a *Reply* packet on the path, traversed by the *Join* control packet, with the minimum number of intermediate forwarding nodes. The newly joining receiver node collects the *Reply* packets from different member nodes and would send a *Reserve* packet on that path that has the least number of forwarding nodes from the member node to itself.

C. Performance Metrics

The performance metrics studied through this simulation are the following:

- **Number of Links per Tree:** This is the time averaged number of links in the multicast trees discovered and computed over the entire multicast session. The notion of "time-average" is explained as follows: Let there be multicast trees T1, T2, T3 with 4, 8 and 6 links used for time 12, 6 and 15 seconds respectively, then the time averaged number of links in the multicast trees is given by (4*12+8*6+6*15)/ (12+6+15) = 5.6 and not merely 6.0, which is the average of 4, 8 and 6.
- **Hop Count per Source-Receiver Path**: This is the time averaged hop count of the paths from the source to each receiver of the multicast group and computed over the entire multicast session.
- **Time between Successive Broadcast Tree Discoveries:** This is the time between two successive broadcast tree discoveries, averaged over the entire multicast session. This metric is a measure of the lifetime of the multicast trees discovered and also the effectiveness of the path prediction approach followed in NR-MLPBR and R-MLPBR.
- **Energy Consumed per Node:** This is the sum of the energy consumed at a node due to the transfer of data packets as part of the multicast session, broadcast tree discoveries as well as the periodic broadcast and exchange of beacons in the neighborhood.

IV. SIMULATION RESULTS

The performance results for each metric displayed in Figures 10 through 13 are an average of the results obtained from simulations conducted with 5 sets of multicast groups and 5 sets of mobility profiles for each group size, node velocity and network density values. The multicast source in each case was selected randomly among the nodes in the network and the source is not part of the multicast group. The nodes that are part of the multicast group are merely the receivers.

A. Number of Links per Multicast Tree

The number of links per multicast tree (refer figure 10) is a measure of the efficiency of the multicast routing protocol in reducing the number of link transmissions during the transfer of the multicast data from the source to the receivers of the multicast group. The smaller is the number of links in the tree, the larger the link transmission efficiency of the multicast routing protocol. If fewer links are part of the tree, then the chances of multiple transmissions in the network increase and this increases the efficiency of link usage and the network bandwidth. Naturally, the BEMRP protocol, which has been purely designed to yield bandwidth-efficient multicast trees, discovers trees that have a reduced number of links for all the operating scenarios. This leads to larger hop count per source-receiver paths for BEMRP as observed in figures 11.

R-MLPBR, which has been designed to choose the predicted paths with the minimum number of non-receiver nodes, manages to significantly reduce the number of links vis-à-vis the MAODV and NR-MLPBR protocols. R-MLPBR attempts to minimize the number of links in the multicast tree without yielding to a higher hop count per source-receiver path. But, the tradeoff between the link efficiency and the hop count per source-receiver path continues to exist and it cannot be nullified. In other words, R-MLPBR cannot discover trees that have minimum number of links as well as the minimum hop count per source-receiver path. Nevertheless, R-MLPBR is the first multicast routing protocol that yields trees with the reduced number of links and at the same time, with a reduced hop count (close to the minimum) per source-receiver path.

For a given network density and multicast group size, we do not see any appreciable variation in the number of links per tree for each of the multicast routing protocols studied. As the network density increases, BEMRP attempts to reduce the number of links per tree by incorporating links that can be shared by multiple receivers on the paths towards the source. On the other hand, both MAODV and NR-MLPBR attempt to choose minimum hop paths between the source and any receiver and hence exploit the increase in network density to discover minimum hop paths, but at the cost of the link efficiency. On the other hand, R-MLPBR attempts to reduce the number of links per tree as we increase the network density. For a given multicast group size, the number of links per tree for R-MLPBR is about 4-15%, 8-18% and 10-21% more than that incurred by BEMRP. This shows that R-MLPBR is relatively more scalable, similar to BEMRP, with increase in the network density. For medium and large-sized multicast groups, the number of links per tree for both MAODV and NR-MLPBR is about 7-15%, 17-28% and 22-38% more than that incurred for BEMRP in low, medium and high-density networks respectively. On the other hand, the number of links per tree for R-MLPBR is about 6-15%, 12-18% and 16-21% more than that incurred for BEMRP in low, medium and high-density networks respectively. This shows that R-MLPBR is relatively more scalable, similar to BEMRP, with increase in the network density.

B. Hop Count per Source-Receiver Path

All the three multicast routing protocols – MAODV, NR-MLPBR and R-MLPBR, incur almost the same average hop count per source-receiver and it is considerably lower than that incurred for BEMRP. The hop count per source-receiver path is an important metric and it is often indicative of the end-to-end delay per multicast packet from the source to a specific receiver. BEMRP incurs a significantly larger hop count per source-receiver path and this can be attributed to the nature of this multicast routing protocol to look for trees with a reduced number of links. When multiple receiver nodes have to be connected to the source through a reduced set of links, the hop count per source-receiver path is bound to increase. The hop count per source-receiver path increases significantly as we increase the multicast group size. The hop count per source-receiver path for BEMRP can be as large as 41%, 57% and 59% more than that of the hop count per source-receiver path incurred for the other three multicast routing protocols.

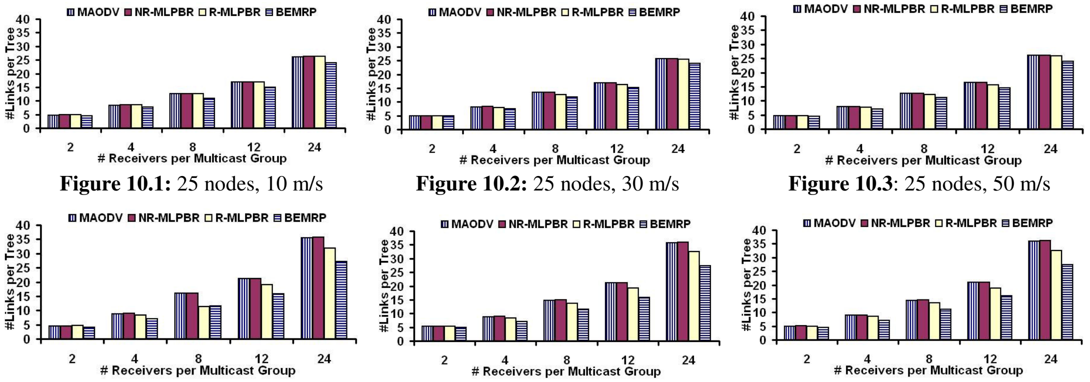


Figure 10.1: 25 nodes, 10 m/s **Figure 10.2:** 25 nodes, 30 m/s **Figure 10.3**: 25 nodes, 50 m/s

Figure 10.4: 75 nodes, 10 m/s **Figure 10.5:** 75 nodes, 30 m/s **Figure 10.6**: 75 nodes, 50 m/s

Figure 10: Average Number of Links per Multicast Tree (Route Discovery Procedure: Flooding)

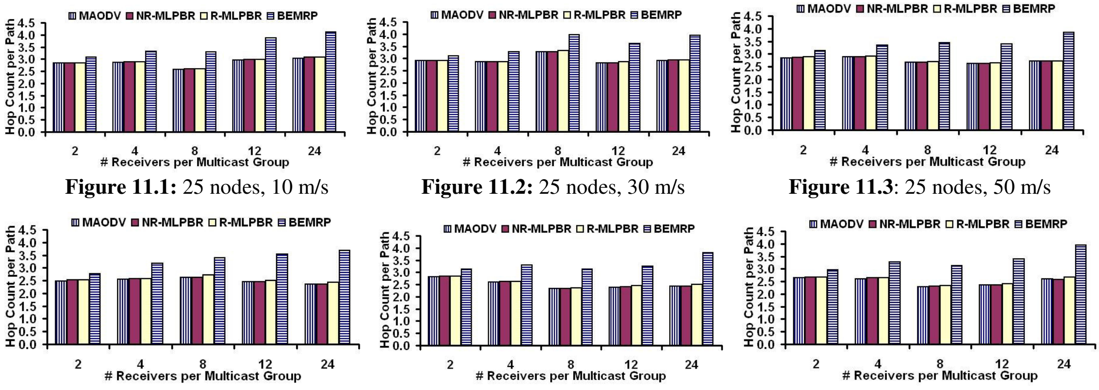


Figure 11.1: 25 nodes, 10 m/s **Figure 11.2:** 25 nodes, 30 m/s **Figure 11.3**: 25 nodes, 50 m/s

Figure 11.4: 75 nodes, 10 m/s **Figure 11.5:** 75 nodes, 30 m/s **Figure 11.6**: 75 nodes, 50 m/s

Figure 11: Average Hop Count per Source-Receiver Path (Route Discovery Procedure: Flooding)

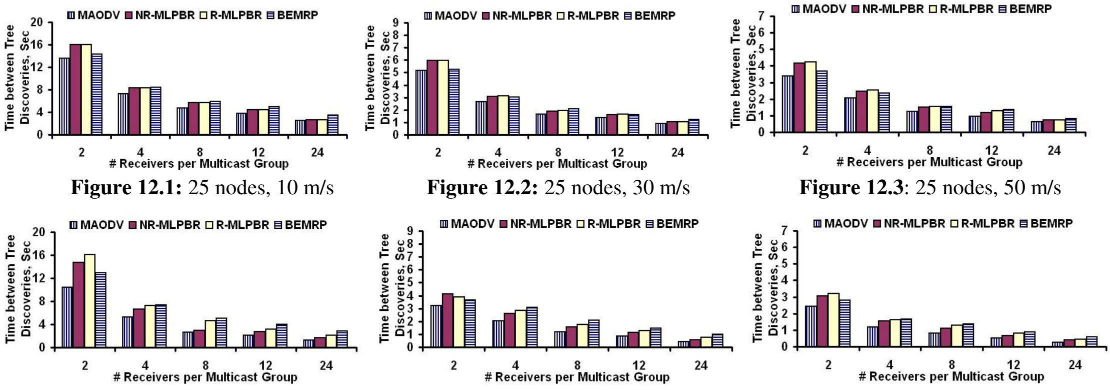


Figure 12.1: 25 nodes, 10 m/s **Figure 12.2:** 25 nodes, 30 m/s **Figure 12.3**: 25 nodes, 50 m/s

Figure 12.4: 75 nodes, 10 m/s **Figure 12.5:** 75 nodes, 30 m/s **Figure 12.6**: 75 nodes, 50 m/s

Figure 12: Average Time between Successive Tree Discoveries (Route Discovery Procedure: Flooding)

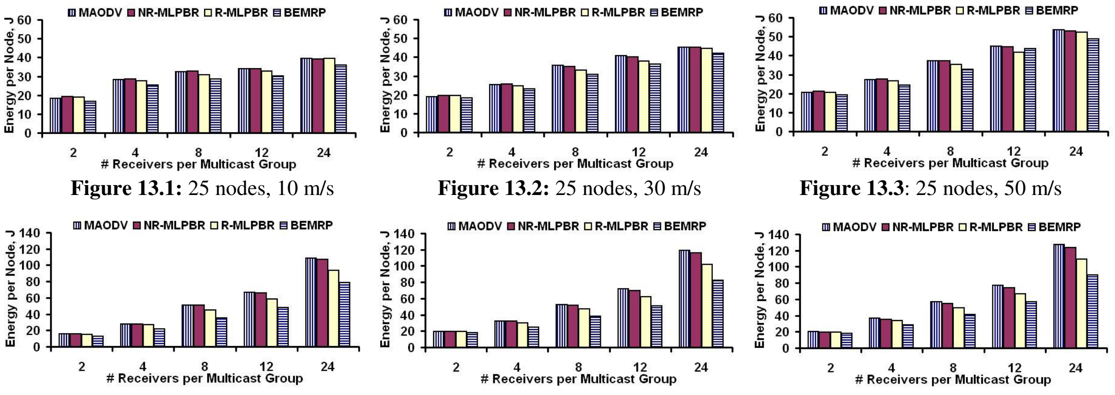


Figure 13.1: 25 nodes, 10 m/s

Figure 13.2: 25 nodes, 30 m/s

Figure 13.3: 25 nodes, 50 m/s

Figure 13.4: 75 nodes, 10 m/s

Figure 13.5: 75 nodes, 30 m/s

Figure 13.6: 75 nodes, 50 m/s

Figure 13: Average Energy Consumed per Node (Route Discovery Procedure: Flooding)

For a given network density and group size, there is no appreciable variation in the hop count per source-receiver path for each of the multicast routing protocols studied. As we increase the network density, the hop count per source-receiver path decreases. This is mainly observed in the case of the minimum-hop based MAODV, NR-MLPBR and R-MLPBR. With BEMRP, the impact of network density on the decrease in the hop count is relatively less as it is a bandwidth-efficient multicast routing protocol attempting to reduce the number of links in the tree. The hop count per source-receiver path for BEMRP increased with increase in the multicast group size, while the hop count per source-receiver path for the other multicast routing protocols almost remained the same with increase in multicast group size.

C. *Time between Successive Broadcast Tree Discoveries*

The time between successive broadcast tree discoveries is a measure of the stability of the multicast trees and the effectiveness of the location prediction and path prediction approach of the two multicast extensions. For a given condition of node density and node mobility, both NR-MLPBR and R-MLPBR incur relatively larger time between successive broadcast tree discoveries for smaller and medium sized multicast groups. MAODV tends to be more unstable as the multicast group size is increased, owing to the minimum hop nature of the paths discovered and absence of any path prediction approach. For larger multicast groups, BEMRP tends to perform better by virtue of its tendency to strictly minimize only the number of links in the tree. On the other hand, NR-MLPBR attempts to reduce the hop count per source-receiver path and ends up choosing predicted paths that increase the number of links in the tree, quickly leading to the failure of the tree. The time between successive tree discoveries for R-MLPBR is 15-25%, 15-59% and 20-82% more than that obtained for MAODV in networks of low, moderate and high density respectively. For a given level of node mobility and network density, MAODV trees become highly unstable as the multicast group size increases. For multicast groups of size 2 and 4, the time between successive broadcast tree discoveries for NR-MLPBR and R-MLPBR is greater than that obtained for BEMRP, especially in networks of low and moderate network density. For larger multicast group sizes, BEMRP tends to incur larger time between successive broadcast tree discoveries compared to NR-MLPBR and R-MLPBR. While using a broadcast strategy that will lead to the discovery of inherently stable trees, we conjecture that R-MLPBR will tend to incur larger time between successive broadcast tree discoveries compared to BEMRP, even for larger group sizes.

For each multicast routing protocol, for a given multicast group size and level of node mobility, as the network density increases, the time between successive broadcast tree discoveries decreases. This is mainly observed for the minimum-hop based multicast protocols (especially MAODV and NR-MLPBR) which incur a reduced hop count per source-receiver path as we increase the network density. But, such minimum hop paths obtained in moderate and high-density networks are relatively less stable than those obtained in low-density networks. For a given multicast group size and low node mobility, the time between successive tree discoveries in networks of high density (75 nodes) is 51-80% for MAODV and NR-MLPBR and for R-MLPBR and BEMRP is 70-90% of those obtained in networks of low-density. For a given network density and node mobility, the time between successive route discoveries decreases as the multicast group size increases. For smaller group sizes, the time between successive broadcast tree discoveries for MAODV and BEMRP is respectively about 80%-90% and 85%-94% of that incurred for NR-MLPBR and R-MLPBR. For larger groups, the time between successive tree discoveries for NR-MLPBR and R-MLPBR is respectively about 57%-76% and 75%-80% of that incurred for BEMRP for all network densities.

D. *Energy Consumed per Node*

Energy consumption in multicast routing is directly proportional to the number of links in the tree. Larger the number of links, more the transmissions and more will be the energy consumption in the network and vice-versa. Simulation results in Figure 13 clearly illustrate this. BEMRP incurs the least energy consumption per node and MAODV incurs the largest energy consumption per node. The energy consumed per node for the two multicast extensions is in between these two extremes. The energy consumed per node for R-MLPBR is less than that of NR-MLPBR as the former also attempts to simultaneously reduce the number of links as well as the hop count per source-receiver path. The energy consumption per node increases as the multicast group size increases. For a given multicast group size and multicast routing protocol, the energy consumed per node increases with increase in network density as well as with increase in node mobility.

For a given multicast group size, network density and multicast routing protocol, the energy consumed per node at higher node velocities of 30 m/s and 50 m/s can grow as large as 10-40% of that obtained at maximal node velocity of 10 m/s. BEMRP and MAODV incur the largest increase in energy consumed per node with increase in node mobility. NR-MLPBR and R-MLPBR incur a relatively lower increase in the energy consumed per node with increase in node mobility. This can be attributed to the tendency of these multicast routing protocols to reduce the number of broadcast tree discoveries using effective tree prediction.

For multicast groups of size 2 and 4, as we increase the network density from 25 to 75 nodes, the energy consumed per node decreases. This is due to the smaller group size, leading to the effective sharing of the data forwarding load among all the nodes in the network. For larger group sizes, all the nodes in the network end up spending more energy (due to transmission/reception or at least receiving the packets in the neighborhood). MAODV and NR-MLPBR incur a relatively larger energy consumed per node at high network densities due to the nature of these routing protocols to discover trees with minimum hop count. R-MLPBR and BEMRP discover trees with reduced number of links and hence incur relatively lower energy consumed per node at high network density.

V. CONCLUSIONS

In this paper, we propose multicast extensions to the location prediction based routing (LPBR) protocol for mobile ad hoc networks (MANETs). The multicast extensions of LPBR (referred to as NR-MLPBR and R-MLPBR) have been proposed to simultaneously reduce the number of tree discoveries and the hop count per path from the source to each of the receivers of the multicast group. NR-MLPBR and R-MLPBR differ from each other based on the type of path predicted and notified to the source. NR-MLPBR determines the minimum hop path to the source and sends a Multicast Predicted Path Message on the minimum hop path to the source. R-MLPBR assumes that each receiver knows the identity of the other receivers of the multicast group and hence attempts to choose a path that will minimize the number of newly added intermediate nodes to the multicast tree. R-MLPBR has been thus designed to also reduce the number of links that form the multicast tree, in addition to the source-receiver hop count and the number of global tree discoveries. Nevertheless, the number of links per tree discovered using R-MLPBR is still about 15-20% more than that discovered using BEMRP, but the hop count per source-receiver path is significantly smaller (by about 40%-60%) than those observed in trees discovered using BEMRP and is the same as that discovered using MAODV (aims to minimize the hop count between source-receiver paths). We conjecture that with the deployment of broadcast tree discovery strategies (such as DMEF [13]) that can discover inherently stable trees, the performance of NR-MLPBR and R-MLPBR with respect to the time between successive tree discoveries and energy consumed per node actually improved relatively more than that observed for BEMRP and MAODV. This can be attributed to the effective path prediction of the two multicast extensions, an idea inherited from LPBR.

REFERENCES

[1] J. Broch, D. A. Maltz, D. B. Johnson, Y. C. Hu and J. Jetcheva, "A Performance of Comparison of Multi-hop Wireless Ad hoc Network Routing Protocols," in *Proceedings of the 4th Annual ACM/IEEE Conference on Mobile Computing and Networking*, pp. 85 – 97, October 1998.

[2] N. Meghanathan, "Location Prediction Based Routing Protocol for Mobile Ad Hoc Networks," *Proceedings of the IEEE Global Communications Conference* (GLOBECOM), New Orleans, Nov-Dec 2008.

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, "Introduction to Algorithms," 2nd Edition, MIT Press/ McGraw Hill, Sept. 2001.

[4] C. S. R. Murthy and B. S. Manoj, "Ad Hoc Wireless Networks: Architectures and Protocols," Prentice Hall, June 3, 2004.

[5] J.-P. Sheu, C.-M. Chao, W.-K. Hu and C.-W. Sun, "A Clock Synchronization Algorithm for Multi-hop Wireless Ad hoc Networks," Wireless Personal Communications: An International Journal, vol. 43, no. 2, pp. 185-200, 2007.

[6] D. Zhou, "A Compatible and Scalable Clock Synchronization Protocol in IEEE 802.11 Ad hoc Networks," Proceedings of the International Conference on Parallel Processing, pp. 295-302, Washington DC, USA, June 2005.

[7] L. Breslau, et. al., "Advances in Network Simulation," *IEEE Computer*, vol. 33, no. 5, pp. 59-67, 2000.

[8] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE Journal of Selected Areas in Communication*, vol. 18, no. 3, pp. 535-547, 2000.

[9] T. Ozaki, J-B. Kim and T. Suda, "Bandwidth-Efficient Multicast Routing for Multihop, Ad hoc Wireless Networks," *Proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies* (INFOCOM), vol. 2, pp. 1182-1192, Anchorage, USA, April 2001.

[10] E. Royer and C. E. Perkins, "Multicast Operation of the Ad-hoc On-demand Distance Vector Routing Protocol," *Proceedings of the 5th ACM/IEEE Annual Conference on Mobile Computing and Networking*, (MOBICOM), pp. 207-218, Seattle, USA, August 1999.

[11] C. Bettstetter, H. Hartenstein and X. Perez-Costa, "Stochastic Properties of the Random-Waypoint Mobility Model," *Wireless Networks*, vol. 10, no. 5, pp. 555-567, 2004.

[12] C. E. Perkins and E. M. Royer, "The Ad hoc On-demand Distance Vector Protocol," Ad hoc Networking, edited by C. E. Perkins, pp. 173-219, Addison-Wesley, 2000.

[13] N. Meghanathan, "A Density and Mobility Aware Energy-Efficient Broadcast Strategy to Determine Stable Routes in Mobile Ad hoc Networks," submitted for journal publication, March 2009.

Dr. Natarajan Meghanathan is currently working as Assistant Professor of Computer Science at Jackson State University, Mississippi, USA, since August 2005. Dr. Meghanathan received his MS and PhD in Computer Science from Auburn University, AL and The University of Texas at Dallas in August 2002 and May 2005 respectively.

Dr. Meghanathan's main area of research is ad hoc networks. He has more than 30 peer-reviewed publications in leading international journals and conferences in this area. Recently, Dr. Meghanathan has received grants from the Army Research Laboratory (ARL) to conduct research on ad hoc routing protocols and from the National Science Foundation (NSF) to conduct a three-year Summer Research Experience for Undergraduates (REU) program at Jackson State University for the years 2009-11. Dr. Meghanathan currently serves as the editor of a number of international journals and also in the program committee and organization committees of several leading international conferences in the area of networks.

Bifurcation analysis and synchronization issues in a short life cycle products supply chain

K.R.Anne
Institute for Smart Systems Technologies
University of Klagenfurt
Klagenfurt, Austria
rao.anne@uni-klu.ac.at

J.C.Chedjou
Institute for Smart Systems Technologies
University of Klagenfurt
Klagenfurt, Austria
Jean.Chedjou@uni-klu.ac.at

K. Kyamakya
Institute for Smart Systems Technologies
University of Klagenfurt
Klagenfurt, Austria
Kyandoghere.kyamakya@uni-klu.ac.at

Abstract— **In today's global market-place, the majority of products have a short product life cycle due to the innovative character of the products and ever changing customer desires. This short product life cycle nature produces various types of uncertainties along the supply chain, e.g. demand uncertainty, supply uncertainty, delivery uncertainty and forecasting uncertainty. These uncertainties make supply chains complex and nonlinear systems as they propagate along the supply chain in both upstream and down stream. This work investigates the dynamical behavior of a three-echelon supply chain. The modeling of this structure is carried out to display its nonlinear dynamical behavior. It is shown that the dynamics (e.g. stability) of the supply chain is very sensitive to external uncertainties. Specifically, the supply chain subjected to these uncertainties can exhibit strange and undesired dynamics/states such as saturation and chaos. An adaptive algorithm for the automatic cancellation of these strange dynamics due to uncertainties is developed by re-adjusting the internal parameters of the supply chain in order to achieve its synchronization. A bifurcation analysis is also carried out. This analysis is essential and useful for strategic decision makers as it allows both the visualization and control of the states/dynamics of the entire supply chain. In order to display the dynamics of a real world supply chain, the bifurcation analysis is performed for the Tamagotchi™ supply chain.**

Keywords— Supply chain synchronization; Supply chain control; Bullwhip effect; Chaos in supply chains; Bifurcation analysis for supply chains.

I. INTRODUCTION

Supply chain network management has been defined as the management of upstream and downstream relationships with suppliers and customers in order to create enhanced value in the final market-place at less cost to the supply chain as a whole [1]. Today's global market-place is increasingly dynamic and volatile. This dynamic and volatile nature produces various types of uncertainties along the supply chain e.g., demand uncertainty, supply uncertainty, delivery uncertainty, forecasting uncertainty. Apart from these uncertainties caused by external sources, uncertainties observed on daily basis (e. g. machine breakdowns, wrong supplies, supply shortages etc.) make supply chains much more complex systems. The uncertainties propagate along the upstream and downstream of the entire supply chain leading to the production of various nonlinear dynamic effects. In addition to these uncertainties, the relation between the various players in the supply chain is often characterized by the mistrust and competition.

In fact, inventory is generally used as insurance against the uncertainties. In the case of a single enterprise based supply chain it is relatively easy to overcome the uncertainties with properly sized inventories at each stage like raw materials, work in process and finished goods inventories. The present day statistical tools and forecasting methods can satisfactorily aid in determining how much must be hold to satisfy the customer demand for the particular product despite the uncertainties [2]. However, the problem is much more complicated when considering the whole network consisting of different players distributed globally. Nearly each player holds some inventory to protect against uncertainties, but the real difficulty is in determining how much must be hold and where to hold it. To date, there is no clear analytical way to calculate the propagation of uncertainties up and/or down the supply chain. Traditionally, firms have relied on the experience and intuition in facing uncertainties.

The decisions made with intuition can make the supply chain exhibit various dynamic states like chaos. Chaos is defined as an aperiodic, unpredictable and bounded dynamics in a deterministic system with sensitivity dependence on initial conditions. Chaos is a disorderly long term evolution occurring in deterministic nonlinear systems. The beer game developed at MIT to introduce the students and industrialists to the concepts of economic dynamics shed further light on supply chain dynamics [3]. It has been found that one in four management teams in the supply chain creates deterministic chaos in the ordering pattern and inventory levels [4]. This clearly demonstrated in practice the occurrence of chaos in supply chains.

The objectives of this paper are: (1) to show the potentials of the concepts of nonlinear dynamics and modelling in supply chain networks; (2) to show the interest of synchronization in supply chain networks; (3) to offer to strategic decision makers an approach or technique to control and stabilize the states of their supply chain networks when subjected to uncertainties; and (4) to demonstrate the dynamics of a Tamagotchi$^{(T)}$ supply chain through bifurcation analysis.

II. RESPONDING TO UNCERTAINTY

The objective of any player in a supply chain network is to achieve maximum profits and give maximum customer satisfaction. In addition to the stated individual objective all players also have a responsibility towards the global objective of resilient supply chain networks (SCN).

Christopher [5] emphasises this by stating:

Competition in the future will not be between individual organizations but between competing supply chains

This realization has made the entities in SCN to look beyond their own boundaries to asses how the resources of each other can be utilized to achieve the global objective without compromising on their own objective. In the midst of pursing towards it, entities have relentlessly restructured and reengineered their internal organizational boundaries and policies with an objective of transforming their relations from "arm's – length" relationship to "durable arm's – length" relationships [6].

The definition of a truly efficient supply chain is when all players involved are communicating correct data – a manufacturer is communicating the correct product information and receiving accurate purchase orders; a retailer is receiving the specific products that were ordered; and the product is available to the end consumer at the right time and at the right price. However, communicating the correct data is always not possible in supply chains as each stake holder has different objectives and constrains. The traditional supply chain management has been based on limited information sharing restricted to the product in consideration and transaction oriented towards that product [7]. It is well studied [2, 8-10] that information sharing, demand pattern, ordering policy, and lead time have a direct impact on the performance of supply chains. Information sharing can reduce the lead time. Lead time reduction is found to be very beneficial and can reduce inventory and demand variability and improve customer service [8].

However, during the process of identifying the ways to mitigate the effects due to uncertainties, companies within the supply chain realized that they need to achieve the self organization of the supply chain they belong to in order to satisfy the stated objective. In accordance with the greater focus on the self-organization of the supply chain, the companies are increasingly focusing on the pre-requisitions like integration, collaboration and synchronization between all entities in the supply chain as shown in Figure 1.

The first step towards self-organized supply chains is the integration stage. A Complex corporate- structure demands the various functional units within a company to be integrated first.. The intra-entity integration also called the functional integration is the basic driver towards the integration of the entire supply chain. The functional integration of purchasing, manufacturing, transportation and warehousing activities gives the very much needed visibility to the supply chain. Besides the functional integration, the inter-temporal integration also called hierarchical integration of these activities over strategic, tactical, and operational levels is also important [11-13]. This inter-temporal integration requires consistency among overlapping of supply chain decisions at various levels of planning. However, the major role of the inter-temporal integration is in designing the supply chain for the product. The improved integration of activities across multiple companies/entities in a supply chain allows the collection of data from the breadth of the supply chain [14]. Information technology is the key enabler for the integration in supply chains. The information systems like EDI (Electronic Data Interchange) and ERP (Enterprise Resource Planning) provide the necessary information needed for the integration [5]. The information systems make the information available; however the effect of integration directly depends on the quality of the

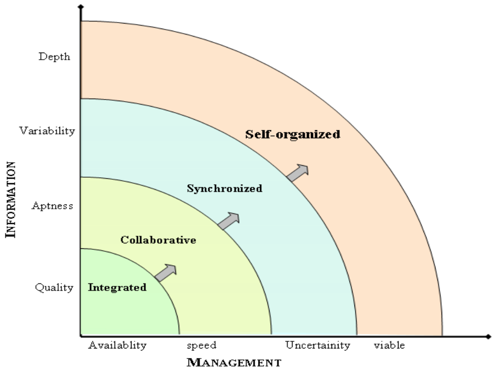


Figure 1. Key stages to the evolution of adaptive supply chain networks

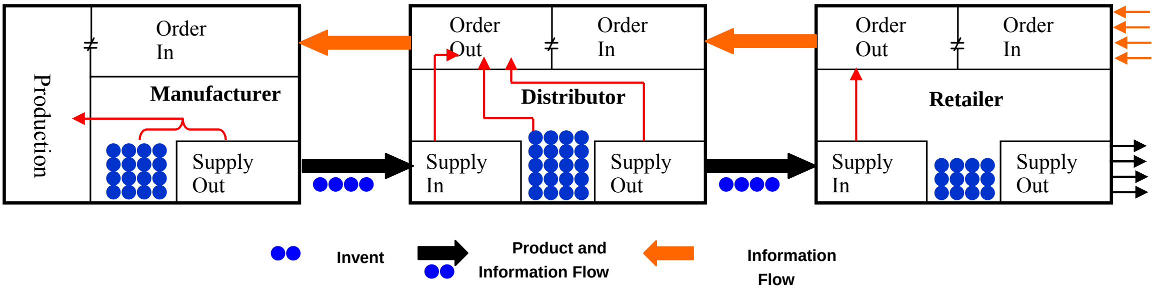


Figure 2. Model of a three echelon supply chain

information made available.

