

International Review on Computers and Software (IRECOS)

Contents

Optimal Selection of Wavelet and Threshold Using Cuckoo Search for Noise Suppression in Speech Signals <i>by Harjeet Kaur, Rajneesh Talwar</i>	1496
Survey and Analysis of Visual Secret Sharing Techniques <i>by L. Jani Anbarasi, G. S. Anandha Mala</i>	1507
Method for Automatic Ontology Building in Costumer Support Expert System for Energy Consumption <i>by A. Stropnik, M. Zorman</i>	1518
Proof of Retrievability Using Elliptic Curve Digital Signature in Cloud Computing <i>by Sumathi D., S. Kathik</i>	1526
Hybridization of ABC and PSO for Optimal Rule Extraction from Knowledge Discovery Database <i>by K. Jayavani, G. M. Kadhar Nawaz</i>	1533
{0, 1, 3}-NAF Representation and Algorithms for Lightweight Elliptic Curve Cryptosystem in Lopez Dahab Model <i>by Sharifah M. Y., Rozi Nor Haizan N., Jamilah D., Zaitun M.</i>	1541
Hybrid Re-Clustering Algorithm for Enhancement of Network Lifetime in Wireless Sensor Networks <i>by Aby K. Thomas, R. Devanathan</i>	1548
Smart Sentinel: Monitoring and Prevention System in the Smart Cities <i>by J. M. Sánchez Bernabéu, J. V. Berná Martínez, F. Maciá Pérez</i>	1554
Hybrid Fusion Technique Using Dual Tree Complex Wavelet Transform for Satellite Remote Sensor Images <i>by G. Dheepa, S. Sukumaran</i>	1560
Effect of Sensing Time Variation on Detection, Misdetction and False Alarm Probabilities in Cognitive Radio-Based Wireless Sensor Networks <i>by J. A. Abolarinwa, N. M. Abdul Latiff, S. K. Syed-Yusof, N. Fisal, N. Salawu</i>	1568
An Efficient Hybrid Segmentation Algorithm for Computer Tomography Image Segmentation <i>by V. V. Gomathi, S. Karthikeyan</i>	1576
Practical Analysis of Impact of Transmitter Hardware Impairments for MIMO Channel Measurement <i>by P. Vijayakumar, S. Malarvizhi</i>	1583
Meta-Classifer Based on Boosted Approach for Object Class Recognition <i>by Noridayu Manshor, Amir Rizaan Abdul Rahiman, Raja Azlina Raja Mahmood</i>	1590
Exploration of Heterogeneous Resources in Embedded Systems <i>by Aissam Berrahou, Nassim Sefrioui, Ouafaa Diouri, Mohsine Eleuldj</i>	1597
A Navigation-Aided Framework for 3D Map Views on Mobile Devices <i>by Adamu Abubakar, Sadegh Ameri, Suhaimi Ibrahim, Teddy Mantoro</i>	1605

(continued)

Model-Driven Transformation for GWT with Approach by Modeling: From UML Model to MVP Web Applications <i>by R. Esbai, M. Erramdani, S. Mbarki</i>	1612
An Adaptive Bilateral Filter for Noise Removal and Novel Hybrid Fuzzy Cognitive Map-FNN Edge Detection Method for Images <i>by T. Karthikeyan, N. P. Revathy</i>	1621
Adaptive Resource Allocation Mechanism (ARM) for Efficient Load Balancing in WiMAX Network <i>by P. Kavitha, R. Uma Rani</i>	1630
Adaptive Algorithm for Beacon and Superframe Values in IEEE802.15.4 Based Networks <i>by Wail Mardini, Abdulaziz Alraddadi</i>	1637



Optimal Selection of Wavelet and Threshold Using Cuckoo Search for Noise Suppression in Speech Signals

Harjeet Kaur¹, Rajneesh Talwar²

Abstract – Application of threshold based techniques to wavelet transformed speech signals can act as the noise suppression algorithm. There are types of wavelets and thresholding algorithms which exist and selection of optimal wavelet and optimal thresholding technique based on the input speech signal is vital for having better noise suppression. Selection of optimal number of decomposition levels is also has an impact on the resultant signal. In this paper, optimal selection of wavelet, level and thresholding technique for noise suppression in speech signals using Cuckoo search is proposed. After finding these, the signal is wavelet transformed and is applied proposed adaptive coefficient process which is then done thresholding. Subsequently, reconstruction is carried out to have the noise suppressed signal. The implementation is carried out with MATLAB and the evaluation metrics employed are Itakura–Saito distance (IS) and MSE. The results are taken under various noise conditions and compared with existing technique. From the results obtained, IS and MSE values for proposed is far lower than existing technique. Total IS average for proposed was about 0.78×10^5 compared with 1.3×10^5 that of existing. Total MSE average for proposed was about 0.22×10^{-3} compared with 1.25×10^{-3} that of existing. The results show the effectiveness of the proposed technique. **Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords: Noise Suppression, DWT Transform, Wavelets, Thresholding Techniques, Optimization, Cuckoo Search

Nomenclature

ψ	Mother Wavelet
d	Translation factor
$\psi_{c,d}(t)$	Shifted and scaled version of mother wavelet
Wf	Continuous Wavelet Transform
Th	Threshold
\hat{b}	Reconstructed signal
En	Entropy
FM	Factor Matrix
Fit	Fitness Function
D_{IS}	Distance
HN	Host Nest

I. Introduction

Noise suppression in speech signals is targeted at fine-tuning the performance of speech communication systems in noisy surroundings which has emerged as a catch-22 situation in signal processing. It is a unique application which is found extensively applied in various domains like as mobile radio communication techniques, speech to text methods, speech detection mechanisms and, inferior quality recordings [1]-[33].

It has also appeared as a boon to the hearing aid users,

efficiently tackling the intricacy involved in comprehension of speech in a noisy environment [10]. Noise diminution and suppression in the speech signal have been a subject matter of investigation for umpteen numbers of years [11], [30].

The deformation in noisy speech signals can be separated into two kinds such as those that have an adverse effect on the speech signal itself known as speech distortion and those that have a negative sway on the background noise known by the term noise distortion [12]. The most widespread distortion in speech is because of additive noise, which does not depend on the clean speech [13].

To tackle these deformations, there exists many a noise suppression algorithm [14]. The current methods at present handling these tasks are the conventional ones like spectral subtraction [15], Wiener filtering [9], and Ephraim Malah filtering [16]. It is pertinent to note that Wavelet-based techniques employing coefficient thresholding methods have also been extensively in vogue for speech enhancement [17], [31]. Wavelet transform (WT) [18] is a highly effective device for a varied number of signal processing [19] and compression applications [20]. Its crucial, and most beneficial, application domains are those that engender or process wideband signals.

This transform creates a time-frequency decay of the signal under investigation, and is competent of dividing

individual signal component more efficiently than the conventional Fourier scrutiny. In this way, this technique symbolizes signals efficiently and in various stages of resolution, which is the ideal requirement for decay and rebuilding objectives. This makes the discrete wavelet transform a proficient technique to digitally eliminate noises from the speech signals.

The wavelet type to be employed for the distinct wavelet analysis is an imperative decision for this dispensation. The wavelets are of two kinds namely the orthogonal wavelets such as Haar, Daubechies (dbxx), Coiflets (coifx), Symlets (symx), and the bi-orthogonal (biorx x), where x represents the order of the wavelet and the upper the order, the smoother the wavelet. As the orthogonal wavelets are not outmoded, they can be effectively employed for signal or image de-noising and compression [1]. A bi-orthogonal wavelet is a wavelet where the related wavelet transform is invertible but not inevitably orthogonal. Therefore, for noise elimination, employment of orthogonal wavelet transform is capable of heaping rich harvests by way of superior outcomes. [1].

When we apply threshold to the DWT processed speech signals, it acts as the noise suppression algorithm. There exist many thresholding algorithms and a majority of them are founded on the threshold definition offered in Donogo's universal theory [21]. Various kinds of prefixed thresholding [22] techniques embrace universal thresholding, universal thresholding level dependent, and modified universal thresholding level dependent, exponential thresholding, exponential thresholding level dependent and minimax thresholding.

There are a significant number of threshold proposals that have been positively assessed [23] and, in certain cases; they amass superior outcomes in relation to those of the prefixed thresholds. These comprise maxcoef threshold [24] and Kurtosis and ECE-based thresholds [23]. For signal de-noising, once the threshold to be executed is chosen, it is rescaled using noise variance in order to achieve superior outcomes and noise riddance [1].

In this paper, optimal selection of wavelet, level and thresholding technique for noise suppression in speech signals using Cuckoo search is proposed. After finding these, the signal is wavelet transformed by the found out optimal wavelet and optimal number of levels. After the wavelet transformation, the adaptive coefficient process is carried out.