After the integration of the supply chain, the next step towards self-organization is the collaboration between entities. The objective of the collaboration is to make the information available when it is needed. Of course collaboration concepts are not new; they exist from the days of traditional business in the form of contracts. However, the effectiveness, execution speed, of the collaborations is highly increased due to the technological advances and the integration tools. Collaboration is strong where business to business relationships are strong. The degree of collaboration varies depending upon the strength of the integration [15]. But true collaboration among and in between all the entities in a SCN is more difficult in practice.

Even though data is made available with the help of integration tools and collaborative agreements, often the collected data paints a false image of an operation due to data entry errors and inconsistent collection procedures. Further, the upstream /downstream requirements are sometimes not clearly understood, not accurate enough and are not up-to-date. The breadth of the supply chain can compound the accuracy problem as the data can be re-worked or re-created in between.

In order to cope up with inaccurate and/or varied data, synchronization is an important step in dealing with uncertainties. Synchronization can be classified into two types, namely the complete synchronization and the partial synchronization. The complete synchronization is observed when the data synchronization is achieved, leading to the real-time access to available data by all players at the same time. Complete synchronization enables the supply chain to react quickly to changes in demand and in product design. This type of synchronization is particularly suitable in just -in-time supply chain networks. To achieve the complete synchronization the complete chain should follow the integration and collaboration methods in true spirit.

Partial synchronization is achieved through a feedback controller item . Apart from the data synchronization as explained in complete synchronization, a controller item is developed to mitigate the effects due to both inaccurate data and uncertainties. In this type of synchronization the major effort is placed in quantifying the effects due to uncertainties. The modelling and quantification of the effects caused by the time lag (i.e. time delay), the information discrepancy, and the individual objectives help in designing the controller/synchronizer element. This controller item can be unique for each entity or supply chain as a whole. Uncertainties and exceptions are identified early and the data for intelligent response are immediately available. This greatly minimizes the bullwhip effect, demand amplification, and saves downstream partners and customers from needless activity [10, 11, 16, 17]. In this paper, we propose an adaptive controller to achieve synchronization phenomena in order to mitigate the effects due to uncertainties.

III. Modeling of a three level supply chain

The theoretical framework for supply chain management underlies the setting, optimization and control of the system model. The system model is not unique for all the supply chains [18]. The system dynamics change for each type of product e.g. food, oil, consumer goods, etc., depending upon the processes involved. The system dynamics based approach to model the business dynamics was first introduced by Forrester [19]. The system dynamics has its origins in control engineering and management. The approach uses a perspective based on information feedback and delays to understand the dynamic behaviour of complex physical and social systems. System dynamics is an approach which is actively used to model the managerial behaviour.

In 1989, Sterman [3] proposed a model that can be used to analyse the supply chains using the industrial dynamics fundamentals proposed by Forrester. The Sterman model was actively used to analyse the supply chain dynamics. Due to its visual nature and simplicity, after a lengthy floppy period, the system dynamics approach is gaining momentum in modelling the inventory management process, the policy development and demand amplification [20]. The system theory based modelling is also used to develop the feedback controllers to mitigate the bullwhip effect (i.e. the demand amplification) to some extent [21]. However, to address the issues emanating from uncertainties and the problems occurring in real time, the system dynamics modelling might not be suitable [8, 20, 21]

In recent years, many researchers are also using an agent based distributed modelling approach to cope with supply chain networks (SCN) [13, 22, 23]. One or several agents can be used to represent each entity in the SCN. Each agent is assigned with a local objective and global objective as well. With the advent of mobile agents which can run on lighter platforms, the use of agents to both collect information and take decisions has become popular. Moreover, the agent paradigm is a natural metaphor for network organizations, since companies prefer maximizing their own profit rather than the profit of the supply chain [22]. The multi agent based modelling approach offers a way to elaborate the supply chain as the agents are autonomous and the agent rules can be defined in advance. The distributed decision nature of the multi agent systems makes it easier to add other entities in the local environment. Entities leaving the SCN in the middle will not affect the entire SCN to a great extent. The agent rule framework provides certain amount of trust among the partners as it eliminates the mistrust and deception among entities.

The major disadvantages with the multi agent based SCN modeling is that, as there is no global view of the system theoretical optimisation, the optimization of the supply chain can't be visualized. As per the system dynamics theory any system can be unstable [19]; this theorem proves that the multi agent system , which is a system with multi autonomous elements/entities, again can be unstable. Further, the multi agent system is also based on the assumption that all the participating entities in the SCN are truly integrated and collaborating. However “true integration and collaboration” is highly difficult [3, 10, 17, 24-26].

By considering these issues (in this work) we envisage the complex nonlinear modelling of a three echelon supply chain to represent the realistic dynamics. A three level model is envisaged (Figure 2) to describe a simple scenario in a very complex supply chain. The nonlinear supply chain models in the literature [18, 27] focus mainly on the specific tasks, and thus becomes a transaction oriented approach. In this work, we mainly focus on building a non linear supply chain model that can exhibit more complexity covering the information distortion, retailer order satisfaction and safety stock. An additional criterion is the extreme sensitivity of the model to both uncertainties and initial conditions. The following notations are introduced to facilitate the description of the model:

i Time period

m Rate of customer demand satisfaction at retailer

r Rate of information distortion of products demanded by retailer

k Safety stock coefficient at manufacturer

x_i The quantity demanded by retailer in current period

y_i The quantity distributors can supply in current period

z_i The quantity produced in current period depend on the order

The orders they make might not be equal to orders they receive. The order-out quantity depends not only on how much inventory you have already, but also how much you want to supply out. The order-out quantity at retailer depends on the ratio m at which the demand is satisfied during the previous order. The distributor needs to take into consideration among other things, the rate of information distortion r that can occur in the received orders. The producer needs to take care about the safety stock k in order to avoid the small production batches. These scenarios/phenomena are described in Figure 2. An in-depth explanation is provided below and a corresponding mathematical model is derived to analyse the dynamics of the SCN.

We consider that the demand information is transmitted within the layers of the supply chain with a delay of one unit time. As illustrated in Figure 1, the ordering quantity is not the same as the requested order quantity at any level. The order quantity at the current period of time at retailer is linearly coupled with the distributor and it is influenced by how much of demand is satisfied in the previous period of time. This is scenario/phenomenon is modelled by Eq. (1).

$$x_i = m(y_{i-1} - x_{i-1}) \tag{1}$$

Here m is the ratio at which the demand is satisfied. The dependency/coupling between the distributor, the producer and the retailer (Figure 1) is not linear. Indeed the distributor needs to take the combined effect of retailer and producer into consideration before making his order, i.e., quadratic coupling. Apart from this, the distributor also needs to take into consideration that the order information received from the retailer might be distorted. This scenario is modelled by Eq. (2).

$$y_i = x_{i-1}(r - z_{i-1}) \tag{2}$$

Here, r is the information distortion coefficient. The production quantity from the producer unit typically depends on the distributor's orders and the safety stock. However the distributors' orders again depend on the retailer's orders, i.e., the producer needs to take the combined effect of retailer and distributor into account before making production decisions. This scenario is modelled by Eq. (3)

$$z_i = x_{i-1}y_{i-1} + kz_{i-1} \tag{3}$$

Eqs. (1) – (3) represent the quantity demanded by customers (Eq. (1)), the inventory level of distributors (Eq. (2)) and the quantity produced by producers (Eq. (3)), where:

$x_i < 0$ denotes that the supply is less than customers demand in the previous period

$y_i < 0$ denotes that the information is severely distorted and no adjustment is necessary at the inventory level

$z_i < 0$ denotes the cases of overstock or return and hence no new productions

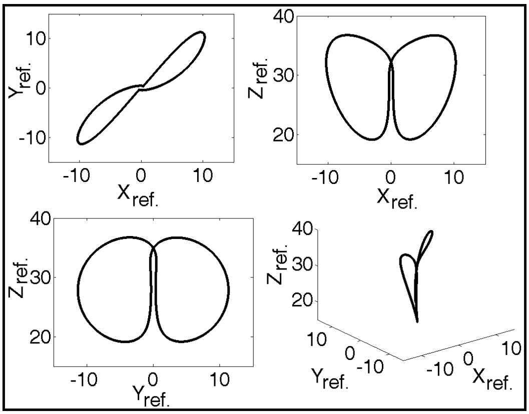


Figure 3. Phase space representation of the reference supply chain for $\sigma = 15$ $r = 29$ and $b = 2/3$

Eqs. (1) - (3) are discrete models describing the dynamics of the SCN of Figure 2. Considering very small time intervals, the continuous model in Eq. (4) can be derived from Eqs. (1) - (3).

$$\begin{cases} \dot{x} = my - (m+1)x \\ \dot{y} = rx - y - xz \\ \dot{z} = xy + (k-1)z \end{cases} \tag{4}$$

If the conditions $\sigma = m + 1$ and $b = 1 - k$ are satisfied, Eq. 4 leads to the *Lorenz* model in Eq. 5.

$$\begin{cases} \dot{x} = \sigma(y - x) \\ \dot{y} = rx - y - xz \\ \dot{z} = xy - bz \end{cases} \tag{5}$$

From the theory of dynamic systems it is proved that this model produces a wide variety of nonlinear features depending upon the parameters values. This model is particularly of interest when dealing with the modelling of scenarios/phenomena which are very sensitive to initial conditions and to uncertainties as well. In the concrete case of SCN, uncertainties, when added at one layer effectively propagate in both upstream and downstream. This is a common dynamics exhibited by realistic SCN. The similar model (Eq. (5)) is also proposed by the authors in [17] to exhibit the supply chain dynamics and mitigate the bullwhip effect. In this work, we are considering the external uncertainties caused by external perturbations. These perturbations can occur due to the market-place dynamics and volatility. We consider the external perturbations are considered to be nonlinear as the market-place behaviour is nonlinear in nature. Assuming that the perturbations can affect any of the three levels of the SCN, a perturbed form of Eq. (5) is proposed in Eq. (6).

$$\begin{cases} \dot{x}' = \sigma(y' - x') + d_1 \\ \dot{y}' = rx' - y' - x'z' + d_2 \\ \dot{z}' = x'y' - bz' + d_3 \end{cases} \tag{6}$$

Where $d_i (i = 1,2,3)$ represent the external perturbations. Before considering the effects of external perturbations on the supply chain we define the reference model with the following parameters values $\sigma = 15$ $r = 29$ and $b = 2/3$. These values illustrate the regular state of the reference supply chain model. The phase space structure of the reference model is shown in Figure 3.

After defining the reference model we analyse the effects of external perturbations on this model. Basically we are concerned with the new dynamics exhibited by the reference supply chain subjected to external perturbations.

IV. SYNCHRONIZATION OF AN EXTERNALLY PERTURBED SUPPLY CHAIN

In this section, we briefly discuss the traditional/classical approach to investigate synchronization issues. Modern synchronization tools provide an automation framework but do not concentrate on what happen if the given data is slightly changed accidentally. Many companies have taken inspiration from the modern web and wireless technologies to make the synchronization in a timeliness manner [1 1, 26]. The integration efforts and the collaboration for the processes within the SCN certainly improved communication by means of EDI and current internet based web information exchanges. Better information (point of sale data and the Collaborative Planning, Forecasting, and Replenishment, CPFR, initiatives), and a general willingness to work more closely together made the timeliness of information possible to certain extent. Nevertheless, the efficiencies have been gained through improvements that any executive can effect at his or her own workplace by putting in place the appropriate company-wide initiatives aimed at improving the internal business process. However, as we have seen the uncertainties propagate in both directions (upstream and downstream) along a SCN, network wide initiatives are necessary to mitigate the effects caused by uncertainties.

In this context, we provide different cases of perturbations affecting the data and present techniques/methods to synchronize or stabilize the new states (or perturbed states) exhibited by the SCN. The causes of instability of the supply chain can be broadly classified into two categories. The first cause is the dynamical and nonlinear character of the motions (i.e. material/products flow, information exchanges, etc.,) between different entities in supply chains. The second cause originates from the effects of both external and internal perturbations [28] to which the supply chain is subjected.

An optimal management of the information flows within the supply chains may be of high importance in order to alleviate the effects leading to negative consequences on the flows within the supply chains. This could be achieved through an adaptive control mechanism which is based on a current comparison of the dynamical data with in the supply chains with the pre-defined data fixed by the requirements of the supply chains. Here, an automatic or adaptive control of the flows with in the supply chains should be able to detect changes in the flows within the supply chains and act accordingly/consequently (by undertaking a given action within the supply chains) in order to alleviate the undesirable effects and therefore stabilize the system behaviour that has been perturbed. The achievement of synchronization is observed when the action undertaken has allowed the recovery of the original behaviour (eventually thresholds or reference requirements) of the supply chain. The schematic description of the adaptive synchronization of supply chain is illustrated in the Figure 4.

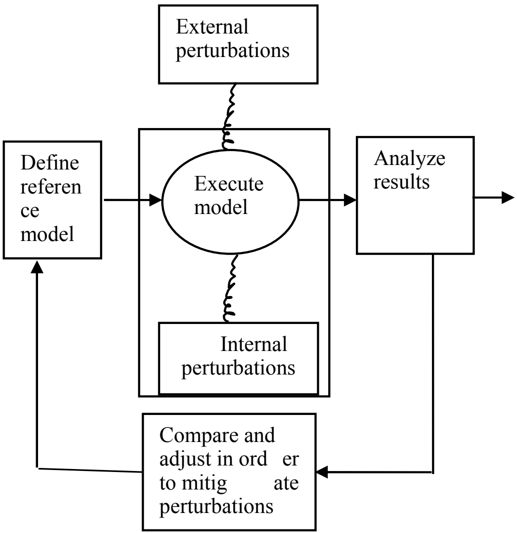


Figure 4. An adaptive feedback control model to mitigate the effects due to uncertainties and perturbations

This paper develops an adaptive method (algorithms and/or tools) for the systematic and automatic control of the flows within the supply chains. In fact, due to the dynamic changes as discussed (in time domain), some pre-defined settings or requirements within the supply chains (thresholds like safety stocks) may be varying accordingly as consequence of these perturbations. It should be worth mentioning that a combination of the simultaneous effects of both internal and external perturbations may be responsible of the dynamic motion variations (e.g. flow of materials, information exchange, etc.) within the supply chain. This is a concrete and/or realistic scenario as the supply chains of many companies are currently exposed to the both types of perturbations. However the analysis in this work is restricted to the case where the reference model is subjected to external perturbations due to the fluctuation in the market demand.

Figure 5 shows a new representation of the phase space structure of the reference supply chain subjected to the following external perturbations: $d_1 = 0.56\cos(3t)$, $d_2 = 20\cos(5t)$ and $d_3 = 50\cos(1020\cos(5t)t)$. The effects of external perturbations on the original (or specific) requirements of the reference supply chain are clearly shown by the well-known chaotic Lorenz attractors exhibited by the reference SCN subjected to perturbations. The cancellation process of these effects is achieved by exploiting the synchronization controller item shown in Figure 4 . The synchronization process concerns two different models of the supply chain: (a) the reference model (unperturbed model) described by Eq.(5), and (b) the externally perturbed model described in Eq. (6). Following the active control approach of Liao [29], for the purpose of synchronization we vary the internal parameters of the perturbed supply chain. In order to vary the internal parameters, we define the state errors between the perturbed system and the reference system by Eq. (7).

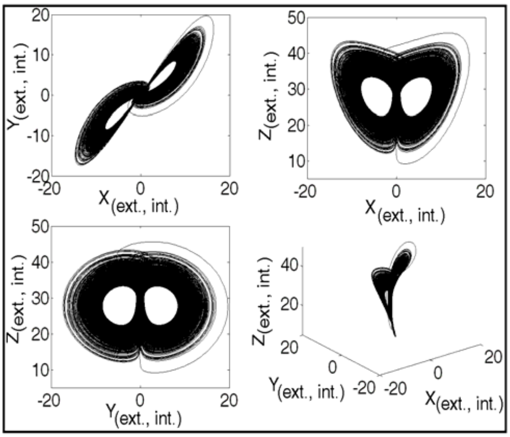


Figure 5. Chaotic state of the supply chain due to external perturbations

$$e_x = x' - x;\ e_y = y' - y;\ e_z = z' - z; \tag{7}$$

x', y' and *z'* are perturbed states and *x, y* and *z* are unperturbed states. The synchronization problem in this context can be equivalent to the problem of stabilizing the system shown in Eq. (7). This is possible through a suitable choice of the internal variables of the perturbed system. In fact, the adaptive control algorithm

Considers the effects of external perturbations and adjusts the values of the internal parameters (σ, r, b) of the three echelon supply chain by tiny variations $d\sigma$, dr and db. The variation of each internal parameter is performed in well defined ranges (or windows) of variation. Performing the parameters variations in these ranges is necessary as we can't vary the parameters beyond the realistic scenario. A threshold error is fixed (which is less than approximately 0.02) under which full alleviation of the effects due to external perturbations is supposed to be effective; this is leads to the achievement of synchronization, which results in the recovery of the behavior of the reference system .

Figure 5 shows the structure in phase space of the supply chain subjected to external perturbations $(d_1 = 0.56\cos(3t), d_2 = 20\cos(5t), d_3 = 50\cos(10t))$. This structure in the phase space shows the occurrence of the well- known bullwhip and chaotic effects in the three level supply chain. The regulation process has been exploited to adjust the internal parameters values in order to achieve synchronization, i.e. the full cancellation of the effects due to perturbations. The corresponding values for the achievement of the regulation process are $d\sigma = 35$, $dr = 15$ and $db = 0.09$. The results of this process are shown in Figure 6. Indeed, below the precision 2%, Figure 6 shows attractors similar to those of the reference or original supply chain system.

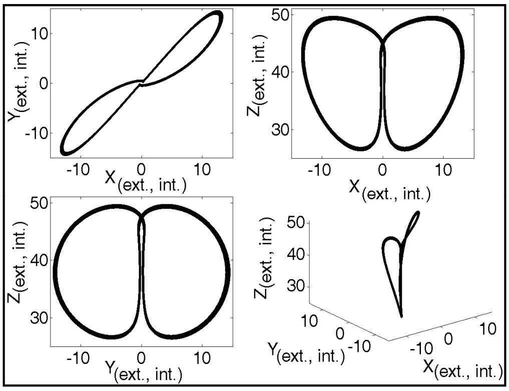


Figure 6. Alleviation of the chaotic effect caused by external perturbation with adaptive synchronization

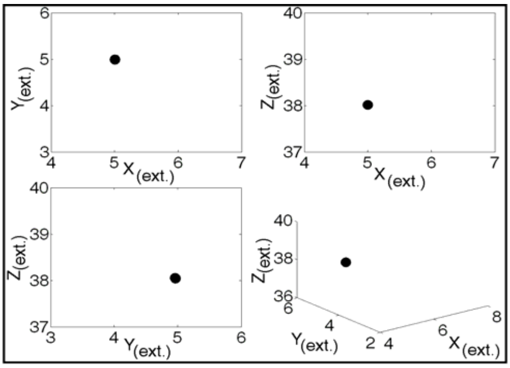


Figure 7. Saturation state of the supply chain due to external perturbations

Further investigations have been performed to show that the supply chain subjected to external perturbations can exhibit the state of saturation which is characterized by non dynamic (or fixed) data in each level of the three levels. Indeed, for $d_1 = 10\cos(5t)$, $d_2 = 5\cos(10t)$ and $d_3 = 10\cos(10t)$, the achievement of the state of saturation is clearly shown in Figure 7. The saturation manifests itself by a sudden exhibition of fixed or constant values/data along each level of the externally perturbed three echelon supply chain. When the state of saturation is obtained, further changes/flows in the supply chain are not represented effectively. To alleviate the effect (i.e. saturation) due to external perturbations, we performed the adaptive regulation process explained before by adjusting the internal parameters of the SCN. The appropriated values of the internal parameters to alleviate the effects due to external perturbations are $d\sigma = 4$, $dr = 2$ and $db = 1$. The result of the regulation process is shown in Figure 8. Indeed, tiny variations of the internal parameters of the supply chain lead to the achievement of synchronization. This is manifested by the abrupt change of the state of the system from the saturation state (Figure 7) to a regular state (Figure 8) which is similar to the state of the reference supply chain (Figure 3).

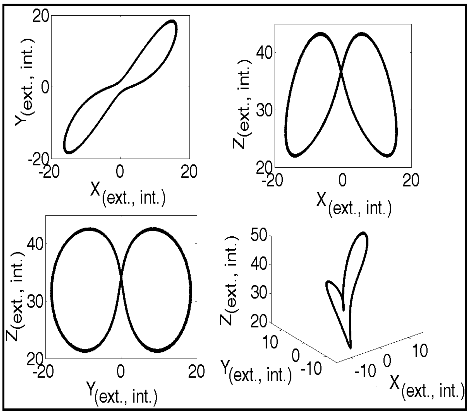


Figure 8. Alleviation of the saturation effect caused by external perturbation with adaptive synchronization

The supply chain subjected to external perturbations can exhibit various striking states such as chaos (Figure 5) and saturation (Figure 7) to name a few. The regulation process can be performed to cancel or mitigate the effects due to external perturbations. This process is based on the adjustment of the internal parameters of the supply chain. Various new and interesting states of the supply chain are discovered towards the achievement of synchronization which is characterized by the cancellation or alleviation of the effects due to external perturbations. Therefore, an interesting and open question must be concerned with the exploration of appropriated methods to control the states of the supply chain. Indeed, the stability of the supply chain is not robust (i.e. very sensitive) to external perturbations. The control process might lead to the derivation of the parameters ranges (windows) in which each of the various states of the supply chain can be found (or defined). The bifurcation analysis is an appropriate method to describe the various states of the supply chain in well specified windows of parameters.

V. Bifurcation Analysis

The bifurcation is a qualitative change observed in the behaviour/state of a system as its parameters settings vary. The bifurcation is observed if the state of the system suddenly changes qualitatively upon small /smooth variation of the parameter values. The bifurcation theory [30] is the analysis/study of the bifurcation scenarios with the aim of defining/determining the states (equilibrium/fixed points, periodic or chaotic states) of the system in a given parameter space. Basically, bifurcation values/points are critical values leading to qualitative changes in the states of the system.

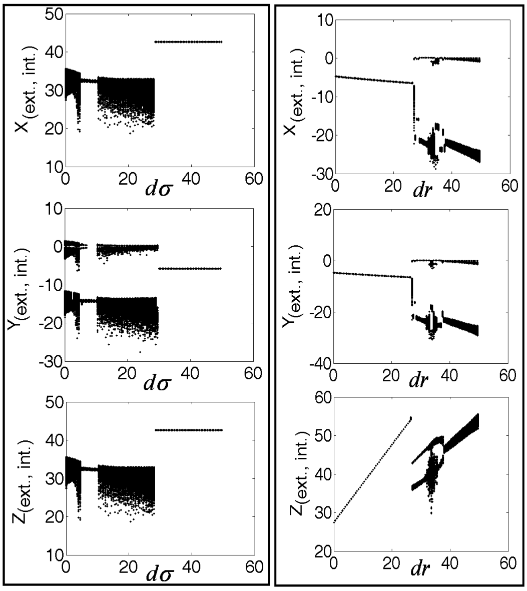


Figure 9. (a) Bifurcation plot showing the sensitivity of the supply chain to the internal variable $d\sigma$ (b) Bifurcation plot showing the sensitivity of the supply chain to the internal variable dr

The preceding section has shown that the annihilation or alleviation of the effects due to external perturbations is possible through the achievement of synchronization. Nevertheless, during the regulation process, we found that the perturbed supply chain system was very sensitive to tiny/small variations of the internal parameters of the supply chain. Indeed, it was observed various states of the supply chain raging from regular to chaotic states. These states were observed when monitoring the internal parameters (e. g. $0 \le d\sigma \le 50$ and $0 \le dr \le 50$) of the supply chain. Therefore, the bifurcation analysis can help to discover the various states towards the achievement of synchronization. This analysis can also be used to control and cancel the effects due to external perturbations.

Figures 9(a)-9(b) are bifurcation diagrams showing the states of the perturbed supply chain. The control parameters $d\sigma$ and dr are obtained by adjusting the internal parameter σ and r respectively. Figure 9 show the extreme sensitivity of the supply chain to the variations of $d\sigma$ and dr. Indeed, windows of regular states are shown which alternate with windows of chaotic states (e.g. period-1, period-3, and chaotic attractors are shown). From Figure 9, windows of parameters can be defined in which each of these states can occur. A control or cancellation of these states is possible and can be achieved through the synchronization analysis.

Bifurcation diagrams are of necessary importance as they can be used to define the ranges of the internal parameters of the supply chain in which the synchronization can be achieved. Two main conditions are important for the achievement of synchronization. The first condition is related to the periodicity. The second condition for the achievement of synchronization is described in Eq. (7). For instance, considering the periodicity of the original supply chain i.e. period-1 attractor (Figure 3), the synchronization (or the annihilation/cancellation of the effects due to external perturbations) to be achieved must lead to the same periodicity. This is the result/consequence of the competition between the internal variation of the system values and the effects of external perturbations. It can be derived from the first condition related to the periodicity the following windows $5 \le d\sigma \le 10$, $30 \le d\sigma \le 50$ and $0 \le dr \le 27$ (Figures. 9-10) in which the synchronization of the supply chain can be achieved. These windows define the set of the internal parameters set tings in which the computation must be performed to fulfil the second condition (Eq. (7)). It should be worth noticing that the bifurcation diagrams were exploited to define the ranges or windows of the internal parameters of the supply chain in which the regulation process can be performed. A random choice of these windows to perform synchronization is possible as well. Nevertheless, when computing in random windows it is not possible to know if the achievement of synchronization is possible. Therefore, the method based on the bifurcation diagram is a systematic tool that can be exploited by strategic decision makers to evaluate how far their supply chain can be affected if the parameters settings are changed.

VI. TAMAGOTCHI™ SUPPLY CHAIN

Tamagotchi™ is the first simulation game of the virtual pet class. In 1996, Bandai & Co. introduced Tamagotchi products into the toy market. Bandai is a famous product vendor for the popular characters, such as POWER RANGERS, GUNDAM, and DIGIMON. Bandai divided its products into eight divisions such as character goods for boys, vending machine products, video games and general toys, models, toys for girls, apparel, snacks and others.

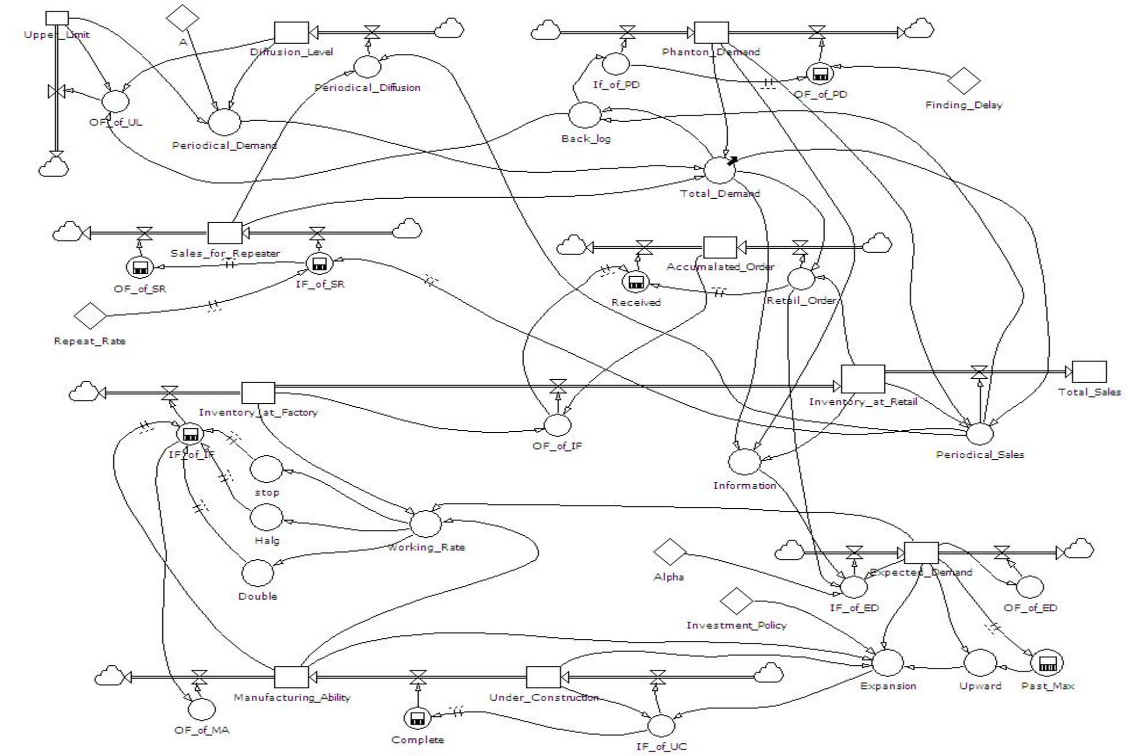


Figure 10. State flow diagram of Tamagotchi™ Supply chain model

Tamagotchi™ appears like an egg-shaped computer game and categorized in the video games and general toys division. The basic idea of Bandai to introduce the Tamagotchi™ is to create a soft pet toy, the way to play with Tamagotchi™ is to take care of it by feeding, giving an injection and so on. Bandai forecasted well in advance at the conceptual stage about the potentiality to strike the toys market even without advertisements using mass media. They initially forecasted that the sales will hit by 300 thousand units by the end of 1996 only in the domestic market. However due to spread of mouth publicity and the immense interest for multiple number of toys Bandai sold about 450 thousand by the end of the 1996 and 4 million by the end of the march, 1997 [31]. However, Bandai was unable to satisfy the customer demand with its supply chain getting affected by the shortage game, copy problems creating phantom demand. In addition, Bandai received more complaints every day about the shortages through wide communication sources and received reports to the police man concerning robberies and aggravated assaults to acquire Tamagotchi™ toys. Finally, Bandai understood the consequences while maintaining overstock and excess capacity and instantly expanded their manufacturing facilities. After expansion of manufacturing facilities Bandai encountered a sharp decline of demand and loses 16 billion yen in fiscal year 1998 [31]. The analysis shows that Bandai was too influenced by the boom and the bullwhip effects.

The above case study illustrates the influence of bullwhip effect and stock out problems. To understand and demonstrate the prevention of these tremendously unfortunate effects Higuchi [31] proposed a system dynamics based approach to conduct sensitivity analysis. To analyze the complex systems Forrester introduced Systems Dynamics concepts [19]. Systems Dynamics is a well-elaborated methodology for deterministic simulation and also analyzes the movements of dynamic systems.

Tamagotchi supply chain model considers three levels, the manufacturer, retail and customer market as described in our theoretical three level model of the supply chain. The scenarios for the market, retail and the factory are described using conceptual framework and then demonstrated the effect of shortages. At market level, the total demand value is calculated as the sum of demands for new customers, phantom demands and sales for repeaters. In addition, the diffusion speed of new products into the market is expressed by using logistic curve i.e. an S-Shaped curve with α and 15% population as initial upper limit. At the retail and factory levels, demands are reviewed and forecasted every week. The issue is here to identify forecasting method that best fits in this case study. Qualitative such as subjective curve fitting, the Delphi method and quantitative methods where experts play a vital role to analyze the next are used to forecast the demand value. During last- period moving average and exponential smoothing are useful methods to invalidate the variations. However, the above statements explain the main advantage of exponential smoothing and significance than the moving average method.

To model the conceptual framework of the Tamagotchi™ case study following assumptions are made: the review of production volume is every week and manufacturing delay is three weeks. The initial manufacturing capability is expected to be 37,500 units per week. The initial maximum manufacturing capability was assumed to be 75,000 units per week by over time and other methods, actual sales is 45,000 units per week in the first six weeks.

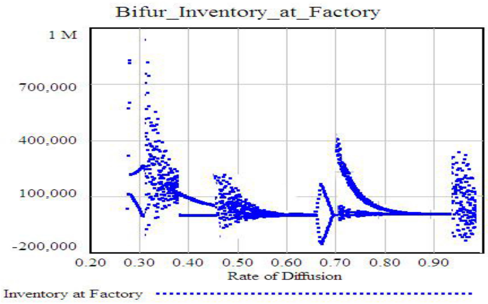


Figure 11. Bifurcation plot showing the sensitivity of the supply chain to the rate of the diffusion

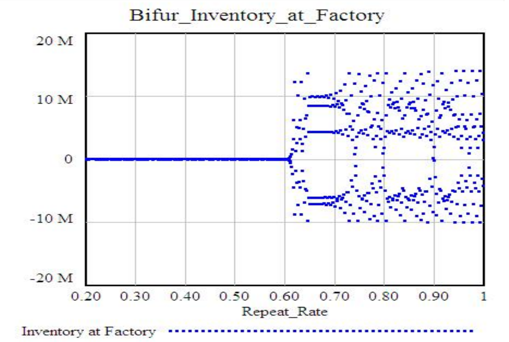


Figure 12. Bifurcation plot showing the sensitivity of the supply chain to the repeat rate

The system dynamics based simulation model as shown in Figure 10 is modeled in powersim$^{\circledR}$ to demonstrate the dynamics of the supply chain. Bifurcation analysis is performed on the key parameters like diffusion speed and rate of repeat for sale for the product with respective to inventory at factory as shown in Figures 11-12. This analysis helps to understand for which parameter range the inventory at the factory can be in steady state. This bifurcation analysis gives the strategists in understanding the sensitive parameter range better as explained in section V.

VII. CONCLUSIONS

The management of supply chains is a complex issue which involves numerous time varying dynamic situations. This work was concerned with the investigation of the effects of uncertainties generated by the dynamic and volatile global market-place on the stability of a three level supply chain. The dynamics of this supply chain was modeled mathematically by the well-known Lorenz oscillator. The uncertainties were considered as periodic external perturbations. The mathematical models were derived by exploiting the structure of the three level supply chain proposed in this work. We have illustrated how a supply chain can exhibit chaotic modes causing forester effect and also saturation modes when subjected to external perturbations. These two modes show the states of instability of the three level supply chain subjected to perturbations. A regulation scheme was designed and exploited to cancel or alleviate the effects due to external perturbations. It was shown that this cancellation or alleviation leads to the achievement of synchronization which is characterized by the reestablishment of the reference data in the supply chain. Two main criteria were defined for the achievement of synchronization. These criteria were exploited to derive some appropriated values of the system parameters for the achievement of synchronization. The regulation process was based on the variation of the internal parameters of the supply chain in well specified ranges/windows. These are ranges/ windows within which the achievement of synchronization is possible. The challenging issue was based on the method to derive or determine these ranges/windows of parameters. We have shown that the bifurcation analysis was appropriate to determine these ranges/windows. The bifurcation analysis was carried out where two important internal parameters (i.e., rate of customer demand satisfaction and rate of information distortion at distributor) were considered as control parameters. Some bifurcation diagrams were obtained showing the extreme sensitivity of the three echelon supply chain when subjected to both external and internal perturbations. It has been found through bifurcations diagrams that the effects due to perturbations can lead to both chaotic and regular states of the supply chain and that these states alternate when monitoring the internal parameters of the supply chain. The bifurcation analysis in this work has been shown to be of necessary importance as it could help the strategic level decision makers in better understanding of the performance of the supply chain over a range of parameter settings. The regulation process exploited in this work was based on an adaptive algorithm for the automatic cancellation of the effects of the external perturbations by re-adjusting the internal thresholds. This process is particularly appealing as it is possible to control or adjust the internal thresholds of the supply chain. The solutions proposed in this paper offer a new range of possibilities for risk managers and provide a future research direction with the aim of considering the concept of nonlinear dynamics.

An open question under investigation concerns the design of "analogue computing" based simulators based on the CNN (Cellular Neural Network) technology to achieve the adaptive synchronization in supply chain networks. This investigation is of high importance due to both the complexity and dynamic character of supply chain networks in practice. These features make the supply chains very difficult to simulate by mean of the classical simulation tools. It would also be of great interest considering the case where the external perturbations to which the supply chain networks are subjected are non-periodic and stochastic. This is a realistic scenario which currently manifests itself in commercial supply chain net works and which can reflect the evolution of the market demand.

VIII. REFERENCES

[1] Uta Jüttne, Martin Christopher, and Susan Baker, " Demand chain management-integrating marketing and supply chain management," Industrial Marketing Management, vol. 36, pp. 377-392, 2007.

[2] Tom Davis, "Effective supply chain management," Sloan Management Review, vol. 34, p. 35, 1993.

[3] John D. Stermann, "Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment," Management Science vol. 35 pp. 321 - 339 1989

[4] Richard Wilding, "The supply chain complexity triangle: Uncertainty generation in the supply chain " International Journal of Physical Distribution & Logistics Management vol. 28, pp. 599 - 616 1998.

[5] M. Christopher, "Logistics and supply chain management: strategies for reducing costs and improving services," Financial Times, Pitman Publishing, London., 1992.

[6] T Skjoett-Larsen, C Thernøe, and C Andresen, " Supply chain collaboration: Theoretical perspectives and empirical evidence " International Journal of Physical Distribution & Logistics Management, vol. 33, pp. 531 - 549, 2003.

[7] James B. Ayers, Handbook of Supply Chain Management, 2nd ed. New York: Auerbach Publications, 2006.

[8] Frank Chen, Zvi Drezner, Jennifer K. Ryan, and David Simchi-Levi, "Quantifying the Bullwhip Effect in a Simple Supply Chain: The Impact of Forecasting, Lead Times, and Information," Management Science vol. 46, pp. 436 - 443 2000.

[9] A. Gunasekaran and E. W. T. Ngai, " Information systems in supply chain integration and management " European Journal of Operational Research, vol. 159, pp. 269 - 295, 2004.

[10] H Brian Hwarng and Na Xie, "Understanding supply chain dynamics: A chaos perspective," European journal of operations research, pp. 1163-1178, 2006.

[11] K. R. Anne, J. C. Chedjou, and K. Kyamakya, "Modelling of a three echelon supply chain: Stability analysis and synchronization issues," in International workshop on nonlinear dynamics and synchronization, Klagenfurt, Austria, 2008.

[12] James B. Ayers, Handbook of Supply Chain Management, 2nd ed. New York: Auerbach Publications, April 2006.

[13] Birgit Dam Jespersen and Tage Skjott-Larsen, Supply Chain Management: In Theory and Practice: Copenhagen Business School Press DK, 2005.

[14] Stanley E. Fawcett and Gregory M. Magnan, "The rhetoric and reality of supply chain integration," International Journal of Physical Distribution & Logistics Management, vol. 32, pp. 339-361, April 2002.

[15] Matthias Holweg, Stephen Disney , Jan Holmström, and Johanna Småros, "Supply Chain Collaboration: Making Sense of the Strategy Continuum," European Management Journal, vol. 23, pp. 170-181 April 2005.

[16] K. R. Anne, J. C. Chedjou, and K. Kyamakya, "Bifurcation analysis and synchronization issues in a three echelon supply chain network," in Logistics Research Network annual conference Liverpool, UK, 2008.

[17] Zhang Lei, Li Yi-jun, and Xu Yao-qun, " Chaos Synchronization of Bullwhip Effect in a Supply Chain," in Management Science and Engineering, 2006. ICMSE '06. 2006 International Conference on, Lille, Date: 5-7 Oct. 2006, pp. 557-560.

[18] Charles C. Poirier, Using Models to Improve the Supply Chain Routledge, USA August 2003.

[19] J. W. Forrester, Industrial Dynamics. Portland: Productivity press, 1961.

[20] Bernhard J. Angerhofer and Marios C. Angelides, "System dyanamics modelling of supply chain management: Research review," in Winter Simulation Conference, 2000.

[21] Heli Laurikkala, Heikki Vilkman, Mikko Ek, Hannu Koivisto, and Guang-Yu Xiong, "Modelling and control of supply chain with system theory," Tampere University of Technology, Institute of Machine Design 2003.

[22] MARK E. Nissen, " Agent-Based Supply Chain Integration," Information Technology and Management, vol. 2, pp. 289 - 312, 2001.

[23] Hartmut Stadtler and Christoph Kilger, Supply Chain Management and Advanced Planning: Concepts, Models, Software, and Case Studies: Springer, 2000.

[24] J. Dejonckheere, S. M. Disney, M. R. Lambrecht, and D. R. Towillb, "Measuring and avoiding the bullwhip effect: A control theoretic approach," European Journal of Operational Research vol. 147, pp. 567-590, 16June 2003 June 2003.

[25] Akintola Akintoye, "A survey of supply chain collaboration and management in the UK construction industry," European Journal of Purchasing & Supply Management, vol. 6, pp. 159-168 December 2000.

[26] M Khouja, "Synchronization in supply chains: implications for design and management," journal of the operations research society, pp. 984-994, 2003.

[27] Laura di Giacomo and Giacomo Patrizi, "Dynamic nonlinear modelization of operational supply chain, " Journal of Global Optimization, vol. 34, pp. 503-534, 2006.

[28] Zhang Lei, Li Yi-jun, and Xu Yao-qun, " Chaos Synchronization of Bullwhip Effect in a Supply Chain," in Management Science and Engineering, 2006. ICMSE '06. 2006 International Conference on, Lille, 2006, pp. 557-560.

[29] TEHILU LIAO, "Adaptive Synchronization of Two Lorenz Systems," Chaos,Solitons & Fractals, vol. 9, 1998.

[30] John David Crawford, "Introduction to bifurcation theory," Reviews of Modern Physics, vol. 63, pp. 991 - 1037, 1991.

[31] Higuchi Toru and D. Troutt Marvin, "Dynamic simulation of the supply chain for a short life cycle product: lessons from the Tamagotchi case," Comput. Oper. Res., vol. 31, pp. 1097-1114, 2004.

An Energy-Efficient Traffic-Aware Key Management Architecture for Wireless Sensor Networks

C.Gnana Kousalya and Dr.J. Raja

***Abstract*— In Wireless Sensor Networks (WSNs), most of the existing key management schemes, establish shared keys for all pairs of neighbor sensor nodes without considering the communication between these nodes. When the number of sensor nodes in WSNs is increased then each sensor node is to be loaded with bulky amount of keys. In WSNs a sensor node may communicate with a small set of neighbor sensor nodes. Based on this fact, in this paper, an energy efficient Traffic-Aware Key Management (TKM) scheme is developed for WSNs, which only establishes shared keys for active sensors which participate in direct communication. The proposed scheme offers an efficient Re-keying mechanism to broadcast keys without the need for retransmission or acknowledgements. Numerical results show that proposed key management scheme achieves high connectivity. In the simulation experiments, the proposed key management scheme is applied for different routing protocols. The performance evaluation shows that proposed scheme gives stronger resilence, low energy consumption and lesser end to end delay.**

***Index Terms*—Wireless sensor Network, Key management, Key Pre-distribution, Re-keying**

I. Introduction

The utilization of wireless sensor networks a tool for data aggregation and data processing has become increasingly efficient and popular. These tools aid in the monitoring of customary activities, environmental conditions and more besides aiding in cost effective administration of remote and hazardous locations. Close interaction of WSNs with their physical environment and unattended deployement of sensor nodes in hostile environment make WSNs highly vulnerable to attacks. Imparting security in wireless sensor networks is considered to be a tedious task.

WSNs is built with a large number of small battery powered device with limited energy, memory, computation and communication capabilities. Due to this insufficient resources in WSNs, Key management approaches used in Ad-Hoc and other wireless network cannot be applied to WSNs. From literature it is found that reasonable and accepted solution for key management in WSNs is to distribute randomly generated keys to each sensor node.

In wireless sensor networks, a sensor node may communicate with a small set of neighbor sensor nodes. Most of the existing key management schemes, did not consider this communication between these nodes. They establish shared keys for all pairs of neighbor sensor nodes. When the number of sensor nodes in WSNs is increased, large number of keys is to be loaded in each sensor node, which in turn causes more energy consumption. If any two close sensor nodes are rarely in the active-state the assignment of shared keys may be unnecessary, since they may be hardly exploited.

In this paper, a Traffic-Aware Key Management (TKM) scheme is proposed for WSNs, which only establishes shared keys for active sensor nodes which participate in direct communication, based on the topology information of the network. To inform about the state of a sensor node RTS/CTS control frames are modified from their original MAC. Proposed scheme reduces energy consumption with higher connectivity and stronger resilience against node capture.

The paper is organized as follows. Section 2 gives brief literature review on various key management schemes for WSN. Section 3 describes proposed key pre-distribution scheme. Section 4 gives the performance evaluation in terms of numerical and simulation results. Section 5 concludes the paper.

II. Related Work

Various key management schemes for WSNs are proposed for past few years. Wenliang Du et al. [2004] proposed key management using deployement knowledge. Alan price et al. [2004] proposed authentication and key distribution in one set of protocols .For Distributed Sensor Network (DSN) an alternative of random key pre-distribution scheme has been proposed by Siu-Ping Chan et al. [2005].Rui Miguel Soares Silva et al.[2006] proposed a scheme to overcome the disadvantages of the real symmetrical based systems using properties of chaotic systems. Grid-group deployment scheme has been proposed by Dijiang Huang et al. [2004]. "PKM", an in-situ key management protocol for sensor networks was proposed by F. Cheng et al. [2005].Jaemin Park et al.[2005] proposed random key pre-distribution scheme.

Neighbor-based authentication is explained briefly in literature. Sanzgiri et al.[2002] proposed the scheme in which the hash value of the packet corresponds to the decrypted value, the previous certificate is removed by the current node followed by the forwarding of the packet with the certificate of the current node.Both the target and intermediary participants were involved in the authentication of the data to be routed according to a fresh approach Ariadne proposed by Hu et al.[2002].Every node present in the source–destination path determines the authentication of the routing information with the aid of a Tesla key proposed by Perrig et al.[2002], in the course of the route discovery process.

Majority of the schemes use public key cryptography to attain security. But as the sensor nodes in wireless sensor networks are resource constraint the usage of public key cryptography in WSNs is not feasible.

Routing protocols in wireless network are explined briefly in literature. Charles E.Perkins et al.[1999] proposed AODV (Ad-Hoc On Demand Distance Vector Routing) reactive type routing protocol. Proactive type routing protocol DSDV (Destination Sequence Distance Vector Routing) is proposed by Charles E.Perkins et al.[1994] and DSR(Dynamic Source Routing) is proposed by David B.Johnson et al.[2002] From the literature it is found that Cluster formation to reduce the energy consumed is proposed in LEACH a hierarchical type routing protocol In another type of routing protocol PEGASIS, each sensor node communicates only with a close neighbor and takes turns in transmitting to the base station , thus reducing energy.

III. Proposed Key Management Scheme

The proposed Key management scheme is based on the state of sensor nodes. State of sensor nodes are categorized in to three types as follows: Current transmitting node (CTN), Transmitting node (TN), transmitting Node (CTN), Non transmitting Node (NTN).

In the proposed scheme RTS/CTS control frames is slightly modified from their original MAC protocol for informing a node the fact that its state is changed to TN or NTN in the corresponding period.

10 Bytes

Length	Type	To Addr	From Addr	Duration (NAV)	CRC

Fig1. a The Original RTS and CTS Frames

Length	Type	To Addr	From Addr	Duration (NAV)	Destination Addr	CRC

Length	Type	To Addr	From Addr	Duration (NAV)	Next TN Addr	CRC

Fig. 1. b) The Modified RTS and CTS Frames

The modified RTS and CTS frame add only one field of two bytes to the original frame. The newly added bytes in RTS is destination address and the newly added bytes of CTS is TN address

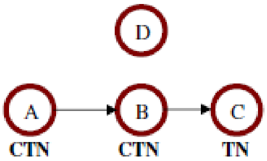


Fig. 2: Classification of Node States

Referring Figure 2, when node B receives A's modified RTS frame including the destination address of sink, its routing agent refers to the routing table for getting the next TN (node C) and informs back to its MAC. The node B then transmits modified CTS frame to node C which changes its state to TN and other neighbor nodes become aware of the fact that they are NTN nodes. Otherwise the routing path is broken or has not yet been established.

The Proposed Key management scheme consists of following phases:

i. Initial setup phase
ii.Pre-distribution phase
iii Shared Key discovery phase
iv.Path key establishment phase
v. Rekeying Phase

A. Initial Setup

Two keys, namely the Node key K and Network key NK are used in this scheme. The latter is utilized by the individual sensor nodes for the encryption and decryption purposes while the former is used by the key server node to unicast the node keys to the sensor nodes.

Sensor nodes agree on the following system parameters used in the protocol. The system parameters include

Global Key Pool: Defined as a pool of random symmetric keys from which a group key pool is generated. Keys are generated using one way function F, where n is chosen to be large.

Ki = F(Ki+1) i=1,2,3….n

Group Key Pool: Defined as a subset of Global key pool for a given group.

Key Ring: Defined as a subset of group key pool, which is independently assigned to each sensor node.

Key-Sharing Graph: Let V represent all the nodes in WSN. A *Key-Sharing graph* G (V, E) is constructed in the following manner: For any two nodes i and j in V, there exists an edge between them if and only if (1) nodes i and j have at least one common key, and (2) nodes i and j can reach each other within the wireless transmission range, *i.e.*, in a single hop.

B. Key Pre-Distribution Phase

This phase is performed off-line and before the deployment of sensor nodes. Primarily Gi (i=1,2…k) group key pools are produced using global key pool S. After this, for each sensor node in a group, a key ring from a group key pool Gi is assigned along with a variable.

C. Shared-Key Discovery Phase

This phase is used to find a secure link between two sensor nodes. Sensor nodes which identify its shared keys in their key rings, then verify that other CTN and TN node contain these keys. Now the shared key turns out to be the key for that link. A key-sharing graph is created by the entire sensor networks following above step. The execution of the shared key discovery phase is completed by a CTN node, if it finds out a TN node as a neighbor.

D. Path -Key Establishment Phase

Sensor nodes can form path keys with their neighbor nodes since they have not shared keys inside their key rings. A path can be established from a source sensor node to other CTN and TN sensor nodes, if the key-sharing graph is connected. A path key can be generated by the source node and send it safely using a path to the target sensor node.

E. Re-keying Phase

This Phase uses two control packets INIT and UPDATE .The command node prepares a control packet INIT which contains

INIT: (L , K_{i+1} , T_{rekey}) , MAC(L , K_{i+1} , T_{rekey})
L – length of the key

Ki - initial key
T_{rekey} - Rekeying interval of Ki

This control packet is encrypted with network key NK and send to every sensor nodes

Command node→ E_{NK}(INIT)

Once the INIT packet is received, a sensor node resets all previous keys. It then calculates new keys Ki , … K1 from K_{i+1}. The subsequent key in the key-sequence is broadcasted by the command node periodically with the aid of UPDATE control packet. The node keys are disclosed by the command node in a periodic manner from the K_{L+2} to all nodes in the group. At time Tstart + T_{rekey} ,the server broadcasts UPDATE packets containing K_{i+L+2}, i=1,2,……n-L-2.

Command node → group: $E_{Ki+1}(K_{i+L+2})$

Where $E_{Ki+1}$is the active encryption key at the time when UPDATE packet is broadcasted.

The UPDATE packet is discarded once the node detects that it is not from its own server. If not, the UPDATE packet is broadcasted to all the neighbors.

IV. Performance Evaluation

A. Evaluation Metrics

In the proposed scheme following evaluation metrics are considered:

Connectivity: The probability that two sensors share at least one common key at a given time-interval should be higher, with smaller number of keys.

Resilience against Node Capture: Exposing of the secret information regarding other nodes should be made certain by the key establishment technique, if a node inside a sensor network is confined.

Any efficient key management scheme for WSNs should have higher connectivity and stronger resilience

B. Numerical Results

Connectivity

It is defined as the probability *(Ps)* that two TN or CTN state sensor nodes share atleast a common key after deployement at a given time interval.

Let φ is the set of all sensor node groups and two nodes Ni and N_J are selected from Gi and Gj of φ. The probability that Ni and N_j are in TN state at given time-interval, and two nodes share at least one common key is given by $P_{S.}$ Using Baye's Theorem,

$$P_s = \frac{\sum_{i\in\varphi} P_1(T_i)\,.\,\mathrm{P}_3(Sh)}{\sum_{i\in\varphi} P_1(T_i)} \qquad (1)$$

Where,

$P_1(T_i)$ – Probability of group Gi at a time interval *T_i*

$P_3(Sh)$ - Probability that two nodes share at least one common key

The probability that two nodes are in TN state at a given time-interval Ti is calculated using

$$f_p(t_m^i) = \frac{e^{-a}a^{t_m^i}}{x!} \qquad (2)$$

Therefore the active-probability of Gi at Ti can be found as follows

$$P_1(T_i) = f_p(t_m^{i+1}) - f_p(t_m^i)$$

$$= \frac{e^{-a}a^{t_m^{i+1}}}{t_m^{i+1}!} - \frac{e^{-a}a^{t_m^i}}{t_m^i!}$$

$$= e^{-a}\left[\frac{a^{t_m^{i+1}}}{t_m^{i+1}!} - \frac{a^{t_m^i}}{t_m^i!}\right] \qquad (3)$$

The probability that two nodes share at least one common key is expressed as

1- Pr[two sensors do not share any key]. (4)

Consider

Total size of each group = M

Shared keys = $Sh(M)$

Non-Shared keys = $M - Sh(M)$

Let n_1, n_2 be two sensor nodes. When n_1 select x keys from $Sh(M)$ keys and y keys from $M - Sh(M)$ keys, then n_2 select z keys from (M-x) Keys.

Pr[two sensors do not share any key] is given by

$$\therefore P_2(NSh) = \frac{\sum_{x=o}^{\min(Z,Sh(M))} \binom{Sh(M)}{\mathrm{x}}\binom{M-Sh(M)}{\mathrm{y}}\binom{M-x)}{\mathrm{z}}}{\binom{M}{z}^2}$$

$$Now\ \mathrm{P}_3(Sh) = 1 - P_2(Nsh)$$

$$= 1 - \frac{\sum_{x=0}^{Min(Z,Sh(M)} \binom{Sh(M)}{\mathrm{x}}\binom{M - Sh(M)}{y}\binom{M - x}{z}}{\binom{M}{z}^2}$$

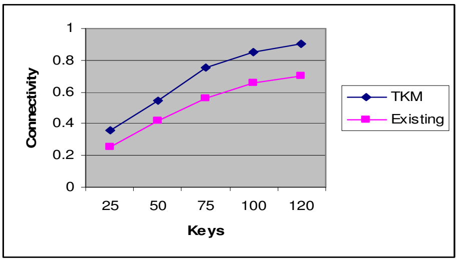


Fig. 3. Connectivity Vs No. of Keys

Figure.3 gives the connectivity with respect to the varied number of keys in each sensor. The proposed scheme is compared with the existing random key pre-distribution scheme of Mohamed F. Younis et al.'s [2006]. It is found that lesser number of keys is involved in the proposed scheme to achieve the same probability.

C. Simulation Results

NS2 simulator is used for simulation with following specifications:

- Maximum Number of nodes is 80
- The deployment area is 500mx500 m.

- Simulation time is 100 seconds.
- The transmission range of 250 meters with Constant Bit Rate (CBR).

The proposed key management is applied with routing protocols AODV, DSDV, LEACH and PEGASIS and simulated to find resilence, energy consumed and end to end delay performance.

Effects of Resilience against Node Capture

An adversary can attack on a sensor node after it is deployed to read the information. To find how a successful attack on n sensor nodes by an adversary affects the rest of the network resilence is used. Resilence is calculated from the fraction of communication among the uncompromised nodes that an adversary can compromise based on the information retrieved from the n captured nodes. Using the routing protocols AODV, DSDV, LEACH, and PEGASIS, resilience is measured for the proposed TKM scheme with varying number of nodes and attackers and compared with SHELL proposed by Mohemed F.Younis et al.[2006].

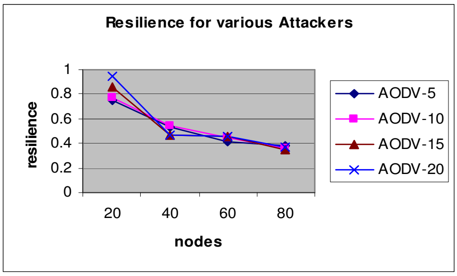


Fig 4.a.Resilence Vs Nodes-AODV

Fig 4.a shows the resilence with TKM using routing protocol AODV.With increase in the number of nodes from 20 to 80 nodes and increase in number of attackers from 5 to 20 attackers the resilence is reduced by 55% to 60%.

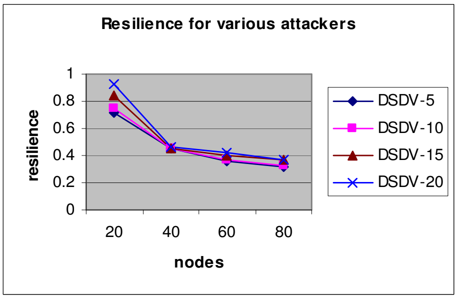


Fig. 4.b.Resilence Vs Nodes-DSDV

Fig 4.b shows the resilence with TKM using routing protocol DSDV. With increase in the number of nodes from 20 to 80 nodes and increase in number of attackers from 5 to 20 attackers the resilence is reduced by 55% to 61%.

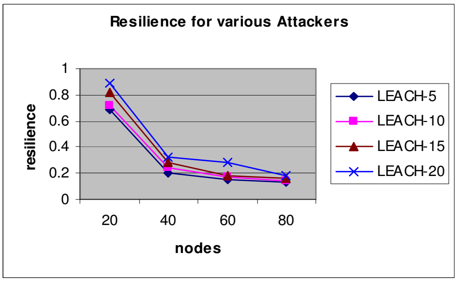


Fig. 4.c.Resilence Vs Nodes-LEACH

Fig 4.c shows the resilence with TKM using routing protocol LEACH.With increase in the number of nodes from 20 to 80 nodes and increase in number of attackers from 5 attackers to 20 attackers the resilence is reduced by 79% to 81%.

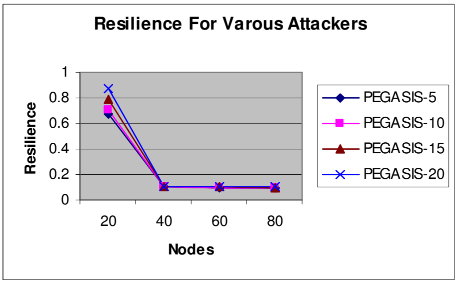


Fig. 4.d.Resilence Vs Nodes-PEGASIS

Fig 4.d shows the resilence with TKM using routing protocol PEGASIS.With increase in the number of nodes from 20 to 80 nodes and increase in number of attackers from 5 to 20 attackers the resilence is reduced by 86% to 88%.

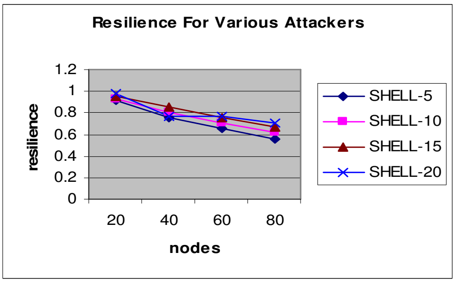


Fig. 4.e.Resilence Vs Nodes-SHELL

Fig 4.e shows the resilence with SHELL. With increase in the number of nodes from 20 to 80 nodes and increase in number of attackers from 5 to 20 attackers the resilence is reduced only by 28% to 38%.

It is found from fig 4.a-e the performance of resilence is best in TKM-PEGASIS and hence more secure when compared with TKM using LEACH, AODV, DSDV and SHELL.

Effects of Energy Consumption against Node Capture

Energy consumed by the network is obtained by varying total number of nodes and attackers with TKM using routing protocols AODV, DSDV, LEACH and PEGASIS. Proposed TKM scheme is compared with SHELL.