Subsequently, it is applied thresholding by the found out optimal thresholding technique and is reconstructed to have the noise suppressed signal. The implementation is carried out with MATLAB and the evaluation metrics employed are Itakura-Saito distance (IS) and Minimum Squared Error (MSE).

The rest of the paper is organized as follows: A brief review of researches related to the proposed technique is presented in section 2. Noise suppression technique is given in section 3 and the proposed technique with contribution is presented in section 4.

The detailed experimental results and discussion are given in section 5. The conclusion is summed up in section 6.

II. Literature Review

Literature presents lot of works for noise suppression in speech and in this section, some of work related to it is reviewed. E. Castillo et al. [1] have shrewdly demonstrated the application of the discrete wavelet transform (DWT) for wandering and noise suppression in electrocardiographic (ECG) signals.

A one-step implementation is offered, which facilitates the upgrade of the entire de-noising procedure. Moreover, a comprehensive investigation is conducted, demarcating threshold restrictions and thresholding regulations for optimal wavelet de-noising employing the innovative method. Li Ruwei et al. [2] have been skilled enough to offer a speech enrichment scheme employing adaptive wavelet threshold and spectral subtraction in the wavelet domain. At first, so as to uphold the linear phase and block the spectral aliasing of reconstructed signals, the noisy signals are decayed by un-decimated bi-orthogonal wavelet packet. Subsequently, spectral subtraction is made use of for decreasing noise from the two low-frequency sub-bands, which eliminated the extreme deformation of low-frequency components triggered by wavelet de-noising. Once this is done, the improved wavelet shrinkage algorithm is executed in the supplementary high-frequency sub-bands. This threshold is revised by adaptively pursuing the intensity of noise. At last, these modified wavelet coefficients are rebuilt to attain the improved speech signals by means of the inverse wavelet packet transform.

Brady Laska et al. [3] has precisely established the fact that the intra-speech residual noise in RBPF enhanced speech is caused mainly from under shrinkage of noise in spectral troughs. Low-order full band TVAR models are not enough to capture the extremely vibrant array of the wideband speech spectrum, while high-order models deeply swell intricacy and consequently, making it very complicated to consistently make assessments. With an eye on scaling up the noise diminution efficiency, they have employed the RBPF algorithm which improves the speech DCT coefficients in place of the time interval for Wiener filtering on account of the significant competence of the DCT to proficiently de-correlate speech. Operating in the transform domain has enabled the statistical model approaches to present minimal intricacy and soaring echelons of noise decline, including colored noises. As a result, undue variations of the noise signal spectrum from its anticipated value has resulted inconsecutive over and under-estimation of the true noise spectrum in a specified frame, with the consequence of the popular time-varying narrowband musical noise works of art.

Slavy G. Mihov et al. [4] fascinatingly investigated the application of wavelet transform for de-noising speech signals tainted with frequent noises.

They employed the fundamental tenets of wavelet transform as a substitute to the Fourier transform. The experimental outcomes gathered were based on processing a huge committed database of reference speech signals polluted with diverse noises in various SNRs. This investigation was an expansion to the realistic investigation for speech signal development for the purposes of hearing-aid devices.

Brady N. M. Laska et al. [5] have profusely put forward a particle filter method to spectral amplitude speech augmentation. Spectral amplitudes were found to demonstrate inter-frame dependencies and non-Gaussian statistics; nevertheless, integrating these properties makes closed-form solutions inflexible. With the use of the particle filter, the framework permitted the offered algorithm to shape the speech spectral amplitudes as an autoregressive procedure with Laplace distributed excitation. All of the particle sampling distributions were forced to take the dimension into consideration, enabling enhancement in the sampling proficiency.

In tests made by using wideband speech and real recorded noise, the projected algorithm variants were observed to present natural-sounding output speech, with objective assessment outcomes superior to those of the current particle filter speech enrichment algorithms.

The multiple model variant demonstrated superior efficiency in inter-speech noise decline, while the phase variant enhanced execution when the signal-to-noise ratio was minimal.

Van den Bogaert T *et al.* [6] valiantly brought to light speech enhancement with multichannel Wiener filter techniques in multi-microphone binaural hearing aids. They assessed speech improvement in binaural multi-microphone hearing aids by noise lessening algorithms based on the multichannel Wiener filter MWF and the MWF with partial noise estimate MWF-N. Both algorithms were specially designed to merge noise diminution with the maintenance of binaural cues. Objective and perceptual estimates were made with diverse speech-in-multi-talker-babble configurations in two dissimilar acoustic surroundings. When the partial noise estimate was supplemented to the MWF to perk up the spatial understanding of the hearing aid user, there occurred limited decrease in the amount of speech improvement. In certain situations the MWF-N was found to outclass the MWF in performance probably on account of improvement of spatial release from masking.

Lollmann, H.W. *et al.* [7] have efficiently presented a blind speech augmentation algorithm for the repression of late echo and noise.

They put forward a speech improvement algorithm for the repression of background noise and late echo without a priori knowledge. A comprehensive spectral weighting regulation based on assessment for the spectral divergences of the late reverberant speech and background noise was employed for speech improvement. This enabled repression of speech deformations on account of late room reflections unaware of the entire room impulse reaction.

As opposed to conventional techniques, the entire quantities required were assessed blindly from the reverberant and noisy speech signal. The algorithm had also a minimal signal delay which was significant such as for speech augmentation in mobile communication devices or hearing aids.

III. Noise Suppression Technique in Speech Signals

The speech signals are initially wavelet transformed and then applied threshold technique to have the noise suppressed signal. Wavelet de-noising has emerged as an effective method requiring no complex treatment of the noisy signal. It is due to the sparsity, locality and multi-resolution nature of the wavelet transform.

The wavelet transform localizes the most important spatial and frequently features of a regular signal in a limited number of wavelet coefficients. Moreover, the orthogonal transform of stationary white noise results in stationary white noise. Thus, in the wavelet domain the random noise is spread fairly uniformly among all detail coefficients.

On the other hand, the signal is represented by a small number of non-zero coefficients with relatively larger values. This sparsity property assures that wavelet shrinkage can reduce noise effectively while preserving the features of the speech. The process is given in Fig. 1.

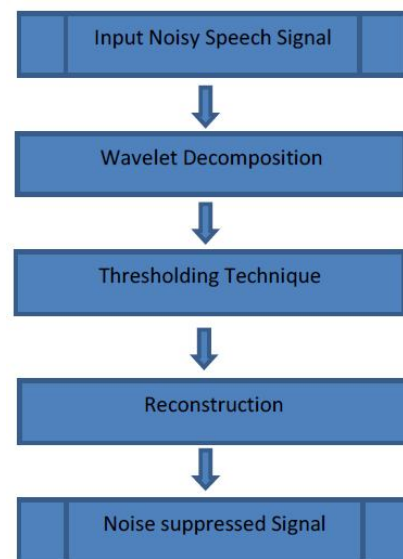


Fig. 1. Noise suppression system architecture

The wavelet denoising technique consists of three phases, namely decomposition phase, thresholding phase and reconstruction phase.

In the decomposition phase, wavelet transform is applied to the noisy speech signals.

Soft or hard thresholding techniques are applied in the thresholding phase and in the reconstruction phase, approximation of coefficients is carried out to have the denoised speech signal.

III.1. Wavelet Decomposition Phase

The input noisy speech signal is initially decomposed with the use of wavelet transformation and normally Discrete Wavelet Transform (DWT) is employed for the purpose. Wavelet series is an illustration of a square-integral function by a specific ortho-normal series created by a wavelet and Discrete Wavelet Transform (DWT) is any wavelet transform for which the wavelets are discretely sampled.

Use of DWT reduces the computation time interval, in addition to decreasing resources required and is very easy to execute. It is competent to decompose a signal at diverse resolutions, which enables the observation of high-frequency events of little duration in non-stationary signals.

The continuous WT of a signal is defined as:

$$Wf(c, d) = \int_{-\infty}^{\infty} f(t) \cdot \psi_{c,d}(t) \cdot dt$$

where:

$$\psi_{c,d}(t) = \frac{1}{\sqrt{c}} \psi^* \left(\frac{t-d}{c} \right)$$

Here, ψ^* denotes complex conjugate of ψ , c a scale factor and d a translation factor. $\psi_{c,d}(t)$ represents a shifted and scaled version of the so-called mother wavelet ψ , which is a window function that defines the basis for the wavelet transformation. A mother wavelet ψ of order mo satisfies the following four properties:

- (1) When $mo > 1$, ψ is $mo-1$ times differentiable?
- (2) $\psi^{(j)} \in L^{\infty}(R)$, for $j = \{1, \dots, mo-1\}$.
- (3) ψ and all its derivatives up to order $mo-1$ decay rapidly.

$$(4) \int t \psi(t) dt = 0, \text{ for } j = \{0, \dots, mo\}.$$

The DWT of a discrete function $fun(n)$ is given by:

$$Wf(c, d) = \sum_n fun(n) \cdot \psi_{j,k}(n)$$

where, $\psi_{j,k}(n) = 2^{-j/2} \psi(2^{-j}n - k)$ and $c = 2^j, d = k2^j$.