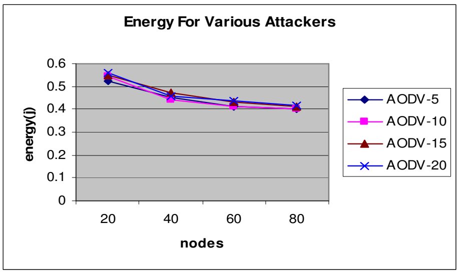


Fig 5.a. Energy Consumption Vs Nodes -AODV

Fig 5.a shows the energy consumed with TKM using routing protocol AODV. It is observed that energy consumed by the network using TKM-AODV is reduced by 44% to 56% when compared with SHELL with increase in nodes from 20 to 80 nodes and attackers from 5 to 20 attackers

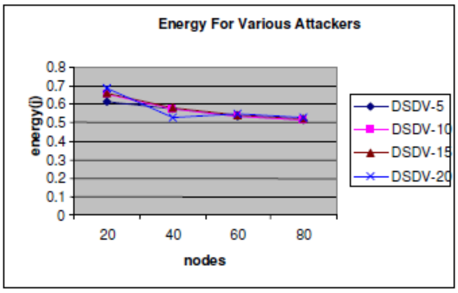


Fig 5.b. Energy Consumption Vs Nodes –DSDV

Fig 5.b shows the energy consumed with TKM-DSDV.With increase in the number of nodes from 20 nodes to 80 nodes and increase in number of attackers from 5 attackers to 20 attackers the energy consumed is reduced by 43% to 47% when compared with SHELL

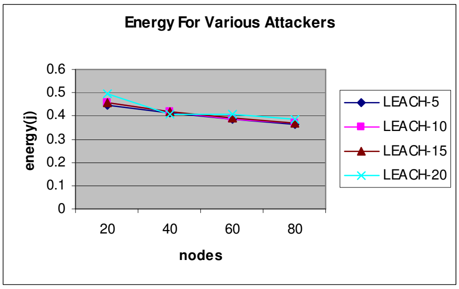


Fig 5.c. Energy Consumption Vs Nodes –LEACH

Fig 5.c shows the energy consumed with TKM- LEACH. Number of nodes is increased from 20 nodes to 80 nodes and the number of attackers is also increased from 5 attackers to 20 attackers and it is observed that the energy consumed is reduced by 58% to 62% when compared with SHELL

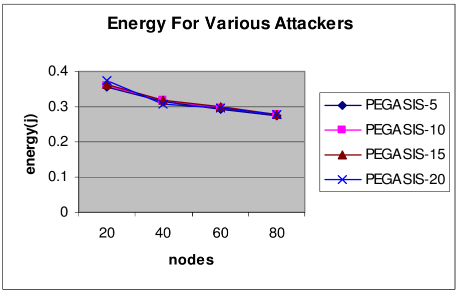


Fig 5.d. Energy Consumption Vs Nodes -PEGASIS

Fig 5.d shows the energy consumed with TKM using routing protocol PEGASIS. With increase in the number of nodes from 20 nodes to 80 nodes and increase in number of attackers from 5 attackers to 20 attackers the energy consumed is reduced by 69% to 71% when compared with SHELL.

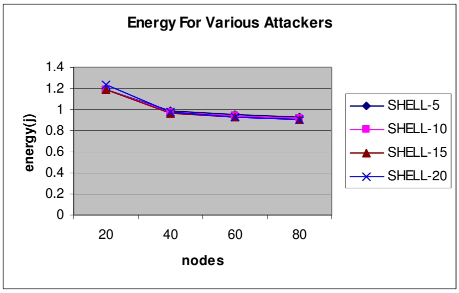


Fig. 5.e. Energy Consumption Vs Nodes –SHELL

From figure 5.a to 5.e it is observed that TKM-PEGASIS consumes less energy for specific transmission when compared with TKM using LEACH, AODV, DSDV and SHELL.

Effects of End to End Delay against Node Capture

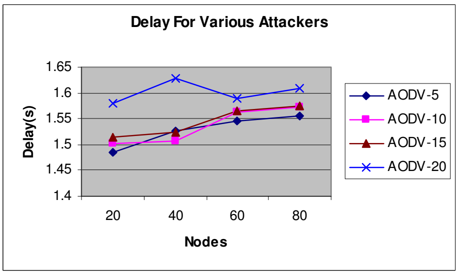


Fig. 6.a. Delay Vs Attackers

Fig 6.a shows that the end to end delay is reduced by 7% to 10%.using TKM-AODV when compared with SHELL with increase in number of nodes from 20 to 80 nodes and number of attackers from 5 to 10 attackers.

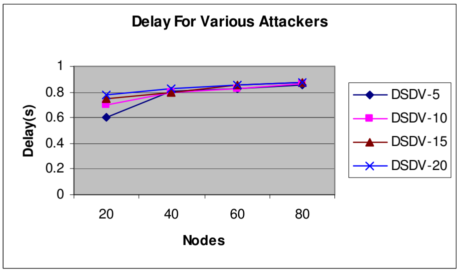


Fig. 6.b. Delay Vs Attackers

Fig 6.b shows that the end to end delay is reduced by 49% to 63% with TKM-DSDV when compared with SHELL with increase in the number of nodes from 20 nodes to 80 nodes and number of attackers from 5 to 20 attackers.

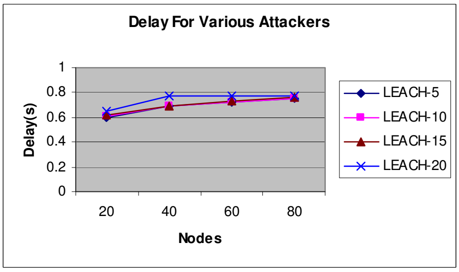


Fig. 6.c. Delay Vs Attackers

Fig 6.c.shows that the end to end delay is reduced by 54% to 61% with TKM-LEACH when compared with SHELL with increase in the number of nodes from 20 nodes to 80 nodes and number of attackers from 5 attackers to 20 attackers.

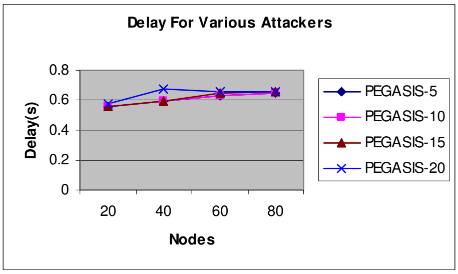


Fig. 6.d.Delay Vs Attackers

Fig 6.d shows that the end to end delay is reduced by 61% to 65% with TKM-PEGASIS when compared with SHELL with increase in the number of nodes from 20 nodes to 80 nodes and number of attackers from 5 attackers to 20 attackers.

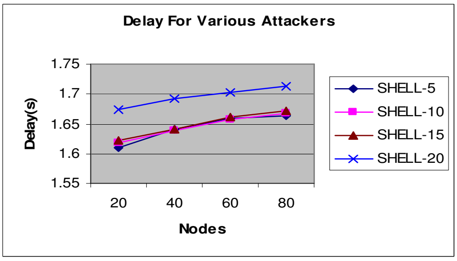


Fig. 6.e. Delay Vs Attackers

From figure 6.a-e it is observed that end to end delay is reduced more in TKM –PEGASIS when compared with TKM using LEACH, AODV, DSDV and SHELL.

V. Conclusion

The proposed scheme establishes shared keys for active sensor nodes which participate in direct communication, based on the topological information of the network. This scheme provides seamless re-keying without disrupting the ongoing security process. Numerical results show that the proposed scheme achieves high connectivity.The simulation is performed for the proposed scheme with different routing protocols. Performance analysis shows that proposed key management scheme TKM with PEGASIS achieves stronger resilience low energy consumption and lesser end to end delay when compared with SHELL.

References

[1] Wenliang Du Jing Deng Han, Y.S. Shigang Chen Varshney, P.K. “A Key Management Scheme for Wireless Sensor Networks Using Deployment Knowledge” INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies 7-11 March 2004.

[2] Alan Price, Kristie Kosaka and Samir Chatterjee “A Secure Key Management Scheme for Sensor Networks” Proceedings of the Tenth Americas Conference on Information Systems, New York, New York, August 2004.

[3] Siu-Ping Chan, Radha Poovendran and Ming-Ting Sun “A Key Management Scheme in Distributed Sensor Networks Using Attack Probabilities” Global Telecommunications Conference, 2005.GLOBECOM '05. 28 Nov.-2 Dec. 2005

[4] Rui Miguel Soares Silva, Nuno Sidónio Andrade Pereira and Mário Serafim Nunes "Chaos Based Key Management Architecture for Wireless Sensor Networks", Australian Telecommunication Networks and Application Conference [ATNAC 2006], December 4-6, 2006.

[5] Dijiang Huang, Manish Mehta, Deep Medhi and Lein Harn “Location Aware Key Management Scheme for Wireless Sensor Networks” Proc. of 2004 ACM Workshop on Security of Ad Hoc and Sensor Networks (SASN'04), pp. 29-42, October 2004

[6] An, F. Cheng, X. Rivera, J. M. Li, J. Cheng, Z. “PKM: A Pairwise Key Management Scheme for Wireless Sensor Networks” Lecture Notes In Computer Science 2005, Numb 3619, pages 992-1001.

[7] Jaemin Park, Zeen Kim, and Kwangjo Kim “State-Based Key Management Scheme for Wireless Sensor Networks” Mobile Adhoc and Sensor Systems Conference, 2005. IEEE International Conference on 7-10 Nov. 2005.

[8] K. Sanzgiri, Bridget Dahill, B. Levine, C. Shields, and E. Belding-Royer.”Secure routing Protocol for Ad Hoc Networks”. In Proceedings of the IEEE International Conference on Network Protocols, 2002

[9] Y. Hu, A. Perrig, and D. Johnson. “Ariadne: A secure on-demand routing protocol for ad hoc networks”. In Proceedings of the International Conference on Mobile Computing and Networking (MobiCom), 2002

[10] A. Perrig, R. Canetti, D. Tygar, and D. Song. "The TESLA Broadcast Authentication Protocol". In RSA CryptoBytes, volume 5(2), pages 2–13, 2002

[11] Charles E. Perkins, Elizabeth M. Royer "Ad hoc On Demand Distance Vector Routing" Mobile Computing Systems and Applications, 1999. Proceedings. WMCSA '99. Second IEEE Workshop on Publication Date: 25-26 Feb 1999.

[12] C.E. Perkins and P.Bhagwat. "Highly Dynamic Destination-Sequenced Distance-Vector routing (DSDV) for mobile computers". In Proceedings of the SIGCOMM'94 conference on Communications, Architectures, Protocols, and Applications, August 1994.

[13] David B. Johnson and David A. Maltz "Dynamic Source Routing in Ad Hoc Wireless Networks" Wiley Series On Parallel And Distributed Computing, Pages: 425 – 450,Year of Publication: 2002 ISBN:0-471-41902-8.

[14] Changsu Suh, Young-Bae Ko and Dong-Min Son, "An Energy Efficient Cross-Layer MAC Protocol for Wireless Sensor Networks," Proc. of the International Workshop on Sensor Networks (IWSN'06) in APWeb06, Jan. 2006. (LNCS),

[15] Mohamed F. Younis , Kajaldeep Ghumman and Mohamed Eltoweissy "Location-Aware Combinatorial Key Management Scheme for Clustered Sensor Networks" , IEEE transactions on parallel and distributed systems, Vol. 17, No. 8, August 2006.

An Effective Scheme for Generating Irrevocable Cryptographic Key from Cancelable Fingerprint Templates

N. Lalithamani and Dr. K.P. Soman

Abstract—**Unswerving information security mechanisms are the need of the hour for fighting the escalating enormity of identity theft in our society. Besides cryptography being a dominant tool in attaining information security, one of the key confronts in cryptosystems is to preserve the secrecy of the cryptographic keys. The incorporation of biometrics with cryptography will be an effective solution to this problem. Recently generating cryptographic key from biometrics has gained enormous popularity in research community due to its improved performance in providing security. Nevertheless, a biometric is enduringly connected with a user and cannot be altered. Thus, when a biometric identifier is compromised, it is lost everlastingly and probably for every application where that particular biometric is employed. Cancelable biometrics intends to resolve this by building revocable biometric templates. In this paper, we have proposed an effective scheme for generating irrevocable cryptographic key from cancelable fingerprint templates. Initially the minutiae points are extracted from the fingerprints. Afterwards, cancelable templates are generated and irrevocable keys are extracted from the cancelable templates. As the cryptographic key is generated in an irreversible manner, obtaining cancelable fingerprint templates and original fingerprints from the generated key is impossible. We have evaluated the effectiveness of our scheme using fingerprints from publicly available sources. We have also presented the security analysis of the proposed scheme.**

Index Terms—**Biometrics, Cancelable Biometrics, Cryptography, Biometric cryptosystems, Key generation, Irrevocable Key, Fingerprint, Minutiae Points.**

I. Introduction

Protecting personal privacy and preventing identity theft are of national precedence. These goals are indispensable to our democracy and our economy, and intrinsically significant to our citizens. Biometrics, a budding set of methodologies, assures an efficient solution. In the domain of computer security, biometrics denotes the authentication techniques that depend on quantifiable physiological and individual features that can be automatically demonstrated. Despite the fact that the field of biometrics is still in its formative years, it's unavoidable that biometric systems will play a significant role in the future of security [1]. A biometric system is fundamentally a pattern recognition system that functions by obtaining biometric data from an individual, extracting a feature set from the obtained data, and evaluating this feature set against the template set in the database [2]. The biometric data comprises of fingerprints [3], facial features [4], iris [5], hand geometry [6], voice [7], signature [8] and the like. Biometrics is extensively employed in forensics, in criminal identification and prison security to quote a few of the instances, and has the prospective to be employed in a wide variety of civilian application areas.

Throughout the last decade biometrics has gained popularity in application employed for identifying individuals. The accomplishment of its relevance in user authentication has signified that numerous benefits could possibly be obtained by integrating biometrics with cryptography [9]. The incapability of human users to keep in mind the powerful cryptographic keys has been an issue restricting the security of systems for decades. This restriction could be resolved in a huge variety range of applications by producing strong cryptographic keys from biometric data, possibly in combination with the entry of a password [10, 11, 12]. Biometric features are highly complicated to duplicate or falsify and impracticable to share. These characteristics of biometrics influence their utilization in cryptographic key generation. In the recent past, researchers have shifted their attention towards merging biometrics with cryptography in order to enhance overall security, by eliminating the necessity for key storage using passwords [8, 4, 9].

The systems that combine biometrics with cryptographic security are known as Biometric cryptosystems, or Crypto-biometric systems [13]. The incorporation of biometrics with cryptography can be widely carried out at two different levels. Considering biometrics-based key release, a biometric matching between an input biometric signal and an enrolled template aids in the release of the secret key. The biometric signals are found to be monolithically bounded to the keys in case of biometrics-based key generation [14]. The principal complexity in biometric cryptosystems lies in the accommodation of the deviations intrinsic in measuring biometrics, or in the biometrics themselves, despite iteratively generating the same key [7]. Lately numerous researchers have attempted to design cryptosystems on basis of biometrics to eradicate some of the problems however have not yet been victorious in exploiting the power of biometrics in a complete manner [15].

Despite possessing benefits including the non-repudiation and convenience of utilization and the like, biometrics comprises certain issues [13] that limit its utilization as a key to a cryptosystem. A significant issue in the utilization of biometrics is that the number of biometrics that can be acquired from a person is restricted and the conciliation of it would mean that that specific biometric is rendered futile perpetually. Once the biometric features are lost, a replacement is unfeasible. It is entirely evident, once an old fingerprint is gone missing, a new fingerprint for the same person is not attainable. It is possible to revoke or replace compromised credit cards and passwords however biometrics are enduringly connected with a user and are impossible to replace. Cancelable biometrics [16] has been projected in literature so as to address this issue. Cancelable biometrics intends to resolve this by building revocable biometric templates [17], [18].

Cancelable biometric templates are necessary for biometric authentication systems, particularly for the ones that are operated under unsupervised and/or over networked environments [19], [20]. In case of cancelable biometrics, the biometric image is distorted in a repeatable yet nonreversible fashion prior to the generation of the template. When the cancelable template is compromised, the distortion characteristics are altered, and the same biometrics is mapped to a fresh template, which is utilized consequently. A cancelable biometric template needs to accomplish four significant criteria before being measured as valuable [21]. Diversity: No equivalent cancelable template can be utilized in two distinct applications. Reusability: Straightforward revocation and reissue at the occurrence of compromise. One-way transformation: Non-invertibility of template computation to avoid recovery of biometric data. Performance: The formulation should not worsen the recognition performance.

This paper discusses an effective scheme for generating irrevocable cryptographic key from cancelable fingerprint templates. As the fingerprints are one of the most widely used biometric modality today, we have employed the fingerprint biometrics in our scheme. Owing to the fact that majority of the fingerprint authentication systems work on basis of minutiae, which are feature points obtained from a raw fingerprint image, we have utilized the minutiae points in the cancelable fingerprint template formation and cryptographic key generation. As discussed above, initially the minutiae points are extracted from the fingerprint images using the approach discussed. Then, the cancelable fingerprint templates are formed from the extracted minutiae points. Subsequently, the irrevocable cryptographic key is generated from the cancelable fingerprint template using the proposed approach. The fingerprint images from publicly available sources are used in evaluating the proposed scheme. The security analysis of the proposed scheme is also presented.

The rest of the paper is organized as follows. A brief review of the works related to the proposed scheme is given in Section 2. The approach to extract the minutiae points from the fingerprint image is discussed in Section 3. The cancelable fingerprint template generation from the extracted minutiae points is explained in Section 4. The irrevocable cryptographic key generation from the cancelable fingerprint template is presented in Section 5. Security Analysis of the proposed scheme is presented in Section 6 and experimental results are given in Section 7. Finally, the conclusions are summed up in Section 8.

II. Review of Related Works

Our work is inspired by a number of previous works related to cryptographic key generation from biometrics and cancelable biometrics. A brief review of some of the works is given below:

A cancelable biometric approach, called PalmHashing was projected by Connie Tee et al [22] in order to solve the problem non-revocable biometric. The method hashes palmprint templates with a set of pseudo-random keys to arrive at a distinctive code known as palmhash. It is possible to store the palmhash code in portable devices like tokens and smartcards for verification purposes. Furthermore, PalmHashing provides numerous advantages over present-day biometric approaches like, unambiguous partition of the genuine-imposter populations and zero EER occurrences. They also delineated the implementation details of the method besides emphasizing its potentials in security-critical applications.

A practical and secure way to integrate the iris biometric into cryptographic applications was presented by Hao, F. et al [23]. They deliberated the error patterns within iris codes and introduced a two-layer error correction technique that merges Hadamard and Reed-Solomon codes. The key was produced from a subject's iris image using auxiliary error-correction data that do not disclose the key and can be stored in a tamper-resistant token, like a smart card. They assessed the system with the aid of iris samples from 70 different eyes, with 10 samples from each eye. They figured out that it is possible to reproduce an error-free key dependably from genuine iris codes with a 99.5 percent success rate.

The application of handwritten signature to cryptography on basis of recent works displaying the probability of key generation employing biometrics was studied by M. Freire-Santos et al [14]. A cryptographic construction known as fuzzy vault was implemented in the signature-based key generation scheme. The utilization of distinguishing signature characteristics suited for the fuzzy vault was conferred and evaluated. The results of experimentation were reported along with the error rates involved in releasing the secret data with the aid of both random and skilled forgeries from the MCYT database.

A two-factor cancelable formulation was proposed by Teoh AB et al [25], where in, the biometric data are distorted in a revocable but non-reversible fashion by initially converting the raw biometric data into a fixed-length feature vector and then projecting the feature vector onto a sequence of random subspaces that were obtained from a user-specific pseudorandom number (PRN). The procedure was revocable and made the replacement of biometrics appear as simple as replacing PRNs. The formulation was established under numerous situations (normal, stolen PRN, and compromised biometrics scenarios) with the aid of 2400 Facial Recognition Technology face images.

A straightforward mechanism for the generation of digital signatures and cryptography communication through the aid of biometrics was proposed by Je-Gyeong Jo et al [26]. It is necessary to generate the digital signature in such a way that it can possibly be verified by the prevailing cryptographic algorithm like RSA/ElGamal without altering its own security requirement and infrastructure. It was anticipated that the mechanism will guarantee security on the binding of biometric information in the signature system on telecommunication environments.

Julien Bringera et al [27] have presented the Cancelable biometrics and secure sketches with an identical purpose in mind: to guard the privacy of biometric templates besides maintaining the ability to match the protected data against a reference. The standard beyond cancelable biometrics was to carry out an irreversible transformation over images and to create matching over transformed images. They demonstrated that applying secure sketch error correction to cancelable biometrics permits one to maintain good matching performance.

The concept of cancelable biometrics was proposed by Andrew B. J. Teoh et al. [28] to express biometric templates that can be cancelled and restored with the addition of another independent authentication factor. A kind of cancelable biometrics that merges a set of user-specific random vectors with biometric features is known as BioHash. The quantized random projection collection on basis of the Johnson-Lindenstrauss Lemma was employed to accomplish the mathematical foundation of BioHash. Depending upon this model, they have explained the characteristics of BioHash in pattern recognition in addition to security view points and provided some methods to resolve the stolen-token problem.

A.T. Beng Jin and Tee Conniea [29] have proposed the Cancelable biometrics in order to describe biometric templates that can possibly be canceled and replaced. A kind of cancelable biometrics that merges a set of user-specific random vectors with biometric features is known as BioHash. The chief disadvantage of BioHash was its immense deprivation in performance when the legitimate token is stolen and is utilized by the pretender to claim as the legitimate user. A modified probabilistic neural network was utilized by them as a classifier to address the aforesaid issue.

Biometric-key generation is a procedure to transform a piece of live biometric data into key with the aid of auxiliary information that is also known as a biometric helper. It is possible to repetitively generate a biometric-key and it is not necessary to physically store the biometric. Beng, A. et al. [30] presented a biometric-key generation system which worked on basis of a randomized biometric helper. The scheme comprises of a randomized feature discretization process and a code redundancy construction. The former facilitates one to manage the intra-class variations of biometric data to the minimal level and the latter additionally lessens the errors. The randomized biometric helper guarantees that a biometric-key was simple to be rescinded when the key was compromised.

The production of biometric keys directly from live biometrics, under specific criteria, by dividing feature space into subspaces and further dividing these into cells, where each cell subspace contributes to the overall key produced, was illustrated by Sanaul Hoque et al. [31]. They assessed the scheme on real biometric data, denoting both genuine samples and attempted limitations. Experimental evaluations illustrated the extent to which the technique can be implemented dependably in possible practical situations.

III. Extraction of Minutiae Points from Fingerprints

The extraction of minutiae points from the fingerprint image is discussed in this section. It is supposed that fingerprints are distinct across individuals and across the fingers of a particular individual [32]. It has been established that even identical twins with identical DNA possess different fingerprints. Since many existing fingerprint authentication systems are based on minutiae points, which are feature points extracted from a raw fingerprint image, we have employed the minutiae points in our scheme as well. A fingerprint can be defined as a pattern of ridges and valleys on the tip of the finger. A fingerprint is therefore described by the distinctiveness of the local ridge features and their relationships. Minutiae points denote these local ridge characteristics that appear either at a ridge ending or a ridge bifurcation. The point where the ridge comes to an abrupt end is known as ridge ending and the ridge bifurcation is denoted as the point where the ridge divides into two or more branches.

The major steps involved in the minutiae points extraction are as follows:

- Segmentation
- Orientation Field Estimation.
- Image Enhancement
- Minutiae Extraction

A. Segmentation

The first step in the minutiae points extraction is segmentation. The input fingerprint image is segmented from the background to actually extract the region comprising the fingerprint, which ensures the removal of noise. Segmentation of an image represents the division or separation of the image into regions that have similar attributes. At first, the image is preprocessed. The preprocessing phase includes the following: histogram equalization and median filtering. Later, the preprocessed image is divided into blocks and segmentation is carried out. The sample fingerprint images are shown in Figure 1.

Fig. 1. Two Sample Fingerprint Images

Preprocessing

The preprocessing of fingerprint images includes the following:

(i) Histogram Equalization
(ii) Median Filtering

(i) Histogram Equalization:

Histogram equalization amplifies the local contrast of the images, particularly when they are represented with very close contrast values. It is possible to distribute intensity through the histogram with the aid of this regulation. Histogram equalization utilizes a monotonic, non-linear mapping that re-assigns the intensity values of pixels in the input image in such a manner that the output image comprises a uniform distribution of intensities (i.e. a flat histogram). The original histogram of a fingerprint image is of bimodal type, the histogram after the histogram equalization transforms all the range from 0 to 255 which results if enhanced visualization effect [33]. The results of histogram equalization are depicted in Figure 2.

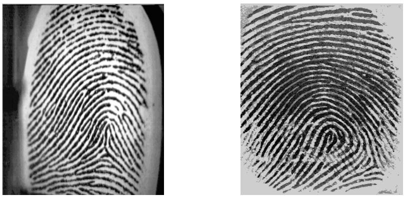


Fig 2: Fingerprint Images after Histogram Equalization

(ii) Median Filtering:

The median filter is a non-linear digital filtering methodology frequently employed to eliminate noise from images or other signals. This is carried out with the aid of a window comprising of an odd number of samples. The values present within the window are arranged into numerical order; the median value, the sample in the center of the window, is chosen as the output. The oldest sample is abandoned, a new sample is obtained, and the calculations are redone [34]. The filtering process is applied to the fingerprint image obtained as a result of the previous step by spatially convolving the image with the filter. The results of median filtering are shown in Figure 3.

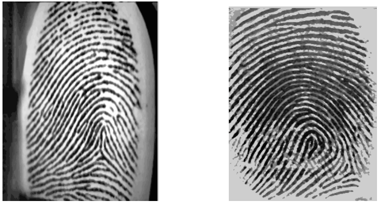


Fig. 3: Fingerprint images after median filtering.

Segmentation

The image obtained after preprocessing has high contrast and enhanced visibility. Subsequently, the preprocessed fingerprint image is divided into non-overlapping blocks of size 16x16 followed by the calculation of gradient of each block. The standard deviation of gradients in X and Y direction is computed and summed. Eventually, the resultant value is compared against a threshold value. If it is greater than the threshold value the block is filled with ones, otherwise the block is filled with zeros. In our scheme, the threshold value is set as 20.

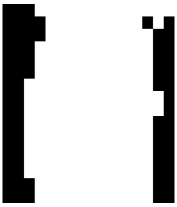


Fig 4: Segmentation result of fingerprint image

B. Orientation Field Estimation

The next step in the extraction of minutiae points is the estimation of orientation field. A fingerprint field orientation map may be described as an ensemble of two-dimensional direction fields. They denote the directions of ridge flows in regular spaced grids. It is possible to neglect the magnitudes of these fields and the angle information alone is of interest [35]. Basically there are two methodologies to compute the orientation field of fingerprint namely the filter-bank based approaches and gradient-based approaches. We have employed gradient based approaches in our scheme. In case of the gradient-based methods, initially, the gradient vectors are determined by considering the partial derivatives of image intensity at every pixel. The gradient vectors can be represented as $[g_x, g_y]^T$. In case of a fingerprint image, the gradient vectors indicate the highest deviation of gray intensity that is perpendicular to the edge of ridge lines [35]. Conventionally, an orientation map is denoted in the form of a matrix $\{\theta_{XY}\}$, where $\theta_{XY} \in [0, \pi]$. The orientation θ is orthogonal to $\overline{\varphi}$, in which $\overline{\varphi}$ is the dominant gradient angle of a local base block.

C. Image Enhancement

Following the orientation field estimation, the fingerprint image is enhanced to extract the minutiae points effectively. The enhancement of fingerprint images involves the following: Average filtering and Gabor filtering. Initially the image is filtered with the help of average filter to correct the frequency of the image. Subsequently, Gabor filtering is applied to the image for further enhancement.

Averaging Filter

The impact of noise can be decreased through simple averaging. Provided a noisy yet bounded measurement sequence it is possible for us to take a huge number of readings of the variable and employ its average to provide improved estimate of its true value (given that there is no systematic error or bias in the measurements). This is in fact the standard process in experimental work, where numerous readings are taken at a sampling instant and the average of these readings utilized as the measurement [36].

Gabor Filter

A Gabor filter can be described as a linear filter whose impulse response is given by a harmonic function multiplied by a Gaussian function. Owing to the multiplication-convolution property (Convolution theorem), the Fourier transform of a Gabor filter's impulse response is the convolution of the Fourier transform of the harmonic function and the Fourier transform of the Gaussian function [37].

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

Where,

$$x' = x\cos\theta + y\sin\theta$$

and

$$y' = -x\sin\theta + y\cos\theta$$

Here, λ denotes the wavelength of the cosine factor, θ denotes the orientation of the normal to the parallel stripes of a Gabor function, Ψ corresponds to the phase offset, γ denotes the spatial aspect ratio and enumerates the ellipticity of the support of the Gabor function. The Gabor filter comprises of both frequency-selective and orientation-selective properties and constitutes optimal joint resolution in both spatial and frequency domains. Thus the Gabor filter is capable of removing the noise and conserve true parallel ridges structures taking the benefit of the local orientation and local frequency [38].

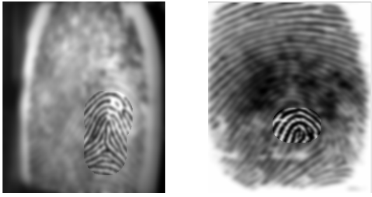


Fig 5: Enhanced Images Clearly showing Delta Region

D. Minutiae Points Extraction

Finally, the minutiae points are extracted from the enhanced fingerprint image. The steps involved in the extraction of minutiae points are as follows:

- Binarization
- Morphological Operations
- Minutiae points extraction

Initially, the enhanced image is binarized. After binarization, morphological operations are performed on the image to remove the obstacles and noise from it. Finally, the minutiae points are extracted using the approach discussed.

Binarization

The binary images with only two levels of interest: The black pixels that denote ridges and the white pixels that denote valleys are employed by almost all minutiae extraction algorithms. A grey level image is translated into a binary image in the process of binarization, by which the contrast between the ridges and valleys in a fingerprint image is improved. Hence, the extraction of minutiae is achievable. The grey-level value of every pixel in the enhanced image is analyzed in the binarization process. Then, the pixel value is set to a binary value one when the value is greater than the global threshold, or else a zero is set as the pixel value. The foreground ridges and the background valleys are the two level of information held by the ensuing binary image. Removal of distortions present in the image is performed followed by the retrieval of the exact skeleton image from the image.

Morphological Operation

The binary morphological operators are applied on the binarized fingerprint image. Elimination of any obstacles and noise from the image is the primary function of the morphological operators. Furthermore, the unnecessary spurs, bridges and line breaks are removed by these operators. Then thinning process is performed to reduce the thickness of the lines so that the lines are only represented except the other regions of the image. Clean operator, Hbreak operator, Spur operator and Thinning are the morphological operators applied.

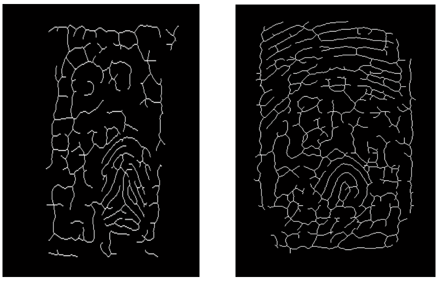


Fig6: Fingerprint Images after Morphological Operations

Thinning is a morphological operation that proficiently wears away the foreground pixels until they become one pixel wide, thus, the thickness of every line of pattern is minimized to a single pixel width [39] the process of removal of redundant pixels till the ridges become one pixel wide is facilitated by ridge thinning. The Ridge thinning algorithm utilized for Minutiae points' extraction in our scheme is employed by the authors of [40]. The image is divided into two dissimilar subfields that bear a likeness to a checkerboard pattern. In the initial sub iteration, only when all three conditions, G1, G2, and G3 are satisfied the pixel p from the initial subfield is erased. Whereas, in the second sub iteration, only when all three conditions, G1, G2, and G3' are satisfied, the pixel p from the foremost subfield is erased.

Condition G1:

$$X_H(P) = 1$$

Where

$$X_H(P) = \sum_{i=1}^{4} b_i$$

$$b_i = \left\{ \begin{matrix} 1 \text{ if } x_{2i-1} = 0 \text{ and } \left(x_{2i} = 1 \text{ or } x_{2i+1} = 1\right) \\ 0 \quad \text{otherwise} \end{matrix} \right\}$$

$x_1, x_2, ..., x_8$ are the values of the eight neighbors of p, starting with the east neighbor and numbered in counter-clockwise order.

Condition G2:

$$2 \le \min\{n_1(p), n_2(p)\} \le 3$$

where

$$n_1(p) = \sum_{k=1}^{4} x_{2k-1} \vee x_{2k}$$

$$n_2(p) = \sum_{k=1}^{4} x_{2k} \vee x_{2k+1}$$

Condition G3:

$$(x_2 \vee x_3 \vee \overline{x}_8) \wedge x_1 = 0$$

Condition G3':

$$(x_6 \vee x_7 \vee \overline{x}) \wedge x_5 = 0$$

One iteration of the thinning algorithm combines the two subiterations.

The fingerprint images with minuatie points marked are shown in Figure 7. The locations i.e) the coordinates of the extracted minutiae points are obtained and used in the subsequent processes.

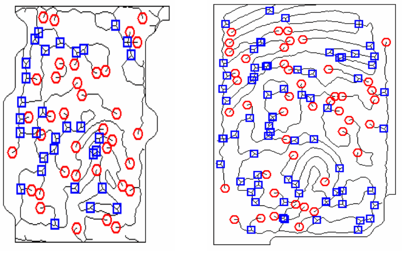


Fig 7: Fingerprint Images With Minutiae Points

IV. Generation of Cancelable Fingerprint Templates

The generation of cancelable fingerprint templates from the extracted minutiae points is explained in this section. The minutiae points extracted from the fingerprint image are represented as follows:

$$M_P = \{P_1, P_2, P_3, ..., P_n\}$$

and their corresponding x, y coordinates are specified separately as

$$M_{P_1}(x_1, y_1), M_{P_2}(x_2, y_2), M_{P_3}(x_3, y_3), ..., M_{P_n}(x_n, y_n)$$

With the aid of these x, y coordinates the distance between each point with respect to the other points is calculated.

$$\begin{Bmatrix} Mp_1 \\ Mp_2 \\ Mp_3 \\ \vdots \\ Mp_n \end{Bmatrix} = \begin{Bmatrix} (P_1, P_j) \\ (P_2, P_j) \\ (P_3, P_j) \\ \vdots \\ (P_n, P_j) \end{Bmatrix} \qquad j = 1 \text{ to } m,\ P_i \ne P_j$$

The distance between two points is computed using the following equation:

$$Distance(P_i, P_j) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Where (x_i, y_i), (x_j, y_j) are the co-ordinates of the points P_i and P_j respectively. Once the calculation of the respective distances for each point is done, the values are sorted in a separate array and the unique values are represented as the array elements. The array is denoted as:

$$\begin{Bmatrix} D_1 \\ D_2 \\ D_3 \\ \vdots \\ D_n \end{Bmatrix} = \begin{Bmatrix} (d_{11}, d_{12}, \cdots, d_{1m}) \\ (d_{21}, d_{22}, \cdots, d_{2m}) \\ (d_{31}, d_{32}, \cdots, d_{3m}) \\ \vdots \\ (d_{n1}, d_{n2}, \cdots, d_{nm}) \end{Bmatrix}$$

and the values obtained are denoted as

$$D = [D_1\ \ D_2\ \ D_3\ \cdots\ D_n]$$

The values thus obtained are sorted further and the formula for sorting them is given as

$$S_D = Sort(D)_{Asc}$$

Where as the unique values are represented as

$$U_D = \cup S_D = [u_{D_1}\ \ u_{D_2}\ \cdots\ u_{D_n}]$$

The UD thus formed is known as the cancelable fingerprint template. This cancelable template is employed in the generation of irrevocable cryptographic key.

V. Generation of Irrevocable Cryptographic Key

The irrevocable cryptographic key is generated from the cancelable fingerprint template formed with the aid of the approach discussed in this section. The cancelable fingerprint template U_D is divided into two equal parts of same size for shuffling purpose. The first part of the divided values are represented as,

$$U_{D_1} = [u_{D_1}\ \cdots\ u_{D_{n/2}}]$$

and the other half values are denoted as

$$U_{D_2} = [u_{D_{\frac{n}{2}+1}}\ \cdots\ u_{D_n}]$$

The elements of U_{D_1} and U_{D_2} are shuffled and stored in SU_{D_1} and SU_{D_2} respectively. The shuffling is performed as follows. An element of U_{D_1} is taken and the modulo operation is performed between the current element and $N/2$, where N represents the total number of elements in U_D. The resultant value is denoted as ind. Subsequently the current element of U_{D_1} is placed in SU_{D_2} at ind^{th} position. The aforesaid process is repeated for all the elements of U_{D_1} and U_{D_2}. Consequently the SU_{D_1} and SU_{D_2} are combined to form a vector SU_D.

$$SU_D = [SU_{D_1} \cup SU_{D_2}]$$

The shuffled vector SU_D is converted into a matrix MU_D of size $sqrt(|SU_D|) * sqrt(|SU_D|)$.

$$MU_D = (a_{ij}) sqrt(|SU_D|) * sqrt(|SU_D|)$$

Finally, the irrevocable key vector IK_V is generated from the matrix MU_D as follows:

$$IK_V = \{k_i : P(k)\}, \mathrm{i} = 1,....,|\mathrm{SU_D}|$$

Where $P(k) = |SM_{ij}| \bmod 2,$

$$\mathrm{SM_{ij}} = MU_D\ \mathrm{i, j : i + size, j + size, -1 < i, j < sqrt(|SU_D|)}$$

The final key thus generated is more secured and irrevocable. Obtaining cancelable fingerprint template from the generated key is impossible.

VI. Security Analysis

The security of the proposed scheme is strengthened by the following two robust features.

- Cancelable Transform
- Irreversible Analysis

A. Cancelable Transform

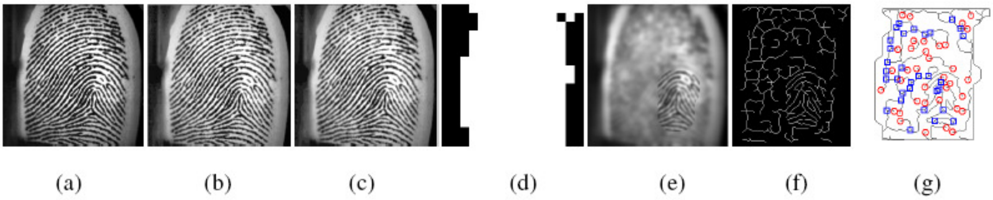
Cancelable transform [24] is employed to produce a cancelable template. The key target of the cancelable transformation is to proffer cancelable skill to a "non-invertible" transform. Generally, it is found to reduce the discriminative power of the original template. Thus, the cancelable templates and the secure templates of an individual in dissimilar applications will be diverse. However, the cross matching across databases will not be possible. Furthermore, it is possible to cancel and reissue the protected template by altering the cancelable transform parameters.

b. Irreversible Analysis

In order for the enunciation of the concept further, the tracing of the matrix with the aid of determinant or reorganizing shuffled data is entirely impracticable, analogous to attempting the generation of original document through hashed bits once hashing function is applied. The innate irrevocable nature reinforces the protection of our scheme. Hence, it is virtually unfeasible to trace the cancelable fingerprint template from the generated keys. The projected scheme is further appropriate and explicit for data such as the ones employed for managing minutiae points arrived at through the aforesaid procedure.

VII. Experimental Results

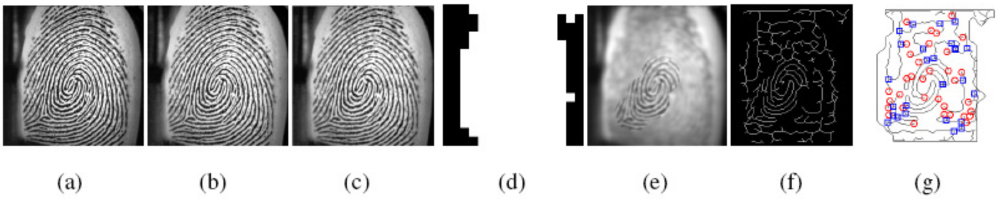
In this section we have presented the experimental analysis of our proposed scheme. Our scheme is programmed in Matlab (Matlab7.4). We have tested the proposed system with diverse fingerprint images from publicly available sources. The minutiae points are extracted from the fingerprint images using the approach discussed in the paper. Initially, the fingerprint image is segmented from the background to extract the region that actually contains the fingerprint. Further the orientation field is estimated and the fingerprint image is enhanced using average filtering and Gabor filtering. Subsequently, the minutiae points are extracted and their coordinates are obtained. The coordinates of the minutiae points are then employed in the generation of cancelable fingerprint template. Eventually, the irrevocable cryptographic key is generated from the cancelable fingerprint template. The input image, extracted minutiae points, the intermediate results and the generated irrevocable cryptographic key of two different fingerprint images are shown in Figure 8 and Figure 9 respectively.



```
10011111111111111100111111111111100011111111111110001111111111111000111111111111100011
11111111111110111111111111111111111111111111111111111111111111111111111111111111111111
111111011111111111111000111111111111111111111111111111111111111111111111111111111111
```

(h)

Fig. 8: (a) Input Fingerprint Image (b) Histogram Equalized Image (c) Median Filtered Image (d) Segmented Image (e) Enhanced Image (f) Morphological Processed Image (g) Fingerprint images with Minutiae Points (h) Generated irrevocable key



```
10101111111111111001111111111111111011111111111110111111111111111001111111111111110111
11111111111001111111111111100011111111111111111111111111111011111111111111100011111111
111111001111111111111011111111111111111111111111111111111111111111111111111111111111
```

(h)

Fig. 9: (a) Input Fingerprint Image (b) Histogram Equalized Image (c) Median Filtered Image (d) Segmented Image (e) Enhanced Image (f) Morphological Processed Image (g) Fingerprint images with Minutiae Points (h) Generated irrevocable key.

VIII. Conclusion

The steadily escalating reports on security infringements have necessitated the increase in concern for the security of information. Despite cryptography being a powerful tool attains information security, one of the chief demands in cryptosystems is to sustain the secrecy of the cryptographic keys. Combining biometrics with cryptography has provided an effective solution to this problem. Generation of cryptographic key from biometrics has gained enormous popularity in research community. Lately, the cancelable biometric systems have been widely recognized in the applications that are highly demanding in terms of privacy and security of biometric templates. We have presented an effective scheme for generating irrevocable cryptographic key from cancelable fingerprint templates in this paper. Initially the minutiae points are efficiently extracted from the fingerprints followed by the generation of cancelable templates and extraction of irrevocable keys from the cancelable templates in a successful manner. As the cryptographic key is generated in an irreversible manner, obtaining cancelable fingerprint templates and original fingerprints from the generated key is impossible. We have evaluated the effectiveness of our scheme using fingerprints from publicly available sources successfully. Moreover, we have as well presented the security analysis of the proposed scheme.

References

[1] John Chirillo and Scott Blaul, "Implementing Biometric Security," Wiley Red Books, ISBN: 978-0764525025, April 2003.

[2] Jain, A.K., Ross, A. and Prabhakar, S, "An introduction to biometric recognition," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 14, No. 1, pp: 4- 20, 2004.

[3] T.C. Clancy, N. Kiyavash and D.J. Lin, "Secure smart card-based fingerprint authentication," Proceedings of the 2003 ACM SIGMM Workshop on Biometrics Methods and Application, WBMA 2003.

[4] A. Goh, D.C.L. Ngo, "Computation of cryptographic keys from face biometrics," International Federation for Information Processing 2003, Springer-Verlag, LNCS 2828, pp. 1–13, 2003.

[5] Wildes, R.P., "Iris recognition: an emerging biometric technology," In Proceedings of the IEEE, Vol. 85, No. 9, pp:1348 - 1363, Sep 1997.

[6] Övünç Polat and Tülay Yıldırım, "Hand geometry identification without feature extraction by general regression neural network," Expert Systems with Applications, Vol. 34,No. 2, pp. 845-849, 2008.

[7] F. Monrose, M.K. Reiter, Q. Li and S. Wetzel, "Cryptographic key generation from voice," Proceedings of the 2001 IEEE Symposium on Security and Privacy, May 2001.

[8] F. Hao, C.W. Chan, "Private Key generation from on-line handwritten signatures," Information Management & Computer Security, Issue 10, No. 2, pp. 159–164, 2002.

[9] Chen, B. and Chandran, V., "Biometric Based Cryptographic Key Generation from Faces," 9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications, pp:394 - 401, 3-5 Dec,2007.

[10] G. I. Davida, Y. Frankel, and B. J. Matt. On enabling secure applications through off-line biometric identification. In Proceedings of the 1998 IEEE Symposium on Security and Privacy, pages 148–157, May 1998.

[11] A. Juels and M. Wattenberg. "A fuzzy commitment scheme". In Proceedings of the 6th ACM Conference on Computer and Communication Security, pages 28–36, November 1999.

[12] F. Monrose, M. K. Reiter, and S. Wetzel. "Password hardening based on keystroke dynamics". In Proceedings of the 6th ACM Conference on Computer and Communications Security, pages 73–82, November 1999.

[13] U. Uludag, S. Pankanti, P. S., and A. Jain, "Biometric cryptosystems: Issues and challenges," Proceedings of the IEEE 92, pp. 948–960, June 2004.

[14] M. Freire-Santosa, J. Fierrez-Aguilara, J. Ortega-Garciaa, "Cryptographic key generation using handwritten signature", In Proc. SPIE, volume 6202, pages 225–231, 2006.

[15] Nagar, A. and Chaudhury, S, "Biometrics based Asymmetric Cryptosystem Design Using Modified Fuzzy Vault Scheme," 18th International Conference on Pattern Recognition, Vol. 4, pp: 537-540, 2006.

[16] M. Savvides, B.V.K. Vijaya Kumar, and P.K. Khosla, "Cancelable biometric filters for face recognition", ICPR, pp. 922-925 Vol.3, 23-26 Aug. 2004.

[17] Nalini Ratha, Jonathan Connell, Ruud M. Bolle, Sharat Chikkerur, "Cancelable Biometrics: A Case Study in Fingerprints", Proceedings of the 18th International Conference on Pattern Recognition, vol:4, Pages: 370 - 373,2006.

[18] Russell Ang, Reihaneh Safavi-Naini, Luke McAven: "Cancelable Key-Based Fingerprint Templates." ACISP, pp: 242-252, 2005.

[19] Cheung King-Hong , Kong Adams ,Zhang David , Kamel Mohamed , You Jane , LAM Toby , LAM Ho-Wang , "An analysis on accuracy of cancelable biometrics based on biohashing" ,International Conference on Knowledge-Based Intelligent Information and Engineering Systems ,September 14-16, 2005.

[20] Ratha, N.K., Connell, J.H., Bolle, R.M.: "Enhancing security and privacy in biometrics- based authentication systems", IBM Systems Journal 40, pp: 614-634, 2001,.

[21] Andrew Beng Jin Teoh, Kar-Ann Toh and Wai Kuan Yip,"2^N Discretisation of BioPhasor in Cancellable Biometrics", Advances in Biometrics, Springer Berlin / Heidelberg, Vol. 4642, 2007.

[22] Connie Tee, Teoh Andrew, Goh Michael, Ngo David, "Palmhashing: a novel approach for cancelable biometrics", Information processing letters, vol. 93, no:1, pp. 1-5, 2005.

[23] F. Hao, R. Anderson, and J. Daugman, "Combining Crypto with Biometrics Effectively," IEEE Transactions on Computers, vol. 55, pp. 1081-1088, 2006.

[24] Y C Feng, Pong C Yuen, and Anil K Jain, “A Hybrid Approach for Face Template Protection,” in Proc. of SPIE Conference of Biometric Technology for Human Identification, Orlando, FL, USA, vol. 6944, 18 March 2008.

[25] Teoh AB, Yuang CT., "Cancelable biometrics realization with multispace random projections.", IEEE Trans Syst., vol:37, no:5, pp:1096-106, 2007.

[26] Je-Gyeong Jo, Jong-Won Seo and Hyung-Woo Lee, "Biometric Digital Signature Key Generation and Cryptography Communication Based on Fingerprint ", Lecture Notes in Computer Science, Springer, Vol: 4613, Pages 38-49, 2007.

[27] Julien Bringera, Hervé Chabannea, Bruno Kindarji, "The best of both worlds: Applying secure sketches to cancelable biometrics", Science of Computer Programming, Volume 74, Issues 1-2,Pages 43-51, 2008.

[28] Andrew B. J. Teoh, Yip Wai Kuan, Sangyoun Lee," Cancelable biometrics and annotations on BioHash ", Pattern Recognition, Vol: 41, Issue 6, pp: 2034-2044, 2008.

[29] Andrew Teoh Beng Jin, Tee Conniea, "Remarks on Bio-Hashing based cancelable biometrics in verification system", Neurocomputing, Vol: 69, no: 16-18, Pages 2461-2464, 2006.

[30] Beng, A., Jin Teoh, Kar-Ann Toh, "Secure biometric-key generation with biometric helper”, 3rd IEEE Conference on Industrial Electronics and Applications, pp: 2145-2150, 2008.

[31] Sanaul Hoque, Michael Fairhurst, Gareth Howells "Evaluating Biometric Encryption Key Generation Using Handwritten Signatures", Bio-inspired, Learning and Intelligent Systems for Security, pp: 17-22, 2008.

[32] S. Pankanti, S. Prabhakar, A.K. Jain, “On the individuality of fingerprints”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 8, pp.1010–1025, 2002.

[33] Jain, A.K.; Prabhakar, S.; Hong, L.; Pankanti, S., "Filterbank-based fingerprint matching", IEEE Transactions on Image Processing, vol. 9, no. 5, pp: 846-859, May 2000, Doi: 10.1109/83.841531.

[34] J. Patrick Fitch, Edward J Coyle and Neal Gallagher, “Median filtering by Threshold Decomposition”, IEEE Transactions on Acoustics, Speech and Signal Processing (ASSP), vol. 32, no.6, pp. 1183 - 1188,1984

[35] Yi Wang, Jiankun Hu and Fengling Han, "Enhanced gradient-based algorithm for the estimation of fingerprint orientation fields," Applied Mathematics and Computation, Special Issue on Intelligent Computing Theory and Methodology, Vol. 185, No. 2, pp. 823-833,15 February 2007.

[36] M.Tham, "Averaging Filter," University of Newcastle from http://lorien.ncl.ac.uk /ming/filter/filave.htm

[37] “Gabor Filter” from http://en.wikipedia.org /wiki/Gabor_filter.

[38] Lin Hong, Wan Yi-fei and A. Jain,”Fingerprint Image Enhancement: Algorithm and Performance Evaluation,” IEEE Transaction on Pattern Analysis and Matching Intelligence, vol. 20, no.8, pp: 777-789, 1998.

[39] Manvjeet Kaur, Mukhwinder Singh, Akshay Girdhar, and Parvinder S. Sandhu, " Fingerprint Verification System using Minutiae Extraction Technique", in proc. of World Academy of Science, Engineering and Technology, vol. 36, December 2008

[40]L. Lam, S. W. Lee, and C. Y. Suen, “Thinning Methodologies-A Comprehensive Survey”, IEEE Transactions on Pattern analysis and machine intelligence, vol. 14, no. 9, 1992.

A Novel Framework Based On Biometrics for Digital Rights Management

Nallamothu.Nagamalleswara Rao and Prof. Trimurthy

Abstract—**Digital Rights Management (DRM) is a technology that aims to stop, or at least ease, the practice of piracy. In this paper, we have developed a framework or protocol for protecting the copyrights of digital images. The approach makes use of the techniques: biometrics and digital image watermarking. The approach aims to embed the biometric feature of the owner as the data to solve the rightful of ownership when an ownership dispute comes. This data is embedded into the digital image using watermarking methods. By using the biometric feature, the information is highly secured and it could not be hacked by the hackers.**

Index Terms—**Biometrics, Digital Watermarking, copyright protection, Digital Rights Management, Fingerprint.**

I. Introduction

SECURING personal privacy and deterring identity theft are national priorities. These goals are essential to our democracy and our economy, and inherently important to our citizens.

Biometrics, an emerging set of technologies, promises an effective solution. Biometrics is the science and technology of interactively measuring and statistically analyzing biological data, in particular, taken from live people. In the area of computer security, biometrics refers to authentication techniques that rely on measurable physiological and behavioral characteristics that can be automatically verified. Biometrics is a rapidly evolving technology that is being widely used in forensics, such as criminal identification and prison security, and that has the potential to be used in a large range of civilian application areas. Although the field of biometrics is still in its infancy, it's inevitable that biometric systems will play a critical role in the future of security [1].

Reliable information security mechanisms are required to combat the rising magnitude of identity theft in our society [2]. The problems that may arise from the attacks on such systems are raising concerns as more and more biometric systems are deployed [6]. Some techniques such as cryptography and watermarking have been introduced to thwart some of these attacks. The idea of digital watermarking is to embed a small amount of secret information - the watermark into the host digital productions, such as image and audio, so that it can be extracted later for the purposes of copyright assertion, authentication, and content integrity verification, etc [15]. Watermarking techniques are gaining more interest by providing promising results [7, 8, 9]. For example, watermarking of fingerprint images can be used to secure central databases from which fingerprint images are transmitted on request to intelligence agencies in order to use them for identification purposes.

Privacy and other copyright violations regarding digital multimedia content represent a significant problem for legal content owners and content distributors. Hence, the protection of intellectual property rights for multimedia content, often referred to as the Digital Rights Management (DRM) for multimedia, recently started receiving a considerable amount of interest. Digital Rights Management solutions enable corporate, government and other organizations to protect confidential information and premium content from unauthorized use even by authorized users [12]. DRM allows the issuer of the media or file to control in detail what can and cannot be done with a single instance. For example, an issuer can limit the number of viewings, number of copies, which devices the media can be transferred to etc. DRM methods combine both encryption and digital watermarking to ensure better security against illegal copying and distribution [3].

Security of digital images has become a great importance with the omnipresence of internet. The advent of image processing tools has increased the vulnerability for illicit copying, modifications, and dispersion of digital images. Techniques like digital watermarking are put into practice to prevent unauthorized replication or exploitation of digital images [4], [13] and [14]. Digital watermarks of ownership embedded onto digital content offer copyright protection, ownership assertion, and integrity checks for digital content [4], and can provide evidence of copyright infringement after an attack. Digital watermarking technology is emerging fields in computer science, cryptography, signal processing, Image Processing and communications [10]. Watermark researchers always claim their watermarking techniques are robust and secure, but they fail to address the fundamental question of just how secure [5]. The idea of digital watermarking is to embed a small amount of secret information the watermark into the host digital productions, such as image and audio, so that it can be extracted later for the purposes of copyright assertion, authentication, and content integrity verification, etc [11].

Watermarking can be used for many security purposes such as copyright protection, fingerprinting, copy protection, data authentication, and so forth. The purpose of watermark is to protect the owner's copyright. The existing techniques by providing counterfeit watermarking schemes that can be performed on water marked image to allow multiple claims of ownerships. Since the copyright protection cannot be achieved in many situations, the rightful owner of a digital document may want the power to detect or prove its misappropriation. Thus one particularly promising copyright protection method is the use of digital watermarking techniques, which embed information identifying the copyright owner's identity within the content.

Digital watermarking ensures the copyright protection by embedding some information about ownership into the digital data. The embedded data can later be extracted from, or detected in, the digital data for different purposes such as copyright protection, access control, and broadcast monitoring. The information about the ownership may be password, logo or any privacy information which uniquely identifies the owner in ownership disputes. The above information could easily be hacked by the hackers and also may be lost or forgotten by the owner. The hackers may brute-force the same information and claim for the ownership.

In this paper, we have developed a novel and efficient framework which focuses on preventing disputes that comes out of ownership claims on digital documents. The proposed method implants the fingerprint feature into the digital content. The information is very much secured and confidential since we use the biometric feature to be embedded into the digital content. As a result the information could not be extracted by the user’s biometric feature because it is unique and it cannot be easily hacked by the hackers.

II. Related Works

Justin Picard et al. [20] have presented a virtually fraud-proof ID document based on a combination of three different data hiding technologies namely digital watermarking, 2-D bar codes, and Copy Detection Pattern, plus additional biometric protection. They have also shown that the combination of data hiding technologies protects the document against any forgery, in principle without any requirement for other security features.

Minerva M. Yeung et al. [21] focused on the study of watermarking on images used in automatic personal identification technology based fingerprints. They investigated the effects of watermarking fingerprint images on the recognition and retrieval accuracy using invisible fragile watermarking technique for image verification applications on a specific fingerprint recognition system.

Mohamed Mostafa Abd Allah [22] has proposed a technique for fingerprint identification using Artificial neural networks. They utilized clustering algorithm to detect similar feature groups from template images generated from the same finger and create the cluster core set. Their proposed feature extraction scheme was based on the reduction of information contents to the required minimum and also it define which part of the image is crucial and then it will be omitted.

Sooyeun Jung et al. [23] presented a user identification method at H.264 streaming using watermarking with fingerprints. The algorithm proposed by them consists of enhancement of a fingerprint image, watermark insertion using discrete wavelet transform and extraction after restoring. Their algorithm can achieve robust watermark extraction against H.264 compressed videos.

Umut Uludag et al., have presented a multimedia content protection scheme based on biometric data of the users and the layered encryption/decryption scheme. Password-only encryption schemes are vulnerable to illegal key exchange problems. By using biometric data along with hardware identifiers as keys, it is possible to alleviate fraudulent usage of protected content. A combination of symmetric and asymmetric key systems was utilized by them for this purpose [24].

Mina Deng, et al. [25] have proposed a model for privacy infrastructures aiming for the distribution channel such that as soon as the picture is publicly available, the exposed individual has a chance to find it and take proper action in the first place. Digital rights management techniques are applied in their proposed infrastructure, and data identification techniques such as digital watermarking and robust perceptual hashing was proposed to enhance the distributed content identification.

Tuan Hoang, et al. [26] have proposed a remote multimodal biometric authentication framework based on fragile watermarking for transferring multi-biometrics over networks to server for authentication. The proposed framework enhances security and reduces bandwidths. In order to reduce error rates from embedding numeric information, they also proposed a method to determine bit priority level in a bit sequence representing the numerical information to be embedded and combine with the current amplitude modulation watermarking method.

Emanuele Maiorana, et al. [27] have proposed two different approaches for the protection of on-line signature biometric templates. In the first one, cryptographic techniques are employed to protect signature features, making impossible to derive the original biometrics from the stored templates. In the second one, data hiding techniques are used to design a security scalable authentication system, embedding some dynamic signature features into a static representation of the signature itself.

III. Embedding The Fingerprint Feature Into The Digital Image

The proposed work makes use of the watermarking and biometrics technologies. Initially

We have fingerprint image and the digital content. From the fingerprint image, we extract the feature points and embed those fingerprint feature points into the digital image. In case of any claim of ownership on the digital content, we extract feature points from the digital image and compare with the claiming owner’s fingerprint feature and decision is performed based on the comparison.

A. Feature Point Extraction

This process extracts the feature points from the fingerprint image. The feature points are Core, Delta points, Bifurcation points and Ridges. The core and delta are also called the singularity points. Delta point is center of the triangular where three different directions flows meet. Ridge can be defined as a single curve segment. The combination of several ridges forms a fingerprint pattern. Ridge Ending and Bifurcation are taken as the distinctive features of fingerprint. The process of feature extraction consists of the following phases.

1.Histogram Equalization: This process turns the embedded finger print image brighter than before.

2.Image Enhancement: Image enhancement process consists the process of applying the fast Fourier transform to the image. This application of fast Fourier transform is done for each block of the image separately.

3.Binarization: This process binarizes the enhanced image into binary image. During this step, a particular threshold is set up, and the pixel values above this threshold are changed into 1 and pixel values below this threshold is marked as 0. Thus finally the binarized image contains only the values 0 and 1. Also the threshold we chose is adaptive threshold, in which the threshold value is automatically set depending upon the image.

4.Determination Region of Interest(ROI) : Determine the ROI of the binarized image. ROI is the regions in the image on which the process is to be done.

5.Applying Binary Morphological Operators: Apply Binary Morphological Operators on the above binarized image. These operators are applied mainly for the purpose of removing obstacles from the image. The morphological operators used are:

i. Thin Operator

While applying Thin operator, the following steps are carried out:

a. Divide the image into two distinct subfields in a checkerboard pattern.

b. In the first subiteration, delete pixel p from the first subfield if and only if the conditions G1, G2, G3 are all satisfied.

c. In the second subiteration, delete pixel p from the second subfield if and only if the conditions G1, G2, G3 are all satisfied.