In DWT, several frequency bands along with diverse resolutions are employed with a view to convert the signal to coarse approximation and detail data. When the signals are passed through the low pass filter, high frequency components are done away with and the approximation components are benefitted.

The scale does not undergo any change and the resolution becomes half of the original resolution.

Thereafter, 50% of the surplus samples are abolished through sub sampling in which it does not affect the resolution which gets doubled, but influence the scale. Similarly, the detail coefficients are obtained by passing the signal through the high pass filter.

The values are multiplied again with the low pass and high pass filter coefficients to get the LL, LH, HL and HH bands. Fig. 2 gives the block diagram to compute the wavelet coefficients and in general, the wavelet transform can be calculated by the formulas:

$$Wf_{Low}[k] = \sum_n Z[n] H[k-n]$$

$$Wf_{HIGH}[k] = \sum_n Z[n] L[k-n]$$

where, Z is the input, H the high filter function and L , the low pass filter function? Block diagram of wavelet transform is given in Fig. 2.

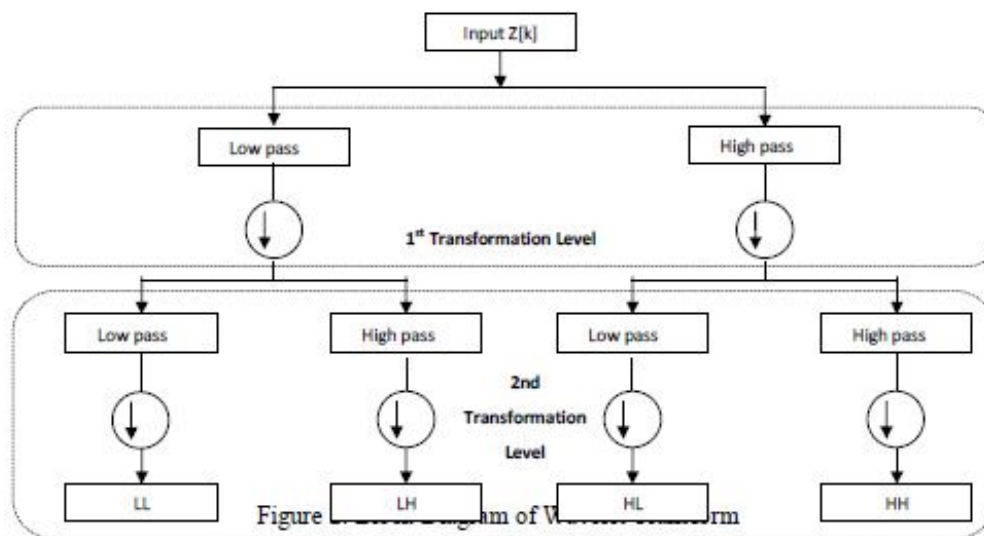


Figure 2. Block Diagram of Wavelet Transform

The process of wavelet decomposition is carried out for Le number of levels the wavelet in order to result in coefficients $c_n^{(Le)}$.

III.2. Thresholding Phase and Reconstruction Phase

After the decomposition of speech wave using wavelet transform, we apply thresholding technique to have the de-noised signal. When speech signals are DWT transformed, the white noise present in the input signal is transformed to white noise with same variance.

And therefore, the coefficients are unchanged or minimized in accordance with the desired signal. It is assumed here that the wavelet transform of smooth functions is such that majority of the coefficients are nearly zero, and white noise is transformed to white noise. Thus, it can be assumed that small coefficients represents the noise and can be set to zero, whereas the large coefficients are retained contributing to original speech signal. It is also possible coefficients having smaller values can also constitute to edge information, hence there is arises a need to have good thresholding techniques to distinguish between noise and original speech signal. There exist many types of thresholding techniques which are mostly based on the threshold definition established in Donogo's universal theory [25].

There exist prefixed thresholds as well as non-prefixed thresholds. Examples of predefined thresholds are universal threshold, universal threshold level dependent universal modified threshold level dependent and minimax threshold. Examples of non-prefixed threshold include maxcoef threshold and Kurtosis and ECE-based thresholds.

Let the signal length be represented by Ns and the level of decomposition be represented by i . Let the coefficient length of i^{th} level be represented by ns_i and let the coefficient be represented by c_i . Commonly used thresholds [26] are:

Universal Threshold:

$$Th_{uni} = \sqrt{2 \log Ns}$$

Universal Threshold Level Dependent:

$$Th_{uni,i} = \sqrt{2 \log ns_i}$$

Universal Modified Threshold Level Dependent:

$$Th_{uni,mod,i} = \frac{\sqrt{2 \log ns_i}}{\sqrt{ns_i}}$$

Minimax Threshold:

$$Th_{minimax} = 0.3936 + 0.1829 \times \left(\frac{\log(ns_i)}{\log(2)} \right)$$

Maxcoef threshold:

$$Th_{maxcoef} = 2^{d-i}$$

where:

$$d = \text{round} \left[\log_2 \left(\max \{ |c_i| \} \right) \right]$$

Kurtosis and ECE-based thresholds:

$$Th_{DF,i} = \frac{1}{\varepsilon_i} \times \frac{\max(c_i)}{F_{iK}}$$

$$\varepsilon_i = \frac{En_i}{En_i}$$

where, En_i is the energy of the i^{th} level, ε_i is the energy contribution efficiency and F_{iK} is the ratio between the Kurtosis value of the signal at i^{th} level to the Kurtosis value of Gaussian noise. There are also predefined Matlab functions for threshold such as *rigsure* (adaptive threshold selection using principle of Stein's Unbiased Risk Estimate) and *heursure* (heuristic adaptive threshold selection using principle of Stein's Unbiased Risk Estimate). Different kinds of thresholds employed give different results based on the wavelet and level of decomposition.

After thresholding, reconstruction is carried out to have the denoised speech signal. The wavelet reconstruction is based on the zeroing approximations which is computed to obtain the bandwidth corrected and denoised signal. The reconstruction formula (Eq. (1) for soft thresholding and Eq. (2) for hard computing) is given by:

$$\hat{b}_n^{(i)} = \begin{cases} \text{sign}(b_n^{(j)}) (|b_n^{(i)} - th_i|) & \text{if } |b_n^{(i)}| \geq th_i \\ 0 & \text{if } |b_n^{(i)}| < th_i \end{cases} \quad (1)$$

$$\hat{b}_n^{(i)} = \begin{cases} b_n^{(j)} & \text{if } |b_n^{(i)}| \geq th_i \\ 0 & \text{if } |b_n^{(i)}| < th_i \end{cases} \quad (2)$$

Here, $\hat{b}_n^{(i)}$ represents the modified i^{th} level coefficient values based on the selected threshold th_i and forms the approximation of the denoised signal.

IV. Proposed Noise Suppression Technique Using Wavelets and Optimization

In this paper, we make two contributions to the existing wavelet thresholding techniques by incorporating adaptive coefficient process and optimal wavelet, level and thresholding techniques.

Contribution 1: Adaptive Coefficient Process

After taking wavelet transform, the coefficients are adaptively changed and the process is discussed in this section. After DWT, we get two sets of coefficients namely detailed coefficients (represented by $d^{(i)}$) and approximation coefficients (represented by $x^{(i)}$). In our technique, we select anyone of these two set of coefficients based on certain criteria. For this we find entropy of both the set. Entropy (En) is a measure of unpredictability or information content and is given by:

$$En = -\sum_j Pr_j \log_2 Pr_j$$

In the above formula, Pr_j is the probability that the difference between 2 adjacent coefficients is equal to j , and \log_2 is the base two logarithms. Let the entropy of detailed coefficients be represented by $D^{(i)}$ and that of approximation coefficients be represented by $X^{(i)}$, where $D^{(i)} = En(d^{(i)})$ and $X^{(i)} = En(x^{(i)})$. Selection of the set of coefficients is based on this found out entropy value and two conditions are given by:

if $(D^{(i)} > X^{(i)})$, Select detailed coefficients set

if $(D^{(i)} \leq X^{(i)})$, Select approximations coefficients set

Hence, the coefficient set (either detailed coefficient set or approximation coefficient set) is selected adaptively based on the entropy measure.

Contribution 2: Optimal Selection of Wavelet and Threshold Using Cuckoo Search

There exist many wavelet types, levels and thresholding techniques and the results vary based on these factors. That is, selection of wavelet, level and thresholding technique has got vital importance and has a large impact on de-noised speech signal obtained. Hence, finding the optimal wavelet, level and thresholding technique for the input noisy speech signal can improve the de-noised speech signal. This would also lead to having good noise suppression.

The selection of wavelet type to use for the discrete wavelet analysis is an important decision for noise removal. The wavelets are of two types: the orthogonal wavelets as Haar, Daubechies (dbxx), Coiflets (coifx), Meyes, Symlets (symx), and biorthogonal (biorx x), where x indicates the order of the wavelet and the higher the order, the smoother the wavelet. The orthogonal wavelets are not redundant and are suitable for signal or image de-noising and compression. In our case, we consider 102 types of wavelet consisting of db1' to 'db45', 'coif1' to 'coif5', 'sym2' to 'sym23', 'dmey', 'bior1.1' to

'bior6.8', 'rbio1.1' to 'rbio5.5'.

The level of decomposition is also important and in our case we consider ten levels. We consider eight thresholding techniques as discussed which are universal threshold, universal threshold level dependent universal modified threshold level dependent, minimax threshold, maxcoef threshold, Kurtosis and ECE-based thresholds, rigrsure (adaptive threshold selection using principle of Stein's Unbiased Risk Estimate) and heursure (heuristic adaptive threshold selection using principle of Stein's Unbiased Risk Estimate).

Hence, we have to find the optimal wavelet within the 102 wavelets, optimal level within ten levels and optimal threshold within eight thresholds. For this, we employ cuckoo search optimization algorithm.

The block diagram of the proposed technique is given in Figure 3. Initially, the upper bounds and lower bounds are set and we consider level (le), wavelet (wa) and threshold (th) to form the factor matrix $FM = [le_i wa_j th_k]$. The lower bound is set with $i=1, j=1$ and $k=1$ to form $FM_{Low} = [le_1 wa_1 th_1]$ and the upper bound is set with $i=10, j=103$ and $k=8$ to form $FM_{Upp} = [le_{10} wa_{103} th_8]$.

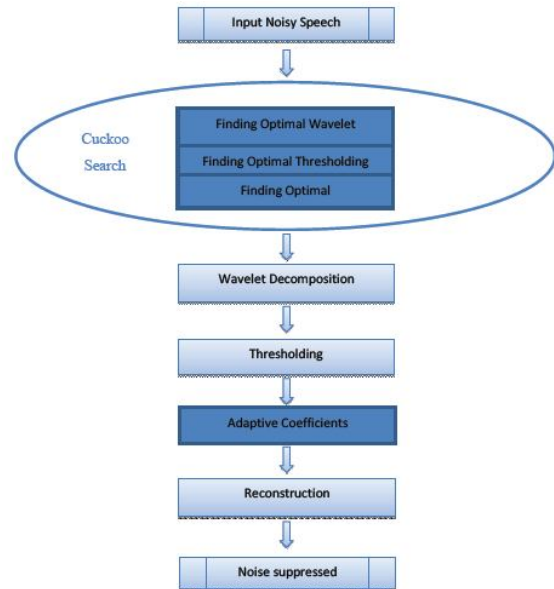


Fig. 3. Block Diagram of proposed technique

Hence the value of i always lie in the range of 1 to 10, j in the range of 1 to 103 and k in the range 1 to 8.

Hence, we employ cuckoo search to find $FM_{Optimal} = [le_i^{optimal} wa_j^{optimal} th_k^{optimal}]$ based on the input noisy speech signal. Cuckoo search (CS)[28] is one of the newest optimization algorithms and has progressed from the motivation that the obligate brood parasitism of some cuckoo variety puts their eggs in the nests of other crowd birds belonging to different species.

Based on the necessitated brood parasitic performance of some cuckoo variety in mixture with the Levy flight performance of some birds and fruit flies, the algorithm is applied.

In Cuckoo Search, three idealized rules are regarded which state that each cuckoo puts one egg at a time, and deposits its egg in an arbitrarily selected nest.

The subsequent rule declares that the finest nests with a superior class of eggs will be taken over to the subsequent generations and the third one says that the number of accessible crowd nests is set and the egg laid by a cuckoo is found out by the host bird with a possibility in the range of 0 to 1. The crowd bird can either heave the egg away or throw out the nest, and make a totally novel nest in this case. Moreover, it is believed that a specified fraction of the nests is substituted by novel nests. The superiority or condition of a solution can merely be comparative to the importance of the objective function for a maximization problem. Cuckoo search is applied to find out optimal wavelet, level and thresholding technique for the input noisy speech signal so as to have the best de-noised signal.

Initially, let the input speech signal be separated into N_{cha} channels where each channel stands for a host nest containing definite number of signals, where each signal symbolizes an egg in the nest. Each egg in the nest is regarded as a novel solution and the final plan of the algorithm is to contain novel and enhanced solutions and to take away the not so fine solutions. Let the host nests be symbolized by $HN = \{hn_1, hn_2, \dots, hn_{N_{cha}}\}$.

Let the signals inside the host nest be represented as $Z_i = \{z_{i1}, z_{i2}, \dots, z_{iN_h}\}$, where N_h is the number of signals in the i^{th} host nest. Then, Levi flight is carried out on Z_i to yield to get a new cuckoo Z_i^* . Suppose the signal z_{i1} in Z_i , and then the updated value is given by:

$$z_{i1}^* = z_{i1}^{(t+1)} = z_{i1}^{(t)} + \Delta \otimes Levy(x)$$

At this point Δ is the step size which is greater than zero and is occupied as one. The Levi flight equation signifies the stochastic equation for arbitrary walk as it depends on the present position and the changeover possibility (second term in the equation). At this point, the levy sharing is specified by:

$$Levy(x) = \sqrt{\frac{c}{2\pi}} \cdot \frac{e^{-\frac{1}{2}(\frac{c}{x})}}{x^{3/2}}$$

where c is arbitrary constant. Therefore, by executing Levi search, we get novel solutions and after that the fitness value (SNR value) of the new solution is discovered. Let the fitness of the Levi achieved nest be Fit_i .

Subsequently, some other nest is regarded as other than the i^{th} host nest and let the nest in deliberation be signified by $Z_j = \{z_{j1}, z_{j2}, \dots, z_{jN_h}\}$ representing j^{th} host nest.

The vigour of the j^{th} nest is defined by means of the fitness function and is symbolized by Fit_j . If the fitness of the Levy flight made i^{th} nest Fit_i is greater than fitness of the j^{th} nest Fit_j , then substitute j^{th} nest signal values $Z_j = \{z_{j1}, z_{j2}, \dots, z_{jN_h}\}$ by the i^{th} host nest Levy performed values $Z_i^* = \{z_{i1}^*, z_{i2}^*, \dots, z_{iN_h}^*\}$.

The course is performed for all the host nests h_i where each of the nests are at first Levi flight performed, related fitness is found out Fit_i , compared to fitness of some other nest Fit_j and the substitution is performed if the condition $Fit_i > Fit_j$ is satisfied.

We have to discard a fraction of the most horrible nests and build novel nests in place of them after comparison and substitution.

By finding the class of all the present nests and examining it, this is done. At this point the fraction of nests to be discarded is set as Fra , so that out of the total nests N_{cha} , number of abandoned nets is $round(N_{cha} - (Fra \times N_{cha}))$ and therefore, it is identical to the similar number of nests to be built. Thus, we keep the most excellent solutions and substitute the worst nests by recently built nests.

Then the solutions are graded and the present best is discovered. The full circle is carried on till some stop criteria is met and the present top in the last circle carried out becomes the top solution.

Hence from cuckoo search, we obtain the $FM_{Optimal} = [le_i^{optimal} wa_j^{optimal} th_k^{optimal}]$ based on the input noisy speech signal.

Thus the optimal levels of decomposition for wavelet and threshold are obtained for the respective input speech signals.

Subsequently, the wavelet transformation is carried out with the found out optimal wavelet and with the found out optimal number of decomposition levels. After that, adaptive coefficient process and thresholding are performed with the found out optimal thresholding which is then reconstructed to obtain the de-noised speech signal.

V. Results and Discussions

In this section, the results obtained for the proposed technique are discussed and analyzed. In section 5.1, experimental set up and evaluation metrics employed are discussed and in section 5.2, simulation results are given. Section 5.3 gives the comparative analysis.

V.1. Experimental Setup and Evaluation Metrics

The proposed technique is implemented in MATLAB on a system having 6 GB RAM and 2.6 GHz Intel i-7 processor. The evaluation metrics employed are IS and MSE. *Itakura–Saito distance (IS)* [27] is a measure of the perceptual difference between an original spectrum and an approximation of that spectrum. Here let the original spectrum be $X(w)$ and the estimated be $\bar{X}(w)$.