Condition G1:

$$X_{H(P)} = 1$$

where

$$X_{H(P)} = \sum_{i=1}^{4} b_i$$

$$b_i \begin{cases} 1 \text{ if } \mathrm{x}_{2\mathrm{i}-1} = 0 \text{ and } (\mathrm{x}_{2\mathrm{i}} = 1 \text{ or } \mathrm{x}_{2\mathrm{i}+1} = 1) \\ 0 \ \ otherwise \end{cases}$$

$X_1, X_2, \ldots, X_8$ are the values of the eight neighbors of p, starting with the east neighbor and numbered in counter-clockwise order.

Condition G2:

$$2 \le \min[n_1(p), n_2(p)] \le 3$$

Where

$$n_1(p) = \sum_{k=1}^{4} x_{2k-1} \ \vee \ \mathrm{x}_{2\mathrm{k}}$$

$$n_2(p) = \sum_{k=1}^{4} x_{2k} \ \vee \ \mathrm{x}_{2\mathrm{k}+1}$$

Condition G3:

$$\left(x_2 \vee x_3 \vee \overline{x_8}\right) \wedge x_1 = 0$$

Condition $G3'$ **:**

$$\left(x_6 \vee x_7 \vee \overline{x_4}\right) \wedge x_5 = 0$$

ii.Clean Operator

The Clean operator removes isolated pixels (individual 1's that are surrounded by 0's), such as the centre pixel in this pattern:

$$\begin{array}{ccc} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{array}$$

iii. Hbreak

The Hbreak operator removes H-connected pixels. For example:

$$\begin{array}{ccc} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{array} \quad \text{becomes} \quad \begin{array}{ccc} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{array}$$

iv. Spur Operator

The Spur operator is used to remove spur pixels. For example:

$$\begin{array}{cccc} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{array} \quad \text{becomes} \quad \begin{array}{cccc} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}$$

The above said process occurs in feature extraction. If any dispute arises, the matching process is done between the feature points which are hided inside the digital image and the fingerprint feature points of the claiming owner.

b. *Watermark Embedding*

The initial phase in our proposed methodology is the watermark embedding phase. In this phase we have to perform certain operations on two categories of images namely cover image and hide image. In watermarking, cover image is the input image which we want to perform watermarking and hide image is an image which is to be embedded on the cover image. In this paper hide image represents the fingerprint feature .In the embedding phase we have to apply certain steps on both cover image and hide image separately.

The following are the steps which we have to perform on cover image:

- Initially we have to apply DCT transform to the cover image. DCT stands for "Discrete Cosine Transform". DCT transformation has its own matrix like any other transformations. In this step we have to perform one mathematical operation which is given below:

DCT Transform matrix * matrix of the cover image * Transpose of DCT

After the application of DCT we will obtain a result which is in matrix format and we can call it as a DCT Transformed matrix which as an input to the next step.

- Next step is to read the DCT transformed matrix in a zigzag fashion. The main reason for this is, after the application of DCT we didn't get the decomposed matrix unlike other transformations. So we are reading our DCT transformed matrix in a zigzag fashion which converts the transformed matrix into a decomposed form. In this step we get four quadrants as a result.
- Next step in this process is the application of SVD. SVD stands for "Singular Value Decomposition". It is a significant factorization of a rectangular real or complex matrix, with several applications in signal processing and statistics. In our work, we have to take the four quadrants which is an output of the previous step as an input and next we have to apply SVD to each block separately.

After the application of SVD, we get [U S V] for each quadrants. In [U S V], S is the diagonal matrix and U and V is unitary matrix. The general equation of SVD is

$$X=USV^T$$

The above specified three steps are the initial process in our work. Next we have performed certain steps on hide image which are as follows:

- Initially as in the cover image, we have to apply DCT to the hide image. After the application of DCT, we get DCT transformed matrix as an output.
- Next we have to read the DCT transformed matrix in zigzag fashion so that we can obtain the matrix in decomposed form. In this step we get four quadrants as output.
- Then we have to apply SVD to the four quadrants separately. After the application of SVD, we get [U S V] for each quadrants.

After obtaining the [U S V] values for four quadrants of both cover image and hide image we have to perform certain operations which were mentioned below:

Initially we have to consider the 'S' value of first quadrant of both cover image and hide image. Then we have to perform a simple mathematical operation which is given below.

$$S_{c(1,1)} + S_{h(1,1)} * 0.25 \quad -> \quad (1)$$

Where

$S_{c(1,1)}$ -> first element in S matrix of the cover image

$S_{h(1,1)}$ -> first element in S matrix of the hide image

But for second, third and fourth quadrants instead of 0.25 we have to multiply the values with 0.01. The reason is, mostly the first quadrant resembles the same as input image but other quadrants seem to be noisy so we have to use the smaller values for performing multiplication.

Next with the result of the first quadrant which is obtained using (1) we have to perform certain operations which were discussed below:

Here we have to consider the first term of U and V values of first quadrant of the cover image. Initially we have to take the result of first quadrant which we got by applying (1) and then we have to multiply it with U and transpose of V of the cover image. After the application of this operation to the whole matrix, i.e. first we find only for the first element in the matrix but the same procedure should be followed for each and every element in the matrix. After the application of this operation to the whole elements in the matrix, we obtain the results for four quadrants.

- Next we have to apply the inverse zigzag function to the four quadrants which we got by the above mentioned step.
- And then we have to perform inverse DCT.

At the end of this process, we will get the original image embedded with the hide image. This step is referred as watermark embedding process.

C. watermark Extraction

Watermark extraction is the second and last step in the first phase of or proposed methodology. Next to embedding we have to perform extraction which removes the embedded image from the original image. The following steps have to be carried out to remove the watermarked image from the original image:

Initially in this process we have to take the watermarked image and cover image as an input. Next first we have to take the cover image first and then we have to apply the following steps which are almost similar to the steps we have performed in the embedding phase.

- Initially we have to apply DCT to the cover image. After the application, we get DCT transformed matrix as an output.
- To decompose the DCT transformed matrix, we have to read the DCT transformed matrix in a zigzag fashion. At the end of this step we get four quadrants as a result.
- Next we have to apply the SVD to the four quadrants separately and as a result we got [U S V] values for the four quadrants.

Next we have to take the watermarked image and then we have to apply the steps given which is also same as the previously explained steps:

- Initially we have to apply DCT to the watermarked image. After the application, we get DCT transformed matrix as an output.
- To decompose the DCT transformed matrix, we have to read the DCT transformed matrix in a zigzag fashion. At the end of this step we get four quadrants as a result.
- Next we have to apply the SVD to the four quadrants separately and as a result we got [U S V] values for the four quadrants.

After the applications of these steps to both cover image and watermarked image the actual extraction phase starts which were explained below:

In this process, we have to consider the S values in first quadrant of both cover image and watermarked image. Then we have to take the first element of S and we have to do a simple mathematical operation which is given below:

$$S_{c(1,1)} - S_{w(1,1)} / 0.25 \qquad \text{->} (2)$$

Where,

$S_{c(1,1)}$ -> first element in S matrix of the cover image.

$S_{w(1,1)}$ -> first element in S matrix of the watermarked image.

As explained in embedding phase we have to multiply the other quadrants with 0.01. And also we have to apply the same mathematical operations for whole matrix. At the end we get four quadrants to that we have to perform certain steps which are as follows:

Here we have to consider the first term of U and V values of first quadrant of the cover image. Initially we have to take the result of first quadrant which we got by applying (2) and then we have to multiply it with U and transpose of V of the hide image. After the application of this operation to the whole matrix, i.e. first we find only for the first element in the matrix but the same procedure should be followed for each and every element in the matrix. After the application of this operation to the whole elements in the matrix, we obtain the results for four quadrants.

- ➢ Next we have to apply the inverse zigzag function to the four quadrants which we got by the above mentioned step.
- ➢ And then we have to perform inverse DCT.

At the end of this process, we will get the embedded image which is hided inside the original image. This step is referred as the extraction phase.

IV. CONCLUSION

Providing security for copyright protection is an emerging field today. In this paper, we have developed a framework for enhancing the security in the field copyright protection by combining both Biometrics and Watermarking Techniques. Biometrics and Watermarking are themselves powerful technologies for providing security when used individually. Since we are using these two secure technologies in copyright protection, they surely provide more security.

REFERENCES

[1] John Chirillo, Scott Blaul, "Implementing Biometric Security," John Wiley Publishers, 1st Edition, ISBN: 0764525026, 2003.

[2] Nandakumar, K.Jain, A.K.Pankanti, S.,"Fingerprint-based Fuzzy Vault: Implementation and Performance", IEEE Transactions on Information Forensics and Security, Vol: 2, No: 4, pp: 744-757, 2007.

[3] E. T. Lin, A. M. Eskicioglu, R. L. Lagendijk, and E. J. Delp., " Advances in digital video content protection.", IEEE: Special Issue on Advances in Video Coding and Delivery, pp:171–183, 2005.

[4] Memon, N. and Wong, P.W., "Protecting digital media content," Communications of the ACM, Vol: 41, pp.35–43, 1998.

[5]Sai Ho Kwok, "Watermark-based copyright protection system security", Communications of the ACM, Vol: 46, No: 10, pp: 98-101, 2003.

[6] Congress of the United States of America, "Enhanced border security and visa entry reform act", 2002.

[7] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Secure data hiding in wavelet compressed fingerprint images," In Proceedings of the ACM Multimedia Workshops , pp. 127–130, USA, 2000.

[8] A. K. Jain and U. Uludag, "Hiding biometric data," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 11, pp. 1494–1498, 2003.

[9] K. Zebbiche, L. Ghouti, F. Khelifi, and A. Bouridane, "Protecting fingerprint data using watermarking," In Proceedings of the 1st NASA/ESA Conference on Adaptive Hardware and Systems , pp. 451–456, 2006.

[10] Saraju, P. Mohanty. "Digital Watermarking: A Tutorial Review". Dept of Computer Science and Engineering, University of South Florida. 1999.

[11] Huayin Si, Chang-Tsun Li, "Copyright Protection in Virtual Communities through Digital Watermarking", Idea Group Publishing, 2005.

[12] Daniel Socek, Michal Sramka, Oge Marques , Dubravko Culibrk, "An Improvement to a Biometric-Based Multimedia Content Protection Scheme", In Proceedings of the 8th workshop on Multimedia and security , pp.135-139, 2006.

[13] G. Voyatzis and I. Pitas, "The use of watermarks in the protection of digital multimedia products," IEEE Proceedings, vol. 87, No. 7, pp 1197-1207, July 1999.

[14] A.B. Kahng, J. Lach, W.H. M-Smith, S. Mantik, I.L. Markov, M. Potkonjak, P. Tucker, H. Wang, and G. Wolfe, "Constraint-based watermarking techniques for design IP protection," IEEE Trans. Comput.-Aided Des. Integrated Circuits Syst., vol.20, no.10, pp.1236–1252, Oct. 2001.

[15] Huayin Si, Chang Tsun Li, "copyright protection in virtual communities through digital water marking", Idea Group Publishing; 2005.

[16] K. Zebbiche, F. Khelifi, "Region-Based Watermarking of Biometric Images: Case Study in Fingerprint Images", International Journal of Digital Multimedia Broadcasting, 2008.

[17] N. Memon, P. W. Wong, "Protecting digital media content," Communications of the ACM, 4, no. 7, pp. 11-24, July 1998.

[18] G. Voyatzis, I. Pitas, "The use of watermarks in the protection of digital multimedia products," IEEE Proceedings, vol. 87, No. 7, pp 1197-1207, July 1999.

[19] Holliman, M., Memon, N, "Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes", Image Processing, IEEE Transactions.

[20] Justin Picard, Claus Vielhauer and Niels Thorwirth,"Towards Fraud-Proof ID documents using multiple data hiding technologies and biometrics", SPIE Proceedings, vol.5306, pp: 416-427, 2004.

[21]Minerva M. Yeung and Sharath Pankanti,"Verification Watermarks on Fingerprint Recognition and Retrieval", SPIE Proceedings, vol. 3657, no. 66, 1999, Doi:10.1117/12.344704.

[22] Mohamed Mostafa Abd Allah, "Artificial Neural Networks Based Fingerprint Authentication with Clusters Algorithm", in proc.of Informatica vol.29, pp: 303–307, 2005.

[23]Sooyeun Jung, Dongeun Lee, Seongwon Lee, and Joonki Paik," Robust Watermarking for Compressed Video Using Fingerprints and Its Applications", in proc. of International Journal of Control, Automation and Systems, vol. 6, no. 6, pp. 794-799, December 2008.

[24] Umut Uludag and Anil K. Jain,"Multimedia Content Protection via Biometrics -Based Encryption", in Proceedings of the International Conference on Multimedia and Expo, vol.3, pp: 237 - 240, 2003, ISBN:0-7803-7965-9.

[25] Mina Deng, Lothar Fritsch, and Klaus Kursawe, "Personal Rights Management - Taming camera-phones for individual privacy enforcement", in proc. of 6th workshop on Privacy Enhancing Technologies, vol.4258, pp: 172-189, December 2006, Doi: 10.1007/11957454.

[26] Tuan Hoang, Dat Tran, and Dharmendra Sharma, "Remote Multimodal Biometric Authentication Using Bit Priority-Based Fragile Watermarking", in Proceedings of the 19th International Conference on Pattern Recognition (ICPR), 2008.

[27] Emanuele Maiorana, Patrizio Campisi, and Alessandro Neri,"Template Protection for On-line Signature-based Recognition Systems", in proc. of BIOD, pp: 170-180, 2008.

A Novel And Efficient Approach For Near Duplicate Page Detection In Web Crawling

V.A.Narayana[1], Dr P. Premchand[2] and Dr A Govardhan[3]

Abstract: **-The drastic development of the World Wide Web in the recent times has made the concept of Web Crawling receive remarkable significance. The voluminous amounts of web documents swarming the web have posed huge challenges to the web search engines making their results less relevant to the users. The presence of duplicate and near duplicate web documents in abundance has created additional overheads for the search engines critically affecting their performance and quality. The detection of duplicate and near duplicate web pages has long been recognized in web crawling research community. It is an important requirement for search engines to provide users with the relevant results for their queries in the first page without duplicate and redundant results. In this paper, we have presented a novel and efficient approach for the detection of near duplicate web pages in web crawling. Detection of near duplicate web pages is carried out ahead of storing the crawled web pages in to repositories. At first, the keywords are extracted from the crawled pages and the similarity score between two pages is calculated based on the extracted keywords. The documents having similarity scores greater than a threshold value are considered as near duplicates. The detection has resulted in reduced memory for repositories and improved search engine quality.**

Index Terms— **Web Mining, Web Content Mining, Web Crawling, Web pages, Stemming, Common words, Near duplicate pages, Near duplicate detection**

I. INTRODUCTION

THE employment of automated tools to locate the information resources of interest, and for tracking and analyzing the same, has become inevitable these days owing to the drastic development in the information accessible on the World Wide Web. This has made the development of server-side and client-side intelligent systems mandatory for efficient knowledge mining [1]. A branch of data mining that deals with the analysis of World Wide Web is known as Web Mining. Web Mining owes its origin to concepts from diverse areas such as Data Mining, Internet technology and World Wide Web, and lately, Semantic Web [2]. Web mining includes the sub areas: web content mining [3], web structure mining [4], and web usage mining [5] and can be defined as the procedure of determining hidden yet potentially beneficial knowledge from the data accessible in the web. The process of mining knowledge from the web pages besides other web objects is known as Web content mining. Web structure mining is the process of mining knowledge about the link structure linking web pages and some other web objects. The mining of usage patterns created by the users accessing the web pages is called Web usage mining [6].

The World Wide Web owes its development to the Search engine technology. The chief gateways for access of information in the web are Search engines. Businesses have turned beneficial and productive with the ability to locate contents of particular interest amidst a huge heap [31]. Web crawling, a process that populates an indexed repository of web pages is utilized by the search engines in order to respond to the queries [20]. The programs that navigate the web graph and retrieve pages to construct a confined repository of the segment of the web that they visit. Earlier, these programs were known by diverse names such as wanderers, robots, spiders, fish, and worms, words in accordance with the web imagery [7].

Generic and Focused crawling are the two main types of crawling. Generic crawlers [9] differ from focused crawlers [10] in a way that the former crawl documents and links of diverse topics whereas the latter limits the number of pages with the aid of some prior obtained specialized knowledge. Repositories of web pages are built by the web crawlers so as to present input for systems that index, mine, and otherwise analyze pages (for instance, the search engines) [8]. The subsistence of near duplicate data is an issue that accompanies the drastic development of the Internet and the growing need to incorporate heterogeneous data [21]. Even though the near duplicate data are not bit wise identical they bear a striking similarity [21]. Web search engines face huge problems due to the duplicate and near duplicate web pages. These pages either increase the index storage space or slow down or increase the serving costs thereby irritating the users. Thus the algorithms for detecting such pages are inevitable [22]. Web crawling issues such as freshness and efficient resource usage have been addressed previously [11], [12], [13]. Lately, the elimination of duplicate and near duplicate web documents has become a vital issue and has attracted significant research [15].

Identification of the near duplicates can be advantageous to many applications. Focused crawling, enhanced quality and diversity of the query results and identification on spams can be facilitated by determining the near duplicate web pages [19, 26, 22]. Numerous web mining applications depend on the accurate and proficient identification of near duplicates. Document clustering [17], detection of replicated web collections [18], detecting plagiarism [29], community mining in a social network site [30], collaborative filtering [16] and discovering large dense graphs [27] are a notable few among those applications. Reduction in storage costs and enhancement in quality of search indexes besides considerable bandwidth conservation can be achieved by eliminating the near duplicate pages [9]. Check summing techniques can determine the documents that are precise duplicates (because of mirroring or plagiarism) of each other [14]. The recognition of near duplicates is a tedious problem.

Research on duplicate detection was initially done on databases, digital libraries, and electronic publishing. Lately duplicate detection has been extensively studied for the sake of numerous web search tasks such as web crawling, document ranking, and document archiving. A huge number of duplicate detection techniques ranging from manually coded rules to cutting edge machine learning techniques have been put forth [21, 22, 23, 34 - 37]. Recently few authors have projected near duplicate detection techniques [38, 39, 40, 25]. A variety of issues such as from providing high detection rates to minimizing the computational and storage resources have been addressed by them. These techniques vary in their accuracy as well. Some of these techniques are computationally pricey to be implemented completely on huge collections. Even though some of these algorithms prove to be efficient they are fragile and so are susceptible to minute changes of the text.

The primary intent of our research is to develop a novel and efficient approach for detection of near duplicates in web documents. Initially the crawled web pages are preprocessed using document parsing which removes the HTML tags and java scripts present in the web documents. This is followed by the removal of common words or stop words from the crawled pages. Then the stemming algorithm is applied to filter the affixes (prefixes and the suffixes) of the crawled documents in order to get the keywords. Finally, the similarity score between two documents is calculated on basis of the extracted keywords. The documents with similarity scores greater than a predefined threshold value are considered as near duplicates. We have conducted an extensive experimental study using several real datasets, and have demonstrated that the proposed algorithms outperform previous ones.

The rest of the paper is organized as follows. Section 2 presents a brief review of some approaches available in the literature for duplicates and near duplicates detection. In Section 3, the novel approach for the detection of near duplicate documents is presented. The conclusions are summed up in Section 4.

II. Related Work

Our work has been inspired by a number of previous works on duplicate and near duplicate document and web page detection.

Sergey Brin et al. [34] have proposed a system for registering documents and then detecting copies, either complete copies or partial copies. They described algorithms for detection, and metrics required for evaluating detection mechanisms covering accuracy, efficiency and security. They also described a prototype implementation of the service, COPS, and presented experimental results that suggest the service can indeed detect violations of interest.

Andrei Z. Broder et al. [35] have developed an efficient way to determine the syntactic similarity of files and have applied it to every document on the World Wide Web. Using their mechanism, they have built a clustering of all the documents that are syntactically similar. Possible applications include a "Lost and Found" service, filtering the results of Web searches, updating widely distributed web-pages, and identifying violations of intellectual property rights.

Jack G. Conrad et al. [36] have determined the extent and the types of duplication existing in large textual collections. Their research is divided into three parts. Initially they started with a study of the distribution of duplicate types in two broad-ranging news collections consisting of approximately 50 million documents. Then they examined the utility of document signatures in addressing identical or nearly identical duplicate documents and their sensitivity to collection updates. Finally, they have investigated a flexible method of characterizing and comparing documents in order to permit the identification of non-identical duplicates. Their method has produced promising results following an extensive evaluation using a production-based test collection created by domain experts.

Donald Metzler et al. [37] have explored mechanisms for measuring the intermediate kinds of similarity, focusing on the task of identifying where a particular piece of information originated. They proposed a range of approaches to reuse detection at the sentence level, and a range of approaches for combining sentence-level evidence into document-level evidence. They considered both sentence-to-sentence and document-to-document comparison, and have incorporated the algorithms into RECAP, a prototype information flow analysis tool.

Hui Yang et al. [38] have explored the use of simple text clustering and retrieval algorithms for identifying near-duplicate public comments. They have focused on automating the process of near-duplicate detection, especially form letter detection. They gave a clear near-duplicate definition and explored simple and efficient methods of using feature-based document retrieval and similarity-based clustering to discover near-duplicates. The methods were evaluated in experiments with a subset of a large public comment database collected for EPA rule.

Monika Henzinger [22] has compared the two algorithms namely shingling algorithm [35] and random projection based approach [14] on a very large scale set of 1.6B distinct web pages. The results showed that neither of the algorithms works well for finding near-duplicate pairs on the same site, while both achieve high precision for near-duplicate pairs on different sites. She has presented a combined algorithm which achieves precision 0.79 with 79% of the recall of the other algorithms.

Hui Yang et al. [39] have presented DURIAN (DUplicate Removal In lArge collectioN), a refinement of a prior near-duplicate detection algorithm. DURIAN uses a traditional bag-of-words document representation, document attributes ("metadata"), and document content structure to identify form letters and their edited copies in public comment collections. The results have demonstrated that statistical similarity measures and instance-level constrained clustering can be quite effective for efficiently identifying near-duplicates.

In the course of developing a near-duplicate detection system for a multi-billion page repository, Gurmeet Singh Manku et al. [25] have made two research contributions. First, they demonstrated that Charikar's [14] fingerprinting technique is appropriate for this goal. Second, they presented an algorithmic technique for identifying existing f-bit fingerprints that differ from a given fingerprint in at most k bit-positions, for small k. This technique is useful for both online queries (single fingerprints) and batch queries (multiple fingerprints).

Ziv BarYossef et al. [40] have considered the problem of DUST: Different URLs with Similar Text. They proposed a novel algorithm, DustBuster, for uncovering dust; that is, for discovering rules that transform a given URL to others that are likely to have similar content. DustBuster mines dust effectively from previous crawl logs or web server logs, without examining page contents.

Chuan Xiao et al. [21] have proposed exact similarity join algorithms with application to near duplicate detection and a positional filtering principle, which exploits the ordering of tokens in a record and leads to upper bound estimates of similarity scores. They demonstrated the superior performance of their algorithms to the existing prefix filtering-based algorithms on several real datasets under a wide range of parameter settings.

III. Novel Approach For Near Duplicate Webpage Detection

A novel approach for the detection of near duplicate web pages is presented in this section. In web crawling, the crawled web pages are stored in a repository for further process such as search engine formation, page validation, structural analysis and visualization, update notification, mirroring and personal web assistants or agents and more. Duplicate and near duplicate web page detection is an important step in web crawling. In order to facilitate search engines to provide search results free of redundancy to users and to provide distinct and useful results on the first page, duplicate and near duplicate detection is essential. Numerous challenges are encountered by the systems that aid in the detection of near duplicate pages. First is the concern of scale since the search engines index hundreds of millions of web-pages thereby amounting to a multi-terabyte database. Next is the issue of making the crawl engine crawl billions of web pages everyday. Thus marking a page as a near duplicate should be done at a quicker pace. Furthermore, the system should utilize minimal number of machines [25].

The near duplicate detection is performed on the keywords extracted from the web documents. First, the crawled web documents are parsed to extract the distinct keywords. Parsing includes removal of HTML tags, java scripts, stop words/common words and stemming of remaining words. The extracted keywords and their counts are stored in a table to ease the process of near duplicates detection. The keywords are stored in the table in a way that the search space is reduced for the detection. The similarity score of the current web document against a document in the repository is calculated from the keywords of the pages. The documents with similarity score greater than a predefined threshold are considered as near duplicates.

A. Near Duplicate Web Documents

Even though the near duplicate documents are not bitwise identical they bear striking similarities. The near duplicates are not considered as “exact duplicates” but are files with minute differences. Typographical errors, versioned, mirrored, or plagiarized documents, multiple representations of the same physical object, spam emails generated from the same template, and many such phenomenons may result in near duplicate data. A considerable percentage of web pages have been identified as be near-duplicates according to various studies [17, 26 and 22]. These studies propose that near duplicates constitute almost 1.7% to 7% of the web pages traversed by crawlers. The steps involved in our approach are presented in the following subsections.

B. Web Crawling

The analysis of the structure and informatics of the web is facilitated by a data collection technique known as Web Crawling. The collection of as many beneficiary web pages as possible along their interconnection links in a speedy yet proficient manner is the prime intent of crawling. Automatic traversal of web sites, downloading documents and tracing links to other pages are some of the features of a web crawler program. Numerous search engines utilize web crawlers for gathering web pages of interest besides indexing them. Web crawling becomes a tedious process due to the subsequent features of the web, the large volume and the huge rate of change due to voluminous number of pages being added or removed each day.

Seed URLs are a set of URLs that a crawler begins working with. These URLs are queued. A URL is obtained in some order from the queue by a crawler. Then the crawler downloads the page. This is followed by the extracting the URLs from the downloaded page and enqueuing them. The process continues unless the crawler settles to stop [28]. A crawling loop consists of obtaining a URL from the queue, downloading the corresponding file with the aid of HTTP, traversing the page for new URLs and including the unvisited URLs to the queue [7].

C. Web Document Parsing

Information extracted from the crawled documents aid in determining the future path of a crawler. Parsing may either be as simple as hyperlink/URL extraction or complex ones such as analysis of HTML tags by cleaning the HTML content [7]. It is inevitable for a parser that has been designed to traverse the entire web to encounter numerous errors. The parser tends to obtain information from a web page by not considering a few common words like a, an, the and more, HTML tags, Java Scripting and a range of other bad characters [24]

1) Stop Words Removal: It is necessary and beneficial to remove the commonly utilized stop words such as "it", "can" ,"an", "and", "by", "for", "from", "of", "the", "to", "with" and more either while parsing a document to obtain information about the content or while scoring fresh URLs that the page recommends. This procedure is termed as stop listing [7]. Stop listing aids in the reduction of size of the indexing file besides enhancing efficiency and value.

D. Stemming Algorithm

Variant word forms in Information Retrieval are restricted to a common root by Stemming. The postulation lying behind is that, two words posses the same root represent identical concepts. Thus terms possessing to identical meaning yet appear morphologically dissimilar are identified in an IR system by matching query and document terms with the aid of Stemming [33]. Stemming facilitates the reduction of all words possessing an identical root to a single one. This is achieved by removing each word of its derivational and inflectional suffixes [32]. For instance, "connect," "connected" and "connection" are all condensed to "connect".

E. Keywords Representation

We posses the distinct keywords and their counts in each of the each crawled web page as a result of stemming. These keywords are then represented in a form to ease the process of near duplicates detection. This representation will reduce the search space for the near duplicate detection. Initially the keywords and their number of occurrences in a web page have been sorted in descending order based on their counts. Afterwards, n numbers of keywords with highest counts are stored in a table and the remaining keywords are indexed and stored in another table. In our approach the value of n is set to be 4. The similarity score between two documents can be calculated if and only the prime keywords of the two documents are similar. Thus the search space is reduced for near duplicates detection.

F. Similarity Score Calculation

If the prime keywords of the new web page do not match with the prime keywords of the pages in the table, then the new web page is added in to the repository. If all the keywords of both pages are same then the new page is considered as duplicate and thus is not included in the repository. If the prime keywords of new page are same with a page in the repository, then the similarity score between the two documents is calculated. The similarity score of two web documents is calculated as follows:

Let T1 and T2 be the tables containing the extracted keywords and their corresponding counts.

T1	K1	K2	K4	K5	…..	Kn
	C1	C2	C4	C5	…..	Cn

T2	K1	K3	K2	K4	…..	Kn
	C1	C3	C2	C4	…..	Cn

The keywords in the tables are considered individually for the similarity score calculation. If a keyword is present in both the tables, the formula used to calculate the similarity score of the keyword is as follows.

$$a = \Delta[K_i]_{T_1}$$

$$b = \Delta[K_i]_{T_2}$$

$$S_{D_C} = \log(count(a) / count(b)) * Abs(1 + (a - b))$$

Here ‘a’ and ‘b’ represent the index of a keyword in the two tables respectively.

If the keywords of T1 \ T2 ≠ φ, we use the following formula to calculate the similarity score. The amount of the keywords present in T1 but not in T2 is taken as N_{T_1}

$$S_{D_{T_1}} = \log(count(a)) * (1 + |T_2|)$$

If the keywords of T2 \ T1 ≠ φ, we use the below mentioned formula to calculate the similarity score. The occurrences of the keywords present in T2 but not in T1 is taken as N_{T2}

$$S_{D_{T_2}} = \log(count(b)) * (1 + |T_1|)$$

The similarity score (SSM) of a page against another page is calculated by using the following equation.

$$SS_M = \frac{\sum_{i=1}^{|N_C|} S_{D_C} + \sum_{i=1}^{|N_{T_1}|} S_{D_{T_1}} + \sum_{i=1}^{|N_{T_2}|} S_{D_{T_2}}}{N}$$

Where $n = |T_1 \cup T_2|$ and $N = (|T_1| + |T_2|)/2$. The web documents with similarity score greater than a predefined threshold are near duplicates of documents already present in repository. These near duplicates are not added in to the repository for further process such as search engine indexing.