So we can calculate the distance D_{IS} by the formula given in the below equation:

$$D_{IS}(X(w), \bar{X}(w)) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{X(w)}{\bar{X}(w)} - \log \frac{X(w)}{\bar{X}(w)} - 1 \right] dw$$

Mean Squared Error (MSE) quantifies the difference between values implied by an estimator and the true values of the quantity being estimated. MSE measures the average of the squares of the errors. Let estimate vector be represented by \hat{S} and the true vector be represented by S , then MSE is defined by:

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{S}_i - S_i)^2$$

V.2. Simulation Results

The simulation results are given in this section. Fig. 4 gives the clean signal, Fig. 5 gives the noise signal, Fig. 6 gives the normal de-noised signal and Fig. 7 gives the de-noised signal using cuckoo search. In each figure, the time graph is given in top and corresponding spectrum is given below. Analyzing Figs. 6 and 7, we can see that optimization using cuckoo search has obtained better noise denoising than conventional denoising.

V.3. Comparative Analysis

In this section, comparative analysis is carried out comparing the proposed technique with the existing technique. The graphs are plotted for IS and MSE for varying noise conditions.

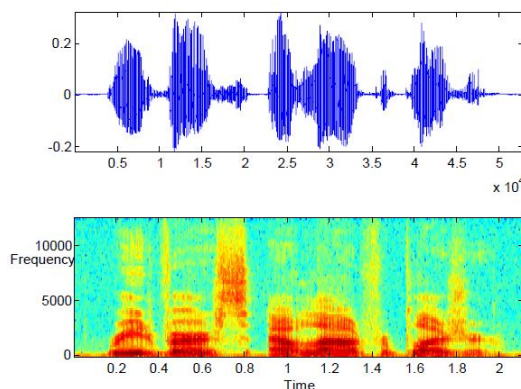


Fig. 4. Clean signal

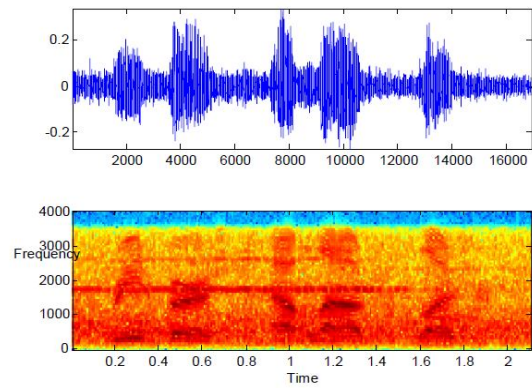


Fig. 5. Noise signal

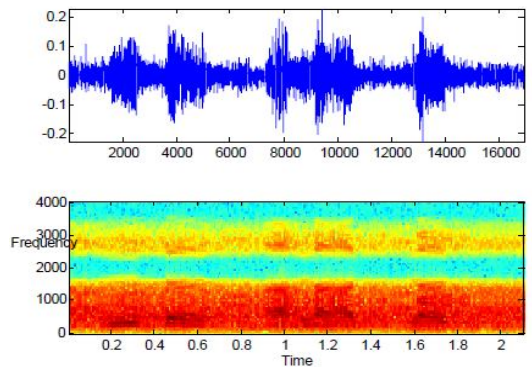


Fig. 6. Normal denoised signal

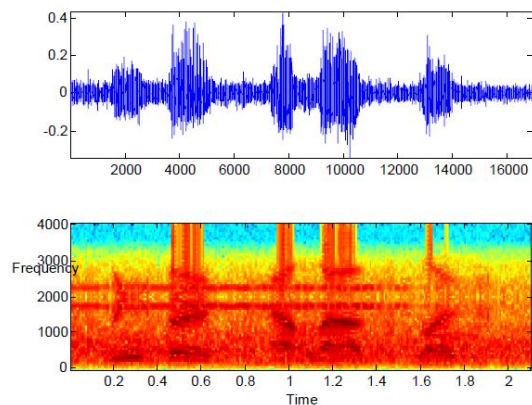


Fig. 7. Denoised signal using cuckoo search optimization

Inferences from Figure 8 to 13

- The graphs give comparative analysis where the proposed technique is compared with the existing technique.
- The graphs are plotted for IS and MSE for varying noise conditions such as airport noise, car noise and babble noise.
- Figs. 8, 10 and 12 give IS graphs for airport noise, car noise and babble noise respectively.
- Figs. 9, 11 and 13 give MSE graphs for airport noise, car noise and babble noise respectively.
- In all graphs, values for noise levels 0, 5dB, 10dB and 15 dB are taken.

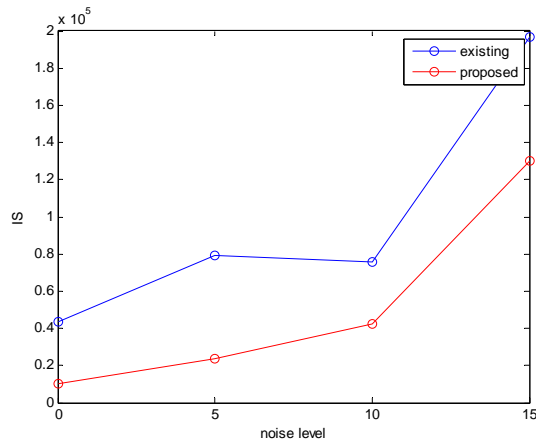


Fig. 8. IS considering airport noise

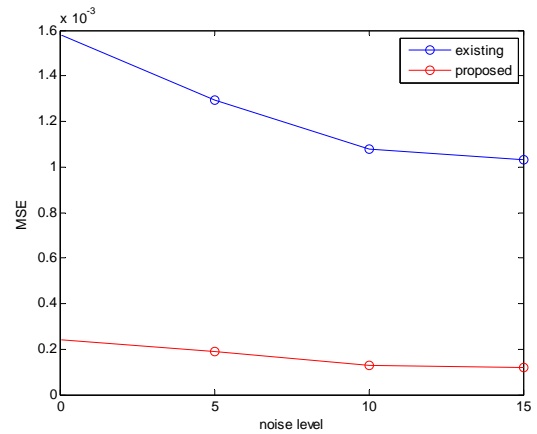


Fig. 9. MSE considering airport noise

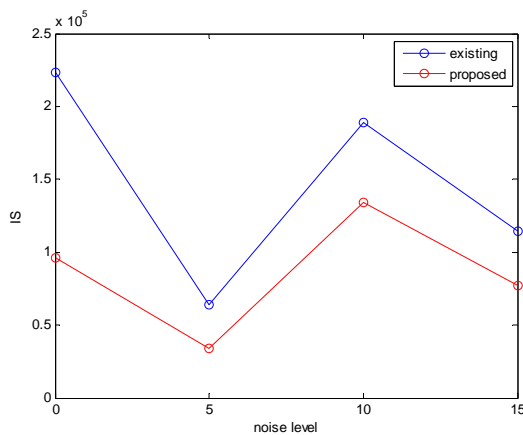


Fig. 10. IS considering car noise

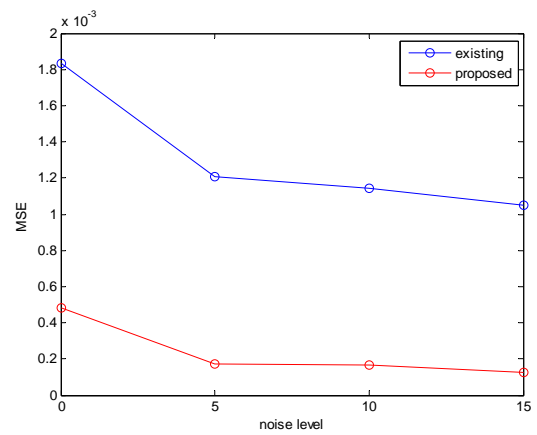


Fig. 11. MSE considering car noise

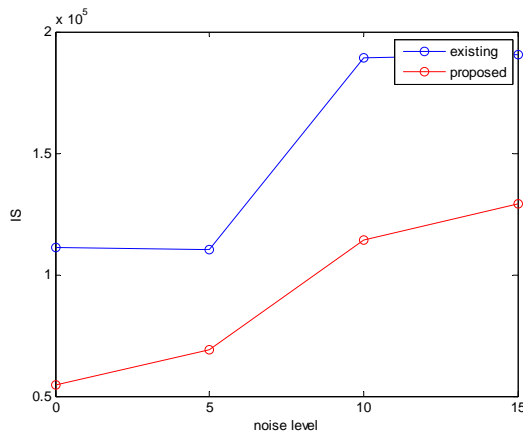


Fig. 12. IS considering babble noise

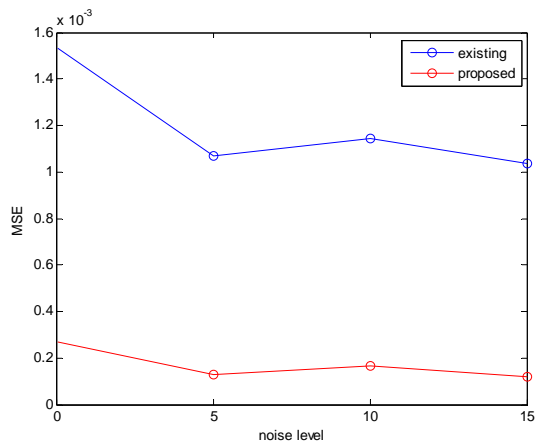


Fig. 13. MSE considering babble noise

- We can infer from the graphs that our proposed technique has achieved lower noise levels at all condition which shows the efficiency of the proposed technique.
- Average values obtained for IS and MSE are calculated for different noise conditions and are given in Table I.