IV. CONCLUSION

Though the web is a huge information store, various features such as the presence of huge volume of unstructured or semi-structured data; their dynamic nature; existence of duplicate and near duplicate documents and the like pose serious difficulties for Information Retrieval. The voluminous amounts of web documents swarming the web have posed a huge challenge to the web search engines making them render results of less relevance to the users. The detection of duplicate and near duplicate web documents has gained more attention in recent years amidst the web mining researchers. In this paper, we have presented a novel and efficient approach for detection of near duplicate web documents in web crawling. The proposed approach has detected the duplicate and near duplicate web pages efficiently based on the keywords extracted from the web pages. Further more, reduced memory spaces for web repositories and improved search engine quality have been accomplished through the proposed duplicates detection approach.

REFERENCES

[1] Cooley, R., Mobasher, B., Srivastava, J., "Web mining: information and pattern discovery on the World Wide Web", Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence, pp: 558 - 567, 3-8 Nov 1997, DOI: 10.1109/TAI.1997.632303.

[2] Bettina Berendt, Andreas Hotho, Dunja Mladenić, Myra Spiliopoulou and Gerd Stumme, "A Roadmap for Web Mining: From Web to Semantic Web", Lecture Notes in Artificial Intelligence , Vol. 3209, Springer-Verlags Heidelberg , Berlin; Heidelberg; New York, pp. 1-22, 2004. ISBN: 3-540-23258-3.

[3] M. Pazzani, L. Nguyen, and S. Mantik, “Learning from hotlists and coldlists: Towards a www information filtering and seeking agent”, In IEEE 1995 International Conference on Tools with Artificial Intelligence, 1995.

[4] David Gibson, Jon Kleinberg, and Prabhakar Raghavan, “Inferring web communities from link topology”, In Conference on Hypertext and Hypermedia, ACM, 1998.

[5] Kosala, R., and Blockeel, H., “Web Mining Research: A Survey,” SIGKDD Explorations, vol. 2, Issue. 1, June 2000.

[6] Ee-Peng Lim and Aixin Sun , “Web Mining - The Ontology Approach “, 2005.

[7] Pant, G., Srinivasan, P., Menczer, F., "Crawling the Web". Web Dynamics: Adapting to Change in Content, Size, Topology and Use, edited by M. Levene and A. Poulovassilis, Springer- verlog, pp: 153-178, November 2004.

[8] S. Balamurugan, Newlin Rajkumar, and J.Preethi, "Design and Implementation of a New Model Web Crawler with Enhanced Reliability", Proceedings of world academy of science, engineering and technology, volume 32, August 2008, ISSN 2070-3740.

[9] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan, "Searching the web", ACM Transactions on Internet Technology, vol. 1, no. 1: pp. 2-43, 2001.

[10] F. Menczer, G. Pant, P. Srinivasan, and M. E. Ruiz, "Evaluating topic-driven web crawlers", In Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 241-249, 2001.

[11] A. Z. Broder, M. Najork, and J. L. Wiener, "Efficient URL caching for World Wide Web crawling", In International conference on World Wide Web, 2003.

[12] S. Chakrabarti, "Mining the Web: Discovering Knowledge from Hypertext Data", Morgan-Kauman, 2002.

[13] J. Cho, H. Garcia-Molina, and L. Page, "Efficient crawling through URL ordering", Computer Networks and ISDN Systems, vol. 30, no. 1-7: pp. 161-172, 1998.

[14] M. Charikar. "Similarity estimation techniques from rounding algorithms". In Proc. 34th Annual Symposium on Theory of Computing (STOC 2002), pp: 380-388, 2002.

[15] H. Garcia-Molina, J. D. Ullman, and J. Widom, "Database System Implementation", Prentice Hall, 2000.

[16] R. J. Bayardo, Y. Ma, and R. Srikant. "Scaling up all pairs similarity search". In WWW, 2007.

[17] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. "Syntactic clustering of the web". Computer Networks, 29(8-13):1157–1166, 1997.

[18] J. Cho, N. Shivakumar, and H. Garcia-Molina. "Finding replicated web collections". In SIGMOD, 2000.

[19] J. G. Conrad, X. S. Guo, and C. P. Schriber. "Online duplicate document detection: signature reliability in a dynamic retrieval environment". In CIKM, 2003.

[20] Pandey, S.; Olston, C., "User-centric Web crawling", Proceedings of the 14th international conference on World Wide Web, pp: 401 – 411, 2005.

[21] Chuan Xiao, Wei Wang, Xuemin Lin, Jeffrey Xu Yu , "Efficient Similarity Joins for Near Duplicate Detection", Proceeding of the 17th international conference on World Wide Web, pp:131--140, 2008.

[22] M. Henzinger, "Finding near-duplicate web pages: a large-scale evaluation of algorithms," Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 284-291, 2006.

[23] F. Deng and D. Rafiei, "Approximately detecting duplicates for streaming data using stable bloom filters," Proceedings of the 2006 ACM SIGMOD international conference on Management of data, pp. 25-36, 2006.

[24] Castillo, C., "Effective web crawling", SIGIR Forum, ACM Press, Volume 39, Number 1, N, pp.55-56, 2005.

[25] Gurmeet Singh Manku, Arvind Jain, Anish Das Sarma, "Detecting near-duplicates for web crawling," Proceedings of the 16th international conference on World Wide Web, pp: 141 - 150, 2007

[26] D. Fetterly, M. Manasse, and M. Najork. "On the evolution of clusters of near-duplicate web pages". In LA-WEB, 2003.

[27] D. Gibson, R. Kumar, and A. Tomkins. "Discovering large dense subgraphs in massive graphs", In VLDB, 2005.

[28] Muhammad Shoaib, Shazia Arshad, "Design & Implementation of web information gathering system," NITS e-Proceedings, Internet Technology and Applications, King Saud University.

[29] T. C. Hoad and J. Zobel. "Methods for identifying versioned and plagiarized documents". JASIST, 54(3):203–215, 2003.

[30] E. Spertus, M. Sahami, and O. Buyukkokten. "Evaluating similarity measures: a large-scale study in the orkut social network". In KDD, 2005.

[31] Aameek Singh, Mudhakar Srivatsa, Ling Liu, Todd Miller, " Apoidea: A Decentralized Peer-to-Peer Architecture for Crawling the World Wide Web ", Proceedings of the SIGIR 2003 Workshop on Distributed Information Retrieval, Lecture Notes in Computer Science, Volume 2924, 2003.

[32] Lovins, J.B. 1968: "Development of a stemming algorithm". Mechanical Translation and Computational Linguistics, 11, 22-31(1968).

[33] M. Bacchin, N. Ferro, Melucci M.,"Experiments to evaluate a statistical stemming algorithm", Proceedings of the intl' Cross-Language Evaluation Forum Workshop, 2002.

[34] S. Brin, J. Davis, and H. Garcia-Molina. "Copy detection mechanisms for digital documents". In Proceedings of the Special Interest Group on Management of Data (SIGMOD 1995), pages 398–409. ACM Press, May 1995.

[35] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. "Syntactic clustering of the web". In Proceedings of WWW6 '97, pages 391–404. Elsevier Science, April 1997.

[36] J. Conrad and C. P. Schriber. "Online duplicate document detection: signature reliability in a dynamic retrieval environment." Proceedings of the twelfth international conference on Information and knowledge management, Pages: 443 - 452 New Orleans, LA, USA, 2003.

[37] D. Metzler, Y. Bernstein and W. Bruce Croft. "Similarity Measures for Tracking Information Flow", Proceedings of the fourteenth international conference on Information and knowledge management, CIKM'05, October 31.November 5, 2005, Bremen, Germany

[38] H. Yang and J. Callan, "Near-duplicate detection for eRulemaking," Proceedings of the 2005 national conference on Digital government research, pp: 78 - 86, 2005.

[39] Hui Yang, Jamie Callan, Stuart Shulman, "Next steps in near-duplicate detection for eRulemaking," Proceedings of the 2006 international conference on Digital government research, Vol. 151, pp: 239 - 248, 2006.

[40] Ziv Bar-Yossef, Idit Keidar, Uri Schonfeld, "Do not crawl in the dust: different urls with similar text," Proceedings of the 16th international conference on World Wide Web, pp: 111 - 120, 2007.

A Novel Approach for Content-Based Image Indexing and Retrieval

Suresh Pabboju

Abstract—**Statistical image analysis is simple to implement, but rarely achieves high levels of significance in image similarity searches except on contrived image collections. Semantic object recognition is, at this time, expensive to implement, especially for the wide variety of objects that humans are interested in, but would probably give accurate and relevant results in an image retrieval system. One of the most important factors in image database systems is indexing, since it affects the speed of data access and also supports the accuracy for retrieval process. The research regarding image indexing has been around for several years. Some of those researched focus from improving tree structures until selecting suitable image's features to be used as an index. The aim of the research is to find out which image's features are suitable for indexing from previous researches with my current data set. We propose a novel approach for content based image indexing and retrieval system based on the global and region features of the images. Tree data structures are planned to be used for the storage of these extracted features. The result from the experiments showed that using features on regional basis will increase to speed up the image retrieval process, effective usage of storage space, internal memory without compromising the accuracy of the retrieval.**

Index Terms—**Content Based Image Retrieval, Indexing, Retrieval, Global Features, Region Features, Edge Density, Color Average, R*-Tree Structure, Euclidean distance**.

I. Introduction

The rapid progress of multimedia computing and applications has brought about an explosive growth of digital images in computer systems and networks. This development has remarkably increased the need for image retrieval systems that are able to effectively index a large amount of images and to efficiently retrieve them based on their visual contents [2]. Content-based image retrieval (CBIR) is a promising technology to enrich the core functionality of picture archiving and communication systems (PACS). CBIR has a potential for making a strong impact in diagnostics, research, and education [9].Content-based image retrieval is a technique which uses visual contents to search images from large scale image databases according to users' interests, has been an active and fast advancing research area since the 1990s. During the past decade, remarkable progress has been made in both theoretical research and system development. However, there remain many challenging research problems that continue to attract researchers from multiple disciplines [8].

Content based image retrieval (CBIR) also known as Query by image content (QBIC) or Content based visual information retrieval (CBVIR), is the application of computer vision to the image retrieval problem. In CBIR, "Content-based" means that the search will analyze the actual contents of the image and the term 'content' refers to colors, shapes, textures, or any other information that can be derived from the image itself [1]. Content- Based image retrieval is the problem of searching large image repositories according to their content [3] [4] [5]. Image content may include both visual and semantic content. Visual content can be very general or domain specific [6]. General visual content include color, texture, shape, spatial relationship, etc [7]. Domain specific visual content, like human faces, is application dependent and may involve domain knowledge. Semantic content is obtained either by textual annotation or by complex inference procedures based on visual content [8].

Even before 1990, CBIR has been in existence, during which very little papers has been published. There has been a considerable increase in the number of papers published since 1997 [3], owing to the increase in interest of people towards this area of research. Those researches performed has resulted in various CBIR algorithms [10], [11], [12], [13]. The majority of those algorithms process image into several layers of tasks. Those layers of tasks consist of extracting the multidimensional features of an image query and compare it with images in the database are perform after the system populate database with images. The performance of retrieval [3] will be affected by the database filled with extracted information from the images which are also indexed suitably. The color, shape, texture and the rest of image's characteristic is contained by the information. Most content-based image retrieval systems do not look at regions in an image. This will cause some problems because many images include some objects and when a user specifies an image for retrieval, the user often pay attention solely an object or a region in the image [20].

The two main approaches to image retrieval are low-level image retrieval commonly known as content based image retrieval (CBIR), and high-level image retrieval based on text retrieval of images using image annotations [17]. Low-level image retrieval is based on the use of an integrated feature extraction/ object-recognition subsystem that automates the process of feature-extraction and object-recognition. This process is based on analysis of the low-level image features color, texture, location and/or shape to index images and later retrieve images based on similarity [16].Some of the techniques used in CBIR are Query by example, Semantic retrieval, content comparison techniques etc. Some of the existing works on CBIR like Query-by example [10], [11], [12], [13], localized CBIR [19] , image segmentation [18].

Image indexing and retrieval has lately drawn the attention of many researchers in the computer science community [24, 23, 22, 21, 25].With large collections of images and video frames, scalability, speed and efficiency are the essence for the success of an image retrieval system. There are two main families of image indexing and retrieval systems: those based on the content of the images (content-based) like color, texture, shape, objects, etc., and those based on the description of the images (description based) like keywords, size, caption, etc. While description based image retrieval systems are relatively easier to implement and to design user interface for, they suffer from the same problems as the information retrieval systems in text databases or Web search engines. It has been demonstrated that search engines using current methods based on simple keyword matching perform poorly. The precision of these search engines is very low and their recall is inadequate. Content-based image retrieval systems [23, 22, 25] use visual features to index images. These systems differ mainly in the way they extract the visual features and index images, and the way they are queried. They give a relatively satisfactory result with regard to the visual clues; however, their precision and recall are still not optimized. Moreover, they lack the power of locating specific objects and identifying their details (size, position, orientation, etc.).

The indexing of images to a storage system and the retrieval of those images for a given query is the prime focus of the existing work on CBIR. In order to produce accurate result for a given image query many researchers have been done on how to retrieve images from the database and for improving the retrieval process later on, many researchers have been done on how the images are being indexed. In this research we planned to focus on the second problem which is developing indexing method that can reduce the time to perform image retrieval and at the same time trying to cut down the usage of storage space.

The aim of this research is to representation of image features and image indexing. The project involve development of an image indexing algorithm based on the study of usefulness of various computational features in describing the visual contents of an image and the study of combination of features leading to successful retrieval results. The approach aims to extract the global features and region features initially from the images. Tree data structures are planned to be used for the storage of these extracted features. The global features we planned to extract are color sigma, edge density, color average and more. The region features planned to be used are region area, moment invariant, and grey level. The global and region features are inserted in to the tree structure. Then for retrieval, distance measures are applied to obtain a degree of similarity between the query image's features and the database image's features. In similarity measure, we planned to use feature weighting scheme, because different features have different amount of discriminatory characteristics. The approach aims to speed up the image retrieval process, effective usage of storage space, internal memory without compromising the accuracy of the retrieval.

The paper is organized as follows; Section 2 describes Global feature extraction including image preprocessing, global index features extraction. Section 3 detailed about region extraction including region preprocessing using k means clustering and region labeling and region index features. Data Structure for indexing using R* - tree structure is detailed in section 4. Section 5 describes similarity measure using Euclidean distance. Section 6 concludes the paper.

II. Global Features Extraction

This section will converse detailed about how to calculate or extract the global features from the image and also what preprocessing is happened before the extraction. The global features are features that are calculated by taking into a count the whole image.

A. Image Pre-Processing

The pre-processing stage of global features contributes two main things, one is quantization and another is edge detection.

1) Color Quantization

Prior to any processing being performed on a color image, color quantization is a very important step, due to the large number of different colors in the image. Color quantization is the procedure used to reduce possible colors to a small number. By using different quantization approaches, such as combining adjacent colors within a predefined range into one single color, the large color set can be reduced to a small number of possible colors. For example, an image can be quantized from true color with 16777216 possible colors, to only 64 possible colors so that any needed processing on it would be easier [26].

To be suitable for computer processing and features extraction (color), an image must be digitized in amplitude. The idea is to reduce the color space while gaining the ability to localize color information spatially. This project applies quantization at HSV color space. With less color space, we can extract the global features easier. In figure 1 we can see the overview 5 steps to perform quantization[27].

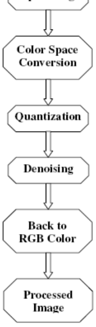


Fig 1. Steps of color quantization

The first step is to convert the color space from RGB into HSV (Hue Saturation Value). The following is the formula used to convert the value [28]:

$$H = \begin{cases} 60 \times \frac{G-B}{MAX-MIN} + 0, & \text{if } MAX = R \\ & \text{and } G \geq B \\ 60 \times \frac{G-B}{MAX-MIN} + 360, & \text{if } MAX = R \\ & \text{and } G < B \\ 60 \times \frac{B-R}{MAX-MIN} + 120, & \text{if } MAX = G \\ 60 \times \frac{R-G}{MAX-MIN} + 240, & \text{if } MAX = B \end{cases}$$

$$S = \frac{MAX - MIN}{MAX}$$

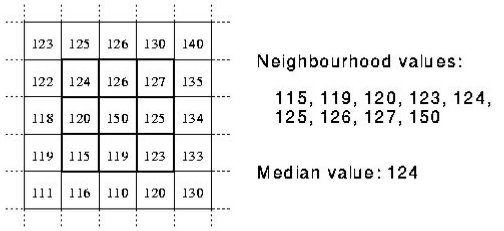
$$V = MAX$$

(1)

Where RGB values ranged from 0 to 1 and MIN/MAX corresponds with RGB values. In original image, each pixel has 3 layers of color which are red, green and blue. For each pixel we convert into HSV value.

The second step is the quantization. We reduce the color by grouping the converted HSV values to the closest predefine HSV value. Value and saturation represent blackness or whiteness and amount of color present respectively, while hue represent tint or tone of the color. The value and saturation divide into 3 values which are 0.33, 0.67 and 1, while the hue value is divided by 18 colors which is 20 degree separation.

The third step is to remove the noises that appear after quantization. We perform median filter to the quantized image. For each color layer, in each pixel is replaced by the median of the color levels in a neighbourhood of that pixel [29]. In figure 2 we can see an illustration of a pixel with its neighbourhood pixels and find the median value [28].



123	125	126	130	140
122	124	126	127	135
118	120	150	125	134
119	115	119	123	133
111	116	110	120	130

Neighbourhood values:

115, 119, 120, 123, 124, 125, 126, 127, 150

Median value: 124

Fig 2. Median filter [29]

The last step is to convert the color space for HSV into RGB. For each of the pixel we have HSV value that we can convert using the following formula back to RGB.

$$H_i = \left\lfloor \frac{H}{60} \right\rfloor \mod 6$$

$$f = \frac{H}{60} - H_i$$

$$p = V(1 - S)$$

$$q = V(1 - fS)$$

$$t = V(1 - (1 - f)S)$$

$$\text{if } H_i = 0 \rightarrow R = V, G = t, B = p$$

$$\text{if } H_i = 1 \rightarrow R = q, G = V, B = p$$

$$\text{if } H_i = 2 \rightarrow R = p, G = V, B = t$$

$$\text{if } H_i = 3 \rightarrow R = p, G = q, B = V$$

$$\text{if } H_i = 4 \rightarrow R = t, G = p, B = V$$

(2)

Where H is in the range [0, 360]; V and S are in the range [0, 1]. If the saturation value is equal to 0 then the RGB value will be equal to V, otherwise use the above formula. We can see from figure 4.3 some examples of the transformation. You can imagine how much of reduction in level of complexity in color. That is why we want to work with simpler color space.

2) Edge Detection using Sobel Operator

Sobel operator is used in image processing, particularly within edge detection algorithms. Technically, it is a discrete differentiation operator, computing an approximation of the gradient of the image intensity function [30]. Sobel operator is one way to do edge detection. This method detects the edges by looking for the maximum and minimum in the first derivative of the image. The Sobel filter performs a 2-D spatial gradient measurement on an image. Usually it finds the approximate absolute gradient magnitude at each point in an input image. It uses a pair of n X n convolution masks. Where the first mask detect changes in horizontal-direction and the other one detect changes in vertical-direction. In that G_X is the mask for detecting horizontal-direction, while G_Y is vertical-direction. We apply those masks into an image which will produce 2 images. The result images represent edge detection in horizontal and vertical direction. We then combine those two images into one image using the following formula:

$$|G| = \sqrt{Gx^2 + Gy^2} \qquad (3)$$

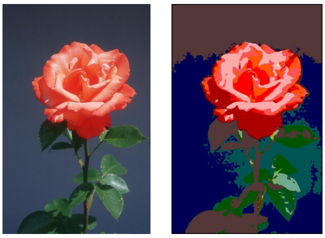
B. Global Index Features

This will use a selection of color based and simple shape based techniques. Given a limited time frame for experimentation and implementation, it was decided to use methods that did not require a segmentation step, or that use a very fast and simple one.

For this system we use six methods of defining regions on the image, and compared these against each other using each similarity algorithm. Each similarity algorithm uses a simple measure of an image property. This becomes a similarity algorithm on an image by computing the measure for all images in the database and for the query image I. The image in the database having a measure mi most like that of the query image Ii is said to be most similar to I. The six measures we used were: Color Average, Color sigma, edge density, Boolean edge density, edge direction, and color histograms.

1) Colour Average

In this feature we want to find an average number of every color layers in an image. Since I am using RGB color space for this feature, we will deal with three color layer



(a) (b)

Fig 3. (a) Original image; (b) transformed image

which are red, green and blue. For each layer we add every pixel value and divide by total number of pixels or width x height of the image.

$$A=1/m \sum_{n=0}^{n=m} (P_n) \tag{4}$$

Where, A is the average value, m is the total number of pixels, P_n is the nth pixel value. In this features we will get 3 values which represent red, green and blue average color values.

2) Colour Sigma

This feature represents pixel intensities in the image or in another word find the standard deviation. Since we are using RGB color space, there will be 3 values of standard deviation for each color layer. The standard deviation consists of intensity mean and variance. First we need to find the average value of each color layer, and then we find the variance. After

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots + x_N}{N} \tag{5}$$

Where, ¯x is the average value, N is the total number of pixels; x_i is the ith pixel value.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{6}$$

Where, σ is the standard deviation, ¯x is the average value, N is the total number of pixels, x_i is the ith pixel value

3) Edge Density

This is found by first using a standard edge detector (E.G. Sobel [30]) to enhance the pixels that belong to edges and boundaries. The result is a set of pixels whose values are in proportion to their residence on an edge; pixels far from an edge are 0, those near and edge increase to a maximum value. The edge density measure consists of the mean pixel value of the edge enhanced image [26]. To calculate this feature, we can use Sobel edge detection formula (3) to enhance the edge in the image and then use color average formula (4).

4) Boolean Edge Density

This feature is the next step from edge density feature. We count the number of pixel that is considered as an edge. First we perform edge density calculations which give us an enhanced image. The image is thresholded so that what could be called edge pixels are white (1) and non-edge pixels are black (0). In this case we use relevance feedback or trial and error to get the threshold. We run 50 images with different threshold and choose the best result. The initial threshold is the mean value of the image.

5) Edge Direction

Some edge detectors, including the Sobel edge detector, operate over a small (3x3) image region. This allows a crude estimate of edge direction to be made. In particular, for a typical 3x3 region in an image:

$$s_y = \begin{matrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{matrix} \qquad \begin{matrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{matrix} = s_x \tag{7}$$

The direction associated with the pixels in the region is an estimate of the gradient:

$$\theta = atan(S_y/S_x) \tag{8}$$

Where, Is the edge direction, S_y is equal with G_y, S_x is equal with G_x.

The edge direction metric computes an overall estimate of the direction of the edges in a region by calculating a resultant vector over all pixels, and using the direction of that resultant.

6) Hue and Intensity Histograms

We use a color histogram technique described in [31] which has the good sense to disregard achromatic information which is often included as noise in traditional color histogram techniques. It does this by calculating s, the standard deviation of the R,G,B components of a color pixel and normalizing to the range [0,1]. The chrominance of a pixel is determined using the function

$$\mu(\sigma) = \begin{cases} 0 & \text{if } 0 \le \sigma < a \\ 2\left(\frac{\sigma - a}{b - a}\right)^2 & \text{if } a \le \sigma < \frac{a+b}{2} \\ 1 - 2\left(\frac{\sigma - a}{b - a}\right)^2 & \text{if } \frac{a+b}{2} \le \sigma < b \\ 1 & \text{if } b \le \sigma < 1 \end{cases} \tag{9}$$

where a and b are constants between 0 and 1, where a<b. In the experiments described here, a=0.05 and b = 0.8 after some empirical trials. These values were computed and used to construct a color histogram with 16 bins. An intensity histogram was also created, having only 4 bins.

III. Region Features

In this section we will see how to cluster an image into regions base on intensity level and identify those regions base on spatial information. Then we calculate the region features for each of the region.

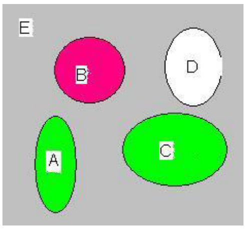


Fig 4. Image's region example

A. Region Pre-Processing

1) K-Mean Clustering

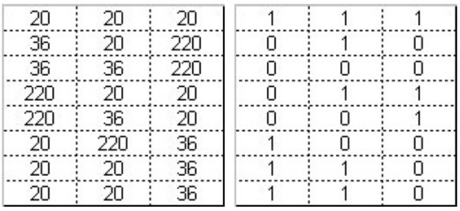
K-mean clustering is widely used for region segmentation which separates the image into regions base on intensity value. However, the traditional k-mean clustering only separates the regions without considering spatial information of the intensity values, so we only get regions in color space. In figure 4 there are 5 regions spatially but only 4 regions in color space, since A and C have the same color. According to [32], we can extend the traditional k-mean clustering to include spatial information which means each intensity value is grouped not only in color space but also into spatial coordinate. In this project we adapted what [32] proposed with what they called K-mean with connectivity (KMC).

We apply the traditional k-mean clustering which will give me K intensity values. We used intensity values instead of color values to reduce the complexity of computation. After getting regions in color space, we applied region labelling to detect the spatial information for each intensity value that will be discussed in more detail later on.

2) Region Labelling

The idea of region labelling is to group all pixels spatially base on an intensity value. For example, in figure 4 we want to be able to detect region A and C, which have the same colour but different location. With region labelling we can give label for each pixel in different region location, where each region has unique label. Jung-Me Park and Chen [2004] proposed a new version of connected component labelling algorithm, where they used divide and conquer approach to perform the labelling. This will improve the speed of the labelling a lot from hours or minutes into seconds. Unfortunately, we do not have time to implement this approach, so we use the traditional 8-connectivities component labeling algorithm [33].

First we used the result image from K-mean clustering, which give us with K intensity values. So, we will process this image K time with each time dealing with 1 intensity value. Given that labelling algorithm is used for binary image, we have to imagine that for each pixel value equal to current $^{-}$Ik, k=1,..., K, which is the k_{th} intensity value, is 1 while the rest are 0. For example, if we have 3 intensity values in the image such as 20, 36 and 220. We will process the image 3 times and the first loop is to image every pixel with value equal to 20 considers as 1 and the others are 0, see figure 5.2 for visualisation.

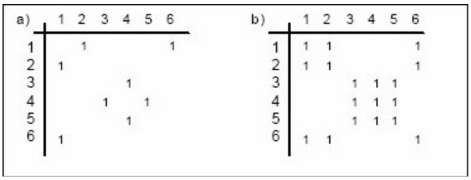


20	20	20
36	20	220
36	36	220
220	20	20
220	36	20
20	220	36
20	20	36
20	20	36

1	1	1
0	1	0
0	0	0
0	1	1
0	0	1
1	0	0
1	1	0
1	1	0

Fig 5. (a) Original image values; (b) Binary version

After convert the image into binary image we apply the labelling algorithm based on [29, 33]. First we initialize the image. Scan an image pixel by pixel, from left to right and from top to bottom. For current pixel p has 4-neigbours denote pixels in north (N), north-west (NW), north-east (NE) and west (W). If p is 0, move on to the next scanning position. If p is 1 and all four neighbours are 0, assign a new label to p. If only one of the neighbours is 1, assign its label to p. If two or more neighbours are 1, assign one of the labels to p and make a note of the appropriate equivalences, what it means is to store information that the two or more neighbours have the same label into some sort of 2D array (matrix); the index of the array represent the label number, see figure 5.3 for clear illustration.



a)	1	2	3	4	5	6
1		1				1
2	1					
3				1		
4			1		1	
5				1		
6	1					

b)	1	2	3	4	5	6
1	1	1				1
2	1	1				1
3			1	1	1	
4			1	1	1	
5			1	1	1	
6	1	1				1

Fig 6. (a) Original matrix; (b) Processed matrix

After completing the initialization, we sort out the equivalent label pairs into equivalence classes, assign unique number or label to each class, and do a second scan through the image, replacing each label by the label assigned to its equivalence class. In another word, we need to group all labels that are equivalence from the matrix that we have in initialize stage and assign new label to the groups which is called equivalence classes. Figure 6 show some illustration of a matrix, where picture (a) is original matrix and picture (b) is the processed one.

To resolve equivalence, there are 2 steps. The first one is to add reflexivity in the matrix; all main diagonals are set to 1. The second step is to obtain transitive closure using Floyd-Warshall algorithm [33]:

```
For j=1 to n
  For i=1 to n
    If L(i,j)=1 then
      For k=1 to n
        L(i,k)=L(i,k) OR L(j,k);
```

Where, L is the matrix. In figure 6 (b) is the result after applying reflexivity and the transitive closure.

B. Region Index Features

1) Region Area

This feature represents a number of pixels in a region. After we applied the region labelling algorithm, we can separate each region spatially. In figure 4, we count how many pixels in region A,B,..., E. The number will represent the region area.

2) Grey Level

When we applied the k-mean clustering we get K intensity value. Which mean there will be K different grey levels. Every region will have a grey value correspond with the K intensity value. In figure 5 the number 20, 36 and 220 is the intensity values; which value is the current region?

3) Moment Invariant

A set of seven moment invariants can be defined by combining the normalized central moments. I will show the formula to find those seven invariants which can be found in more detail in [29].

Moments:

$$m_{pq} = \sum_{x,y} x^p y^q f(x,y) \tag{10}$$

Where, p q the (p + q)th order of moment; f(x,y) is the pixel value at coordinate (x,y).

Centroid (balance point):

$$\bar{x} = \frac{1}{m_{00}} \sum x f(x,y) = \frac{m_{10}}{m_{00}}$$

$$\bar{y} = \frac{1}{m_{00}} \sum y f(x,y) = \frac{m_{01}}{m_{00}} \tag{11}$$

Where, ¯x is the balance in x coordinates; ¯y is the balance in y coordinates; xf(x,y) is the x coordinates in the region

Central Moments:

$$\mu_{pq} = \sum_{x,y} (x-\bar{x})^p (y-\bar{y})^q f(x,y) \tag{12}$$

Where, µpq is the central moments; x,y are the coordinates; ¯x is the balance in x coordinates; ¯y is the balance in y coordinates

Normalized Central Moments:

$$\gamma = 1 + \frac{p+q}{2}$$

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}{}^{\gamma}} \tag{13}$$

Where, _ is the normalized central moments; p q the (p + q)th order of moment.

Invariant Moments:

$$\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 \\
&\quad - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \\
&\quad [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
\phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
&\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\phi_7 &= (3\eta_{21} - \eta_{30})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 \\
&\quad - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03}) \\
&\quad [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2].
\end{aligned} \tag{14}$$

Where, _n is the nth of moment invariant, n=17.

IV. Data Structure for Indexing

A. R - Tree Structure:*

An image has high dimensional metric space; a tree structure is commonly used to store that information. Some common tree structure that can be used are R-tree [Guttman, 1984[34]], R*-tree [N. Beckmann and Seeger, 1990 [35]], VP-tree structure [Yianilos, 1993 [36]] and Hybrid Tree [Chakrabarti and Mehrotra, 1999 [37]]. For the sake of simplicity, this project is using quadratic R*-Tree structure for better performance and efficiency, which will be discussed in below section.

R* tree is a variant of R tree for indexing spatial information. R* tree supports point and spatial data at the same time with a slightly higher cost than other R-trees. It was proposed by Norbert Beckmann, Hans-Peter Kriegel, Ralf Schneider, Bernhard Seeger in 1990. Their performances are Likely significant improvement over other R tree variants, but there is overhead due to the reinsertion method, and efficiently supports point and spatial data at the same time. R*-Tree can be efficiently used as an access method in database systems organizing both multidimensional points and spatial data. R* -Tree are based on the reduction of the area, margin and overlap of the directory rectangles. The R* tree uses the same algorithm as the R-tree for query and delete operations. The primary difference is the insert algorithm, specifically how it chooses which branch to insert the new node into and the methodology for splitting a node that is full [35].

V. Similarity Measure

A. Euclidean Distance

We use Euclidean distance to find similarity between two images. This method is easy to use and quick to compute. Although Euclidean distance measure is not the very accurate according to [38], but it can does the job for retrieval in this project significantly. Due to some possibility for individual feature distances may have vastly different values, so the distance need to be normalised to avoid some features over power the others. We use Gaussian Normalisation to reduced the bias that would be introduced by an unusually large distances [39]. We calculate the similarity between index features of two images using the following formula [38]:

$$D_e(p,q) = \sqrt{\sum_{i=1}^{n}(Q_i - P_i)^2}\; W''_i \tag{16}$$

Where, p is the database image.
q is the query image.
Pi is the database images ith features.
Qi is the query's ith features.
n is the number of features.
W" is the weight for ith feature.

VI. Conclusion

We mention earlier that indexing give information to the retrieval process that can be used to find similarity between a query images with the database. Without appropriate information of image representation, the retrieval process cannot achieve good retrieval result. It is a fact that a system that performs best with one database does not have 100% certainty to perform best with different database. So, it is clear that indexing is affected by retrieval and images in the database. Since there is no universal indexing system that can handle every types of database, what we can do is to create a system that is as robust as possible. In this research multiple combinations of available features has been investigated at some point. It is shown in the results that region features are more significant than global features. In term of the indexing, the more features used as an index will increase precision of describing the image but will increase space and time to process it. Sometimes more features do not do more good, since redundancy can happen. Content-based image indexing is every complex matter, since it involves image processing, statistical analysis and other field that demand high level of mathematical calculation.

References

[1] "Content-based Image Retrieval" from http://en.wikipedia.org/wiki/CBIR.

[2] Yihong Gong, "Advancing Content-based Image Retrieval by Exploiting Image Color and Region Features", Multimedia Systems, Volume 7, Issue 6, Pages: 449 - 457, November 1999.

[3] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval: The End of the Early Years,"IEEE Trans. Pattern. Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

[4] N. Vasconcelos and M. Kunt, "Content-Based Retrieval from Image Databases: Current Solutions and Future Directions," Proc.Int'l Conf. Image Processing, 2001.

[5] J. Smith and S. Chang, "Visually Searching the Web for Content," Multimedia, vol. 4, no. 3, pp. 12-20, July-Sept. 1997.

[6]David Feng, et al,." Multimedia Information Retrieval and Management", Springer 1st edition, April 2003, ISBN-10: 3540002448, ISBN-13: 978-3540002444.

[7] Prakash, K.S.S.; Sundaram, R.M.D,"Combining Novel features for Content Based Image Retrieval", in proc. of EURASIP Conference on Speech and Image Processing, Multimedia Communications and Services,pp:373 - 376,June 2007.

[8]Dr. Fuhui Long, Dr. Hongjiang Zhang and Prof. David Dagan Feng,"Fundamentals of Content-Based Image Retrieval", chapter 1.

[9] Thomas M. Deserno, Sameer Antani, and Rodney Long," Ontology of Gaps in Content-Based Image Retrieval", in proc. of Journal on Digital Imaging", Feb.2008.

[10]J. De Bonet and P. Viola, "Structure Driven Image Database Retrieval," Proc. Conf. Advances in Neural Information Processing Systems, vol. 10, 1997

[11]J. Hafner, H. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient Color Histogram Indexing for Quadratic Form Distance Functions," IEEE Trans. Pattern. Analysis and Machine Intelligence, vol. 17, no. 7, pp. 729-736, July 1995.

[12] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Spatial Color Indexing and Applications," Int'l J. Computer Vision, vol. 35, no. 3, pp. 245-268, Dec. 1999.

[13] N. Vasconcelos and A. Lippman, "Library-Based Coding: A Representation for Efficient Video Compression and Retrieval," Proc. Data Compression Conf., 1997.

[14] Mehtre, B., Kankanhalli, M., Lee, W.: "Shape measures for content based image retrieval: A comparison". Information Processing Management 33 (1997) 319–337.

[15] Cullen, J., Hull, J., Hart, P.: "Document image database retrieval and browsing using texture analysis". In: Proc. 4th Int. Conf. Document Analysis and Recognition. (1997) 718–721.

[16] Lu, G., "Multimedia database management systems" in proc. of Artech House computing library. 1999.

[17]Christian Hartvedt,"Utilizing Context in Ranking Results from Distributed CBIR", in proceedings of NKIC, 2007.

[18]Chad Carson,et al,."Blobworld: Image Segmentation Using Expectation Maximization and Its Application to Image Querying", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 8, Aug.2002.

[19]Rouhollah Rahmani, Sally A. Goldman, Hui Zhang, Sharath R. Cholleti, and Jason E. Fritts,"Localized Content Based Image Retrieval",IEEE Transactions on Transactions On Pattern Analysis and Machine Intelligence, Special Issue, Nov. 2008.

[20] Suematsu, N., Ishida, Y., Hayashi, A., Kanbara, T., Region-based image retrieval using wavelet transform. In: Proc. 15th International Conf. on Vision Interface, May 27- 29, Calgary, Canada, pp. 9-16, 2002.

[21] P. Aigrain, H. Zhang, and D. Petkovic. Content-based representation and retrieval of visual media: A state-of-the-art review. Int. J. Multimedia Tools and Applications, 3:179– 202, November 1996.

[22] C. Frankel, M. J. Swain, and V. Athitsos. Webseer: An image search engine for the world wide web. Technical Report 96-14, University of Chicago, Computer Science Department, August 1996.

[23] M. Flickner, et al. Query by image and video content: the QBIC system. IEEE Computer, 28(9):23–32, 1995.

[24] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Tools for content-basedmanipulation of image databases. In SPIE Storage and Retrieval for Image and Video Databases II, volume 2, 185, pages 34–47, San Jose, CA, 1994.

[25] J. Smith and S. Chang. Visually searching the web for content. IEEE Multimedia, 4(3):12–20, 1997.

[26] J. R. Parker,Brad Behm, "Use of Multiple Algorithms in Image Content Searches",Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04), Volume 2, Page: 246, 2004

[27] Smith, J. and Chang, S. [1995]. Single color extraction and image query. In Proc. IEEE Int. Conf. on Image Proc., pages 528-531, 1995.

[28] Wikipedia [2006]. Hsv color space, on website. page was last modified 07:29, 12 June 2006. URL: http://en.wikipedia.org/wiki/HSV color space

[29] Gonzalez, R. and Woods, R. [1992]. Digital Image Processing, third edn, Addison-Wesley Publishing Company.

[30] J. Parker, Algorithms for Image Processing and Computer Vision, John Wiley & Sons Ltd., New York, 1996.

[31] [13] M. Tico, T. Haverinen, and P. Kuosmanen. A method of color histogram creation for image retrieval, In Proceedings of NORSIG, 2000.

[32] Kompatsiaris, I., Triantafillou, E. and Strintzis, M. [2001]. Region-based color image indexing and retrieval, International Conference on Image Processing (ICIP2001), pp. 658– 661.

[33] Jung-Me Park, C. G. L. and Chen, H.-C. [2004]. fast connected component labeling algorithm using a divide and conquer technique, Technical report, Computer Science Dept. University of Alabama, Tuscaloosa and University of Nevada, Reno.

[34] Guttman, A. [1984]. R-tree: a dynamic index structure for spatial searching, ACM SIGMOD Int. Conf. Management of Data, Boston, MA, pp. 47–54.

[35] N. Beckmann, H.-P. Kriegel, R. S. and Seeger, B. [1990]. The r*-tree: An efficient and robust access method for points and rectangles, ACM SIGMOD Int'l. Conf. on Management of Data, pp. 322–331.

[36] Yianilos, P. N. [1993]. Data structures and algorithms for nearest neighbor search in general metric spaces, SODA: ACM-SIAM Symposium on Discrete Algorithms (A Conference on Theoretical and Experimental Analysis of Discrete Algorithms).

[37] Chakrabarti, K. and Mehrotra, S. [1999]. The hybrid tree: An index structure for high dimensional feature spaces, ICDE, pp. 440–447.

[38] Hore, E. S. [2000]. Content-based image retrieval, Honours thesis, Monash University, Australia Clayton Campus.

[39] Syrjsuo, M. T. and Donovan, E. F. [2005]. Using relevance feedback in retrieving auroral images, Int. Conf. on Computational Intelligence (CI 2005), pp. 420–425.

A Novel Software Reliability Growth Model of Imperfect Debugging

Yanjun Shu, Hongwei Liu, Zhibo Wu, and Xiaozong Yang

***Abstract*—Most software testing processes are imperfect debugging ones and the removal efficiency measures how perfect the debugging process is. In many software reliability growth models (SRGMs), the removal efficiency is assumed as a constant during the debugging. However in practice, the removal efficiency is usually varying with the debugging time because of the debugging personnel's learning ability. In this paper, we investigate the varying trend of removal efficiency by studying the relationship between the number of detected faults and the number of removed faults. The removal efficiency is found to be S-shaped and the logistic curve is appropriate to characterize it. Then, a new imperfect debugging SRGM is derived based on the removal efficiency function. Numerical experimental results show that the proposed model performs very well on the real data sets.**

***Index Terms*—imperfect debugging; learning process; software reliability growth models; debugging process.**

I. INTRODUCTION

Computers are now widely used to control safety-critical and civilian systems. Software reliability is a primary concern for both software developers and software users, and it must be evaluated carefully. Many mathematical models called software reliability growth models (SRGMs) have been developed to describe the software debugging process [1]-[6].

It is widely recognized that debugging processes are imperfect. Software faults may not be removed perfectly and instantly because of the difficulty in locating them, so the number of removed faults is less than the number of detected faults during debugging processes. In other words, the probability of removing the detected fault is not always 100%, and it can be defined as "removal efficiency". The removal efficiency measures how perfect the detected fault can be removed. Although many SRGMs have been proposed to study the imperfect debugging phenomenon [7]-[11], only a few of them have taken the removal efficiency into account. Goel and Okumoto [9] considered the removal efficiency in their Markov model. They assumed that after a failure the residual faults reduce to one less than current value with probability *p*. X. Zhang *et al.* [10] introduced a similar conception in their general NHPP software reliability model. They defined the percentage of faults eliminated by reviews and tests as *p*.

The removal efficiency is assumed to be a constant which is independent of the debugging time. However, this assumption is not very realistic because the learning ability of the debugging personnel can influence the software debugging process. In some SRGMs, the learning process has been incorporated into imperfect debugging [10], [12]-[17].

In practice, when a fault is detected, the debugging personnel study the software fault and modify codes to remove it. The removal efficiency depends on the learning process of debugging personnel. It is not a constant during the debugging process. In this paper, we propose an imperfect debugging SRGM with varying removal efficiency. The rest of this paper is organized as follows. In Section II, the relationship between the number of the detected faults and the number of removed faults with debugging time is discussed. And then the logistic curve is found to be appropriate to describe the removal efficiency. In Section III, a novel imperfect debugging SRGM is derived based on the removal efficiency function. Numerical examples and comparison results are presented in Section IV. Finally, the paper is concluded in Section V.

Acronym & Notations

NHPP	non-homogeneous poisson process
G-O	Goel-Okumoto
MVF	mean value function
SSE	sum of squared errors
$m(t)$	the expected number of faults detected in time $(0, t)$
$m_r(t)$	the expected number of faults removed in time $(0, t)$
$p(t)$	the removal efficiency function
a	the expected number of initial faults
b	the fault detection rate
y_k	the number of actual faults observed at time t_k

Manuscript received May 10, 2009. (Write the date on which you submitted your paper for review.) The work described in this paper is jointly supported by Hi-Tech Research and Development Program of China (2008AA01A201) and National Nature Science Foundation of China (60503015).

Yanjun Shu is a lecturer in Harbin Institute of Technology, Harbin, Heilongjiang 150001 China (corresponding author to provide phone & fax: 0451-86414093; e-mail: yjshu@hit.edu.cn).

Hongwei Liu is an associate professor in Harbin Institute of Technology. (e-mail: lhw@ftcl.hit.edu.cn).

Zhibbo Wu is a professor in Harbin Institute of Technology. (e-mail: wzb@ftcl.hit.edu.cn).

Xiaozong Yang is Director of Computer Science and Technology School in Harbin Institute of Technology. He is also Vice Director of Fault Tolerant Computing Committee under China Computer Federation.. (e-mail: xzyang@ftcl.hit.edu.cn).

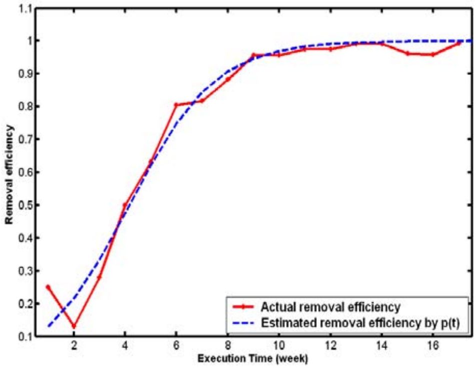


Fig. 1 Removal efficiency function *p(t)* on a middle size software project

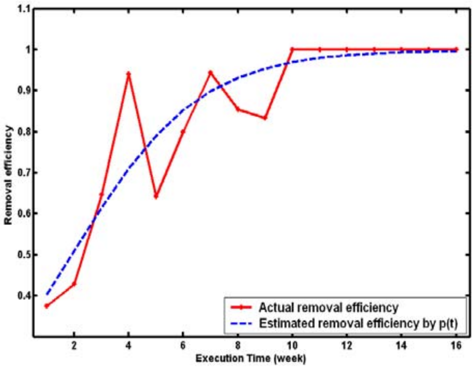


Fig. 2 Removal efficiency function *p(t)* on System T1

II. REMOVAL EFFICIENCY FUNCTION

The removal efficiency is always assumed as a constant in the imperfect debugging SRGMs, which means the detected faults are removed with a fixed probability and the number of removed faults is linear with the number of detected faults. In fact, the relationship between the number of removed faults and the number of detected faults can not be simply characterized as linearity. Wu et al. [6] recorded the testing data of a middle size software project which includes both the detected fault number and the removed fault number. We calculate the actual ratio value of this data set and plot it in the Fig. 1 as a solid line. As show in the figure, the actual ratio is increasing with debugging time, which indicates that the removal efficiency is enhanced during the debugging process. Shbita et al. [18] reported a data set of System T1 which includes the number of detected and corrected faults. Fig. 2 depicts the actual removal efficiency of T1 as a solid line. We also can see from the figure that the removal efficiency has an “increasing” scenario.

The removal efficiency is very low at the beginning of software testing process because the testing personnel are not familiar with the software structure and they may not be able to remove the detected faults quickly. As time going on, the removal efficiency increases rapidly in the middle of testing process. This phenomenon is caused by the leaning process. The testing personnel gain more experience and grasp the software better than before and they can remove faults quickly. At the late stage, the removal efficiency increases little by little because the faults in the software are almost completely detected and corrected, the fault correction process is very close to the fault detection process. The ratio reaches the maximum value which approximates to 100%. The increment of *p(t)* versus the testing time appears S-shaped which represents the testing personnel’s learning ability. This kind of non-decreasing trend can be described by logistic function, as denoted in (1).

$$p(t) = \frac{1}{1+\eta e^{-\partial \cdot t}} \tag{1}$$

III. SOFTWARE RELIABILITY MODELING

In this section, the imperfect debugging phenomena incorporate into the software reliability growth modeling based on Non-Homogeneous Poisson Process (NHPP). The fault removal efficiency is used to characterize the effectiveness of the debugging process. As studied in the section II, the fault removal efficiency has an “increasing” feature, and the logistical curve is selected to represent it. Most of NHPP SRGMs have the following assumptions on the debugging process [6], [7]-[10], [12]:

1) The failures of software debugging obey NHPP.
2) The software system is subject to failures at random time caused by software faults remaining in the software.
3) All faults are independent and equally detectable.
4) The software failure rate at any time is affected by the number of faults remaining in the software at that time.

According to these assumptions, we have

$$\frac{dm(t)}{dt} = b(t)[a(t) - m_r(t)] \tag{2}$$

The fault removal efficiency measures the effectiveness of the debugging, and it can be transformed as:

$$m_r(t) = \frac{1}{1+\eta \cdot e^{-\partial \cdot t}} \cdot m(t) \tag{3}$$

Substituting (3) into (2), we obtain a new differential equation as

$$\frac{dm(t)}{dt} = b(t)\left[a(t) - \frac{1}{1+\eta \cdot e^{-\partial \cdot t}} \cdot m(t)\right] \tag{4}$$

To solve the simultaneous differential equations simply and easily, we assuming the *a*(*t*), *b*(*t*) are constant, which means non fault is introduced and the failure rate is not changing during the software testing. So (5) is changed into:

$$\frac{dm(t)}{dt} = b\left[a - \frac{1}{1+\eta \cdot e^{-\partial \cdot t}} \cdot m(t)\right] \tag{6}$$

when $\partial = 0$, then $r(t) = 1/(1+\eta)$, it means the detected fault is $m(t) = a(1+\eta)[1-\exp(-\frac{bt}{1+\eta})]$, it becomes the SRGM which removed as a fixed probability. In this case, solve (6) withmarginal conditions *m*(0)=0, we can get is proposed by M. Ohba [7]. when $\eta = 0$, then $r(t) = 1$, it means the detected fault can be removed perfectly. In this case, solve (6) with marginal condition $m(0) = 0$, we can get $m(t) = a(1 - e^{-bt})$, the model become the Goel-Okumoto (G-O) model, which is a fundamental SRGM. Otherwise, solving (6), we get a new imperfect debugging SRGM by considering the fault removal efficiency:

$$m(t) = a\left[1 - \left(\frac{1+\eta}{1+\eta e^{-\partial t}}\right)^{b/\partial} e^{-bt}\right] - \frac{ab}{\partial}\left[\ln(1+\eta e^{-\partial t}) - \ln(1+\eta)\left(\frac{1+\eta}{1+\eta e^{-\partial t}}\right)^{b/\partial} e^{-bt}\right] \quad (7)$$

IV. EVALUATION OF THE PROPOSED MODEL

In this section, we will evaluate goodness-of-fit ability and predictive power of the proposed model by using two real data sets. The first data set (DS1) is from a middle size software project [6], and the second data set (DS2) is from System T1 of the Roma Air Development Center project [18]. Table I summarizes some famous software reliability models which considered the imperfect debugging in their modeling.

A. Criteria for Model Comparison

The goodness of fit of the curve is measured by the sum of squares errors, SSE. SSE sums up the squares of the residuals between the actual data and the mean value function $m(t)$ of each model in terms of the number of actual faults at any time point. The *SSE* function can be expresses as follows:

$$SSE = \sum_{k=1}^{n} \left(y_k - \hat{m}(t_k)\right)^2 \quad (8)$$

where y_k is the total number of faults observed at time t_k according to the testing time and $\hat{m}(t_k)$ is the estimated cumulative number of faults at time t_k obtained from the fitted value function, $k = 1,2,...n$. Therefore, the lower the SSE value, the better the model performs.

B. Performance Analysis

1) Example on DS1

To test the goodness of fit ability and the predictive power, researchers use a subset of actual data to fit the models and then predict the future failure. The first 12 data points are used for the goodness of fit test, whereas the remaining data are used for the predictive power test. Table II lists the results of models in terms of *SSE*. Although the proposed model does not give the lowest *SSE* value neither fit nor prediction, it provides the lowest sum *SSE* value. On the average, the proposed model shows better than other models.As can be seen from Table II, K.Y model and P.Z model also have good performance on DS1. Then, we plot the fit and predict data by using the proposed model, K.Y model, and P. Z model in Fig. 3. We can see from the figure that the proposed model not only can fit the data sets well, but also can predict the failure occurrence accurately.

2) Example on DS2

We also fit the models by 70% of DS2 (from week 1 to week 11), and predict the future failure from week 12 to week 16. Table III lists the results of models in terms of *SSE*. We can see from the table that the proposed model provides the lowest value of *SSE* for both fit and predict. For the sum of *SSE*, the proposed model, Delay S-shaped model, P.Z model, and P.N.Z model show better results then other models. Since P.Z model and P.N.Z model have same sum value of *SSE*, we only plot the fit and predict results by using the proposed model, Delay S-shaped model, and P.N.Z model in Fig. 4. As can be seen from the figure, the proposed model shows better performance than other two models. Overall, we can conclude that the proposed model has the best result for both goodness of fit and predictive test on DS2.

TABLE I
SUMMARY OF SOFTWARE RELIABILITY MODELS

Model Name	MVF (*m(t)*)
G-O model	$m(t) = a(1 - e^{-bt})$
Delay S-shaped	$m(t) = a(1 - (1 + bt)e^{-bt})$
P.Z model	$m(t) = \frac{1}{1+\beta e^{-bt}}[(c+a)(1-e^{-bt}) - \frac{a}{b-\alpha}(e^{-\alpha t} - e^{-bt})]$
P.N.Z model	$m(t) = \frac{a}{1+\beta e^{-b(1+\beta)t}}\left[(1-e^{-b(1+\beta)t})\left(1 - \frac{\alpha}{b(1+\beta)}\right) + \alpha \cdot t\right]$
K.Y model	$m(t) = a\frac{b}{c}\ln\left(\frac{b}{(b-c)+ce^{-bt}}\right)$
Z.T.P model	$m(t) = \frac{a}{p-\beta}\left[1 - \left(\frac{(1+\alpha)e^{-bt}}{1+\alpha e^{-bt}}\right)^{\frac{c}{b}(p-\beta)}\right]$

TABLE II
THE COMPARISION RESULTS OF MODELS ON DS1

Model	*SSE*(fit)	*SSE*(predict)
G-O model	770.40	62.08
Delay S-shaped model	763.94	246.60
K.Y model	722.98	62.12
P.Z model	142.98	454.86
P.N.Z model	675.79	668.22
Z.T.P model	432.27	530.22
Proposed model	198.91	73.50

TABLE III
THE COMPARISION RESULTS OF MODELS ON DS2

Model	*SSE*(fit)	*SSE*(predict)
G-O model	79.96	301.24
Delay S-shaped model	117.50	17.92
K.Y model	77.47	240.16
P.Z model	70.85	74.79
P.N.Z model	70.85	74.79
Z.T.P model	95.22	222.79
Proposed model	57.11	5.79

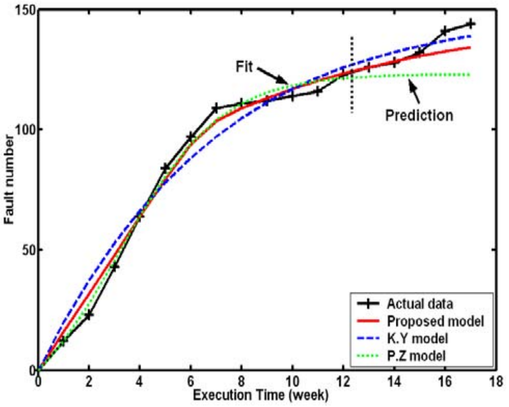


Fig. 3 The comparison results of the proposed model, K.Y model and P.Z model on DS1

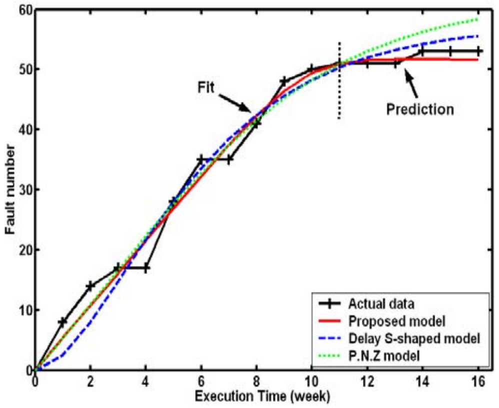


Fig. 4 The comparison results of the proposed model, Delay S-shaped model and P.N.Z model on DS2

V. CONCLUSIONS

In this paper, learning process is considered in imperfect debugging by using variable removal efficiency which is usually ignored or treated as a constant in previous published SRGMs. We investigate the varying trend of removal efficiency on some real data sets, and find the removal efficiency is increase with debugging time. Based on the analysis of the removal efficiency, the logistic curve is selected to characterize it. Then, a novel imperfect debugging SRGM is proposed. Finally, numerical examples are presented and experimental results show that the proposed model has better goodness of fit and predictive power than other models. For convince, our proposed model has temporarily ignored the situation that new faults may be introduced while removing the existing detected faults. The future work is to consider the possibility of fault introduction to improve the imperfect-debugging SRGM.

REFERENCES

[1] M. Lyu, editor, *Handbook of Software Reliability Engineering*, McGraw-Hill and IEEE Computer Society, New York, 1996.

[2] J. Musa, A. Iannino, and K. Okumoto, *Software Reliability Measurement, Prediction and Application*, McGraw-Hill, New York, 1987.

[3] M. Ohba, "Software reliability analysis models", IBM J. Research & Development, 1984, 28, pp. 328-443.

[4] H. Pham and X. Zhang, "An NHPP software reliability models and its comparison", The International Journal on Systems Science, 1996, 27(5), pp. 455~463.

[5] J. Zhao, H. Liu, G. Cui, and X. Yang, "Software reliability growth model with change-point and environment function," *The Journal of Systems and Software*, 2006, (79), pp. 1578-1587

[6] Y. Wu, Q. Hu, M. Xie, and S. Ng, "Modeling and analysis of software fault detection and correction process by considering time dependency", *IEEE Transactions on Reliability*, 2007, 56(4), pp. 629-642.

[7] S. Yamada, K. Tokuno, and S. Osaki, "Imperfect debugging models with fault introduction rate for software reliability assessment", *International Journal System Science*, 1992, 23(12).

[8] M. Ohba and X. M. Chou, "Does imperfect debugging affect software reliability growth?", *11th International Conference on Software Engineering*, 1989, pp. 237-244.

[9] A. Goel and K. Okumoto, "A Markovian model for reliability and other performance measure of software systems", *AFIPS* conference, 1979.

[10] X. Zhang, X. Teng, and H. Phan, "Considering fault removal efficiency in software reliability assessment", *IEEE Transaction on System, Man, and Cybernetics-Part A: Systems and Human*, January 2003, 33(1), pp. 114-120.

[11] S. Gokhale, M. Lyu and K. Trivedi, "Incorporating fault debugging activities into software reliability models: a simulation approach", *IEEE Transaction on Reliability*, 2006, 55(2), pp. 281-292.

[12] H. Pham, L. Nordmann, and X. Zahng, "A general imperfect debugging with S-shaped fault detection rate," *IEEE Transaction on Reliability*, 1997,14(3), pp. 269-282.

[13] P. Kapur and S. Younes, "Modeling an imperfect debugging phenomenon in software reliability", *Microelectronics and Reliability*. 1996, 36(5), pp. 645-660.

[14] S. Yamada, M. Ohba, and S. Osaki, “S-shaped reliability growth modeling for software fault detection”, *IEEE Transaction on Reliability*, 1983, 12, pp. 475-484.

[15] M. Ohba, “Inflextion S-shaped software reliability growth models”, in *Stochastic Models in Reliability theory*, S. Osaki and Y. Hatoyama, editors. Berlin, Germany: Springer-Verlag, 1984, pp. 144-162.

[16] K. Chiu, Y. Huang, and T. Lee, “A Study of reliability growth from the perspective of Learning Effects”, *Reliability Engineering & System Safety*, 2008, (93), pp. 1410-1421.

[17] G. Abu, and J. Cangussu. “A quantitative learning Model for software testing process”, *38th International Conference on System Science*, Hawaii, USA, 2005, pp. 78-88

[18] K. Shibata, K. Rinsaka, T.Dohi, and H. Okamura. “Quantifying software maintainability based on a fault-detection/correction model”, *Proceedings of 13th IEEE International Symposium on Pacific Rim Dependable Computing*, Merbourne, Vectoria, Australia, 2007, pp. 35-42.

[19] C. Huang, and C. Lin. Software reliability analysis by considering fault dependency and debugging time lag. IEEE Transaction on Reliability, 2006, 55(3), pp: 436-450.

Yanjun Shu received BS degree (2002) and MS degree (2004) in computer science and technology from Harbin Institute of Technology, China..She is a lecture at Harbin Institute of Technology from 2004. research interests include software testing, software reliability evaluation, fault tolerance computing. Lecture Shu is a member of China Computer Federation.

Hongwei Liu received BS degree (1994), MS degree (1999) and PhD degree (2004) in computer science and technology from Harbin Institute of Technology, China. He is an associate professor at Harbin Institute of Technology. His research interests include software testing, reliability evaluation, fault tolerance computing, *etc*. Dr. Liu is a senior member of China Computer Federation

Zhibo Wu received PhD degree (1987) in computer science and technology from Harbin Institute of Technology, China. He is an professor at Harbin Institute of Technology. His research interests include software testing, reliability evaluation, fault tolerance computing, *etc*. Dr. Wu is a senior member of China Computer Federation

Xiaozong Yang received BS degree (1964) in computer science and technology from Harbin Institute of Technology, China. He is an professor at Harbin Institute of Technology. His research interests include software testing, reliability evaluation, fault tolerance computing, *etc*. Dr. Wu is a senior member of China Computer Federation