- We can see that average IS and MSE values for proposed is far lower than existing technique.
- Total IS average for proposed was about 0.78×10^5 compared with 1.3×10^5 that of existing.
- Total MSE average for proposed was about 0.22×10^{-3} compared with 1.25×10^{-3} that of existing.
- The total average IS and MSE values are given in Figure 14.

- The obtained lower IS and MSE values indicate the effective functioning of the proposed technique.

TABLE I
AVERAGE IS AND MSE VALUES

	Average IS ($\times 10^5$)		Average MSE ($\times 10^{-3}$)	
	Previous	Proposed	Previous	Proposed
<i>Airport Noise</i>	0.98	0.51	1.26	0.19
<i>Car Noise</i>	1.52	0.89	1.32	0.26
<i>Babble Noise</i>	1.51	0.92	1.18	0.20

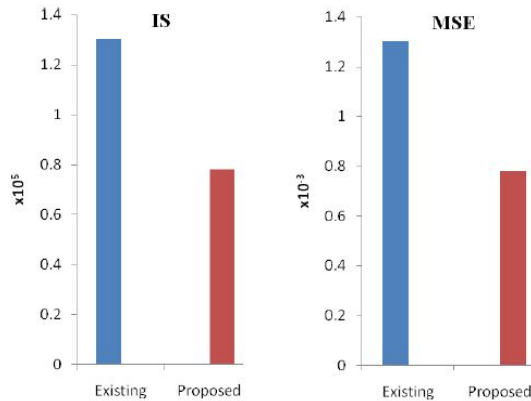


Fig. 14. The total average IS and MSE

VI. Conclusion

Optimal selection of wavelet, level and thresholding technique for noise suppression in speech signals using Cuckoo search is presented in this paper. After finding the optimal values, the signal is wavelet transformed using optimal wavelet which is then done adaptive coefficient process.

Afterwards optimal thresholding technique is carried out and finally, reconstruction is performed to have the noise suppressed signal. The implementation is done with MATLAB and the evaluation metrics of Itakura-Saito distance (IS) and MSE are used. Various noise conditions are considered while taking results and comparison is also made with existing technique. We can also find IS and MSE values for proposed is far lower than existing technique. Total IS average for proposed was about 0.78×10^5 compared with 1.3×10^5 that of existing. Total MSE average for proposed was about 0.22×10^{-3} compared with 1.25×10^{-3} that of existing. The results show the effectiveness of the proposed technique.

References

- [1] E. Castillo, D. P. Morales, A. García, F. Martínez-Martí, L. Parrilla, and A. J. Palma, "Noise Suppression in ECG Signals through Efficient One-Step Wavelet Processing Techniques", *Journal of Applied Mathematics*, pp. 1-13, 2013.
- [2] Li Ruwei; Bao Changchun; Xia Bingyin and Jia Maoshen, "Speech enhancement using the combination of adaptive wavelet threshold and spectral subtraction based on wavelet packet decomposition", *proceedings of International Conference on Signal Processing (ICSP)*, Vol. 1, pp. 481 - 484, 2012.
- [3] Brady Laska, Miodrag Bolic, Rafik Goubran, "Discrete cosine transform particle filter speech enhancement", *Speech Communication*, vol. 52, pp. 762-775, 2010.
- [4] Slavy G. Mihov, Ratcho M. Ivanov and Angel N. Popov, "Denoising Speech Signals by Wavelet Transform", *Annual Journal Of Electronics*, 2009.
- [5] Brady N. M. Laska, Miodrag Bolic, and Rafik A. Goubran, "Particle Filter Enhancement of Speech Spectral Amplitudes", *IEEE transactions on audio, speech, and language processing*, Vol. 18, No. 8, November 2010.
- [6] Van den Bogaert T, Doclo S, Wouters J and Moonen M, "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids", *J Acoust Soc Am.*, vol. 125, no. 1, pp. 360-371, 2009.
- [7] Lollmann, H.W. and Vary, P, "A blind speech enhancement algorithm for the suppression of late reverberation and noise", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3989-3992, 2009.
- [8] Jamal Ghasemi and Mohammad Reza Karami Mollaei, "A New Approach for Speech Enhancement Based On Eigenvalue Spectral Subtraction", *Signal Processing: An International Journal*, Vol. 3, no. 4, 2009.
- [9] J. R. Deller, J. H. L. Hansen, J. G. Proakis, "Discrete Time Processing of Speech Signals", second ed. IEEE Press, New York, 2000.
- [10] Van den Bogaert T, Doclo S, Wouters J and Moonen M, "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids", *J Acoust Soc Am.*, vol. 125, no. 1, pp. 360-371, 2009.
- [11] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109-1121, 1984.
- [12] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 153-156, 2006.
- [13] Achintya Kundu, Saikat Chatterjee, A. Sreenivasa Murthy and T.V. Sreenivas, "GMM Based Bayesian Approach To Speech Enhancement In Signal / Transform Domain", In *33rd IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.
- [14] Yi Hu and Philipos C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement", *IEEE transactions on audio, speech and language processing*, vol. 16, no. 1, 2008.
- [15] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction". *IEEE Transactions Acoustics Speech Signal Process*, no. 27, pp. 113-120, 1979.
- [16] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator". *IEEE Transactions Acoust. Speech Signal Process. ASSP*, vol. 32, no. 6, pp. 1109-1121, 1984.
- [17] D.L. Donoho, "Denoising by soft thresholding". *IEEE Transactions Information Theory*, 41(3), 613:627, 1995.
- [18] S. G. Mallat, "Theory of multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674-693, 1989.
- [19] E.-B. Lin and P. C. Liu, "A discrete wavelet analysis of freak waves in the ocean," *Journal of Applied Mathematics*, vol. 2004, no. 5, pp. 379-394, 2004.
- [20] R. Sameni and G. D. Clifford, "A review of fetal ECG signal processing, issues and promising directions," *The Open Pacing, Electrophysiology & Therapy Journal*, vol. 3, no. 1, pp. 4-20, 2010.
- [21] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200-1224, 1995.
- [22] B. N. Singh and A. K. Tiwari, "Optimal selection of wavelet basis function applied to ECG signal denoising," *Digital Signal Processing*, vol. 16, no. 3, pp. 275-287, 2006.
- [23] L. N. Sharma, S. Dandapat, and A. Mahanta, "ECG signal denoising using higher order statistics in Wavelet subbands," *Biomedical Signal Processing and Control*, vol. 5, no. 3, pp. 214-222, 2010.
- [24] Inc. The MathWorks, Denoising: Wavelet shrinkage, block thresholding, multisignal thresholding, 2013: <http://www.mathworks.es/es/help/wavelet/denoising.html>.

- [25] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.
- [26] B. N. Singh and A. K. Tiwari, "Optimal selection of wavelet basis function applied to ECG signal denoising," *Digital Signal Processing*, vol. 16, no. 3, pp. 275–287, 2006.
- [27] Alan H. S. Chan, Sio-Iong Ao, "Advances in industrial engineering and operations research" Springer, Page 51, 2008.
- [28] X.-S. Yang, S. Deb, "Cuckoo search via Levy flights", in: *Proc. Of World Congress on Nature & Biologically Inspired Computing IEEE publications*, pp. 210-214, 2009.
- [29] Logeshwari, G., Anandha Mala, G.S., An efficient speaker recognition system for separating the single channel speech using frequency modulation, (2013) *International Review on Computers and Software (IRECOS)*, 8 (2), pp. 632-641.
- [30] Z. Sakka, A. Kachouri, M. Samet, Speech Denoising and Arabic Speaker Recognition System Using Subband Approach, (2007) *International Review on Computers and Software (IRECOS)*, 2 (3), pp. 264-271.
- [31] N. Seman, J. Kamaruzaman, Performance of Adapting Non-Native Speech in Isolated Speech Recognizer, (2008) *International Review on Computers and Software (IRECOS)*, 3 (3), pp. 324-328.
- [32] Djebbari, A., Bereksi-Reguig, F., A new chirp-Based wavelet for heart sounds time-Frequency analysis, (2011) *International Journal on Communications Antenna and Propagation (IRECAP)*, 1 (1), pp. 92-102.
- [33] Manasra, G., Najajri, O., Rabah, S., Arram, H.A., DWT based on OFDM multicarrier modulation using multiple input output antennas system, (2012) *International Journal on Communications Antenna and Propagation (IRECAP)*, 2 (5), pp. 312-320.

Authors' information



H. Kaur (Harjeet Kaur) obtained her Bachelor's degree in Electronics & Communication from Agra University. Then she obtained her Master's degree in Electronics & Communication and PhD (pursuing) in Electronics in Signal Processing from Punjab Technical University, Jalandhar, Punjab, India. Currently, she is a Assistant Professor at the Indira College of Engg. & Mgmt. Pune. Her specializations include Digital Signal Processing, Adaptive Digital Signal Processing, Speech Processing.



Dr. R. Talwar (Dr. Rajneesh Talwar) received the B.tech. Degree in 2001 from Aurangabad University, the M.tech. Degree from Thapar University, in 2002, and the Ph.D. Degree from the Thapar University in 2010. He has an Professional experience of more than ten years in teaching and research. He is currently work as a Principal at Chandigarh Group of Colleges, Landran. His research interests are fiber optics, semiconductor devices, communication He has a U.S patent "FIBER OPTIC POINT TEMPERATURE SENSOR" to his credit, He is Editorial board member of International Journal of Engg. Science and Technology, Nigeria, was Reviewer of MAEJO International Journal of Engineering Science and Technology, Thailand and "Materials and Design", a ELSEVIER International Journal.

Survey and Analysis of Visual Secret Sharing Techniques

L. Jani Anbarasi¹, G. S. Anandha Mala²

Abstract – Security is an important issue in information technology, which is ruling the internet world today. The aim of this paper is to present an overview of the emerging techniques for secret Sharing. A (t, n) secret sharing scheme refers to a method of distributing a secret among a group of n participants, each of whom is allocated with a meaningless share of the secret. When t or more participants pool their shares together the secret can be reconstructed, but less than that cannot. Secret sharing is widely used due to the remarkable growth in security awareness by individuals, groups, agencies etc. This paper provides a state-of-the-art review and analysis of the different existing methods of secret sharing, along with some common standard algorithms and guidelines drawn from the literature, and concludes with an analysis of the various functionalities of the different techniques. **Copyright** © 2014 Praise Worthy Prize S.r.l. - All rights reserved.

Keywords: Threshold Secret Sharing, Steganography, Image processing, Cryptography

I. Introduction

Recent technological advances in computer networks have turned the transmission of digital data into a popular task, and digital images are no exception. Security has become an inseparable issue, as information technology is ruling the world now. The necessity of transferring confidential images over open channels such as the internet leads to security threats. Visual secret sharing is a process of encryption procedure which encrypts a secret image into 'n' cipher images called shadows, which can be transmitted or distributed over an untrusted communication channel. The dealer distributes the secret shadows to n participants, and the shares of t participants are pooled to retrieve the secret.

Any secret image sharing mechanism must comply with the following essentials:

- (i). The participants involved must be able to detect cheats.
- (ii). At least t authorized participants can cooperate to reconstruct the secret image.
- (iii). The retrieved secret image must be lossless.
- (iv). The shadow image must be meaningful.
- (v). The distortion of the shadow image must be slight.
- (vi). The size of the embedded secret image must be large enough.

Secret sharing has various techniques such as the Shamir method [1] which uses the Lagrange Interpolation polynomial, Hyperplane interpolation in the Blakely method, and the Mignotte and Asmuth bloom, based on the Chinese Remainder Theorem.

These techniques securely encrypt a secret into shadows and retrieve it by the reverse process of the same algorithm. Noar proposed Visual Cryptography in 1994; this encrypts a secret into n meaningless shares (transparencies), and the secret is retrieved by stacking these transparencies.

A codebook S_0 and S_1 is generated for both 0 and 1, using which the secrets are shared among the participants:

$$S_0 = \left\{ \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \right\} \quad S_1 = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}$$















Pixel	White 		Black 	
Prob.	50%	50%	50%	50%
Share 1				
Share 2				
Stack share 1 & 2				

Fig. 1. Basic Concept of Visual Cryptography

II. Secret Sharing Techniques

II.1. Visual Secret Sharing

In the year 1979, Shamir proposed a (t, n) scheme for dividing a data D [1] into n pieces in such a way, that D is easily reconstructable from any t pieces, but even a complete knowledge of $(t - 1)$ piece reveals absolutely no information about the data D . Shamir's approach uses a secret S and a prime number m to generate a $(t-1)^{\text{th}}$ degree polynomial, which is given below:

$$F(x) = S + C_1X^1 + \dots + C_{t-1}X^{t-1} \text{ mod } m \quad (1)$$

The Coefficients $C_1, C_2 \dots C_{t-1}$ are random integers within the range $[0, m-1]$. $Y_1=F(K_1)$ $Y_2=F(K_2)$... $Y_n=F(K_n)$ where Y_i ($1 \leq i \leq n$) represents the computed

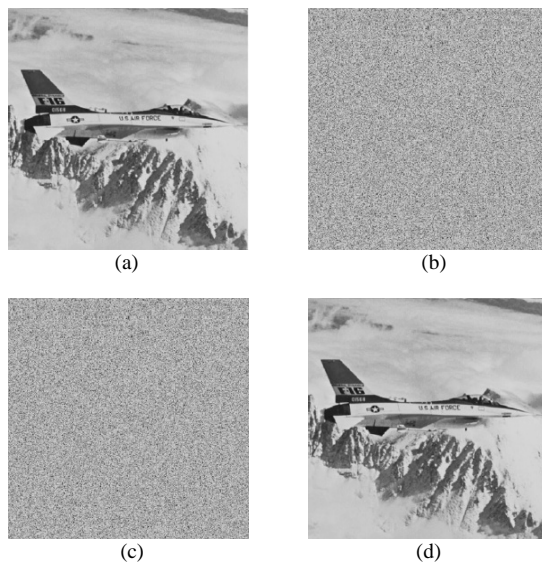
shadow value, which is calculated using the secret key K_i of each participant using (1), and are securely issued to the participants by the dealer.

The secrets are reconstructed using Lagrange's interpolation formula given in Eqs. (2) and (3). This technique enables the construction of robust key management schemes for cryptographic systems:

$$L = \sum_{j=0}^t y_j l_j(x) \quad (2)$$

$$l_j(x) = \prod_{\substack{0 \leq m \leq t \\ m \neq j}} \frac{x - x_m}{x_j - x_m} \quad (3)$$

In recent days, various methods are proposed by researchers using Shamir's secret sharing approach.



Figs. 2. Secret sharing using Shamir Model (a) Secret (b)-(c) Secret shadows (d) Reconstructed secret

The secret image is encrypted into n smaller meaningless shadows by [2] where r meaningless shadow images can reconstruct the secret. The size of each shadow image is smaller than the secret image. Thien and Lin used the image pixels as co-ordinates in the polynomial, instead of random numbers, because of which the size of the shadow is reduced to a great extent.

In order to prevent the visibility of data the pixels are permuted before secret sharing [3] improved the security of secret sharing by providing three levels of security such as steganography, authentication and verification.

By using the steganographic technique, N shadows are embedded into N camouflage images, in order to avoid the attention of intruders.

The image watermarking technique is used to embed fragile watermark signals into the camouflage images.

The capability of authenticating the veracity of each stego image was tested using parity-bit checking. During the recovery process, the pooled shadows are verified by

inspecting the constancy of the parity conditions before reconstructing the secret image. This technique is employed for the secret greyscale image and can be easily extended to the secret color image. The experimented result shows that the PSNR value of 39.16 DB is achieved.

The host image expands to four times that of the secret image which is the shortcoming of this method. In order to avoid the correlation among the pixels, [4] created a difference image by subtracting the neighbouring pixels, and then using the Huffman encoding to reduce the size of the image. The values of the encoded image are used as a co-efficient of the Shamir polynomial.

The size of the generated shadow images is less than half the size of the secret. The calculations for the sharing function are done in $GF(2^t)$ instead of $GF(p)$, as adapted by Thien and Lin ; this prevents the loss of secret data .[5] proposed a secret sharing using the multi secret mode, the progressive mode and the priority mode, based on the Shamir method, where it uses the $GF(2^8)$ irreducible polynomial prime for solving. In the multiple secret modes, a secret is divided into sub images and then shared. The progressive and priority modes make use of digits, which is a lossless version of secret sharing. This method is restricted to participants holding the same amount of information, and the use of a monochromatic secret image. The secrets are reconstructed in a scalable manner without loss, only if all n shadows are available.

The scalability and flexibility of this scheme shows the wide range of potential applications for secret image sharing [6] extended the general (t, n) scalable secret image sharing, which generates smaller shadow images.

This method reconstructs different versions of the secret image when different combinations are pooled. The size of the scheme is $(2n - t)/n^2$ times of the original image, which outperforms [7] scalable secret image scheme, in which each shadow image is half the size of the original image [8] proposed a scheme with teganography and authentication, based on the Chinese remainder theorem to enhance the quality of the stego images; this is the short coming in [3] and [9]. This scheme improved the methods of [3],[113] and predicted the PSNR values as $PSNR_{Chang} > PSNR_{Lin} > PSNR_{Yang}$ which was further disproved in [10]. [10] compared the PSNR values of the stego images of [3],[9],[8] image secret sharing schemes, and proved that the result of the Chang et al, scheme $PSNR_{Chang} > PSNR_{Lin} > PSNR_{Yang}$ is wrong , and showed that the actual result is $PSNR_{Yang} > PSNR_{Lin} > PSNR_{Chang}$. In order to avoid the secret image from eavesdroppers Chang and Wu's gradual search algorithm for a single bitmap BTC (GSBTC) and Shamir's (t, n) threshold concept were combined, by [4] proposed a new model using a reduced shadow size. The secret is reduced using the GSBTC technique, and is then encoded using Shamir's method.

The experimental results confirm that this scheme successfully reduces the shadow size, and that prevents the leakage of information about the secret color image. The horizontal and vertical correlation between adjacent

pixels is found, and on an average, it has a value of NPCR = 0.414% and AUCI = 32.78%. Thus, a one-pixel difference could cause a significant difference in the corresponding shadows. Therefore, the security of the presented scheme is also confirmed. This is a lossy secret sharing, where the secrets are lost to a certain percentage.

Lin, P Y., et.al., proposed the (t, n) -threshold mechanism, [11] in which the generated shadows [12]-[17] are embedded into the host image, using the modular operator to obtain a meaningful shadow image with satisfactory quality. Moreover, Rabin's signature algorithm is used [18] to detect cheat. The security of the algorithm is the same as that of the factorization problem [19], [20]. Specifically, these methods allow the participants to retrieve a lossless secret image and a distortion-free host image.

Tsai, D S., et.al., combined the neural network and the feed forward network [21] to generate shares for a true color image. Shares are camouflaged into a cover, using the XOR operation during encoding, and with low computation the secret and the cover image are retrieved.

The difference between the restored true image and the original secret is not visually perceptible. Chang, C C., et.al., developed a meaningful secret sharing scheme [22] which includes both authentication and remedial abilities that allow the detection of the corrupted area by using the hidden information. The corrupted secret is repaired using the remedial information to obtain reasonable visual quality. A novel scheme is proposed which can losslessly reconstruct the secret image without additional pixels, regardless of the original pixel. In the proposed scheme, the pixels are shuffled, segmented and transformed into three integer parameters used for creating shares along with the secret information used for the remedial purpose. The secret is reconstructed even if the shares are tampered with but with little loss, whereas other algorithms do not do that. A good PSNR value of more than 3DB is achieved when compared to other research works.

Ulutas, M., et.al., proposed a (t, n) secret sharing scheme, [23] which shares a Medical image and EPR among n clinicians, and reconstructs without loss. The meaningless shares are embedded into the meaningful host images, using the LSB OPAP technique of steganography. To authenticate the shares during reconstruction, authentication certificates are generated for each participant, using the MD5 hash algorithm. The proposed method emphasizes and provides three desired capabilities like confidentiality, authenticity and EPR hiding. Anbarasi, L J., et.al., securely shared a color image using the Shamir model, [24] making use of a modular operation, where meaningful shares are generated using the LSB, and the secrets are reconstructed without loss based on the chosen prime.

In Blakley, G.R., n nonparallel $(n-1)$ -dimensional hyper planes intersect [25] at a specific point. The secret may be encoded as any single coordinate of the point of intersection. Each participant is given a hyper plane passing through the point of intersection.

The secret can be uniquely identified when t or more participants come together, but less than that cannot.

Tso, H.K., round off errors due to quantization. Round off errors [26] work reconstructs lossy secret images with some reduce the quality of the reconstructed image. Ulutas, G., et.al., presents a geometry based distortion free secret image sharing approach [27], which uses a difference image to accommodate and compensate round errors, resulting in a distortion free secret image sharing scheme. Reconstructed secret images are distortion free as required by some applications.

Eslami, Z et.al., proposed a secret sharing scheme which used dynamic embedding authentication chaining [28] that uses 1 to 3 bits of authentication, whereas [29] used a one parity bit, while [30] used the keyed hash function to generate a one bit value, and [8] generated a 4 bit authentication using the CRT. A scalable secret image sharing is proposed by [31] to encrypt a secret into n shares where the information of the reconstructed secret is proportional to the number of shadows used for reconstruction; i.e., if an image is divided into 10 parts, the image is retrieved only if all the parts are joined together.

II.2. Secret Sharing of 3D Model

Elsheh, E., et.al., proposed two secret sharing models for a single 3D model [32], using the [25] and [2] Schemes. Lossless compression algorithms like Huffman encoding and ZLIB techniques were used to reduce the size before secret sharing. This is the first paper to introduce secret sharing for 3D models.

II.3. Secret Sharing for Multiple Secrets

Yang, C C., et.al., proposed a (t, n) multi secret sharing based on Shamir's secret sharing method, which uses the multi linear interpolation polynomial to share p secrets [33]. It is a multiuse scheme which allows a parallel secret reconstruction, and the secret holder can dynamically determine the number of distributed secrets. Secrets chosen for decryption can be less than, equal to, or greater than the threshold.

The linear interpolation polynomial uses a one way function to avoid cheating by both the participants and the dealer. [33] uses $(n + p - t + 1)$ or $(n + 1)$ public values, $2(t - 1)$ or $2(p - 1)$ storages, and employs the Lagrange interpolation polynomial to share p secrets, which improves [34] where $(n + p - t + 1)$ public values, $(2(n + p) - t) * (n + p)$ storages, and solved $(n + p - t)$ simultaneous equations to share p secrets. Jun Shao [35] developed a multi secret sharing model based on the YCH where the secret shadows are verified by only one equation, whereas in the verifiable secret sharing proposed by [36] $n!/(n-t)!t!$ equations are used.

Therefore, the complexity of verification is reduced to a lower level. So no secure channel is needed for transmission of secrets. [37] enhanced the security constraints provided by [38] method, in terms of

computational complexity in verifiable multi secret sharing. The computational complexity is reduced to a lower level and cheating between dealer and participant is avoided to a greater extent.

Zhao J et.al., proposed a practical verifiable multi secret sharing scheme using the intractability of the discrete logarithm[39]. Since each participant chooses a shadow on his own, a security channel is not necessary.

This allows an efficient parallel reconstruction of multiple secrets. Verification is done by both the participants and the dealer. [40] dealt with an efficient and verifiable multi secret sharing scheme based on the Lagrange polynomial and the LSFR based public key cryptosystem.

It uses the homogeneous LFSR, the third order LFSR and the LFSR public key cryptosystems. Since it is a verifiable approach, participant cheating, and conspiracy attacks are avoided with good strength in security. It is a good approach for practical situations, and future work may deal with inserting a new secret, or how an old secret can be deleted without altering the rest.

The merits of the system are its less computation and storage.

II.4. Secret Sharing Based Access Structure

A non-perfect secret sharing scheme is a method to distribute a secret among a set of participants, in such a way that some qualified subsets of participants, pool their information, to reconstruct the secret, whereas, other

subsets of participants may have some information about the secret. [41] proposed a secret sharing based on access structure which constructs a circular share with a smaller size than that obtained by previous research.

The length of the circular shares is found by the recursive algorithm which takes exponential time. The reduction of exponential time to polynomial time was left as an open problem. This problem was analyzed by [42] who proved that finding a polynomial time is not possible; they found a better method where the share size is reduced than [41]. It is a non perfect secret sharing scheme. So this algorithm is not secure.

Chan, C W.,et.al., proposed a threshold multi secret sharing scheme, which packs secrets of different threshold access structures into one such that it still follows the same access structure [43]. More than one secrets are shared among participants where a master shadow and sub shadows are created for each participant.

A generalized polynomial form is generated using the Chinese Remainder Theorem. In this scheme, the dealer generates a master shadow for each participant. Participants pool their sub shadows to generate their secrets but not the master secret.

The optimal information about bounds and the characterization of an ideal structure is studied by [44]

The complete characterization of an ideal access structure with intersection number to one and for non ideal case the bounds for the optimal information rate is analyzed.

TABLE I
FUNCTIONALITY COMPARISONS BETWEEN RELATED VISUAL SECRET SHARING MECHANISMS

Ref.	Meaningful Shares	Threshold scheme	Image Type	Lossless Secret	Perfect	Capacity	Cheater Detection	Lossless host	Shadow size	Quality of shadow(DB)
[2]	No	Yes	Gray	Yes	Yes	-	-	-	Small	-
[9]	Yes	Yes	Gray	No	Yes	$m \times n/4$	Yes	No	Expand	39.31
[3]	Yes	Yes	Gray	No	Yes	$m \times n/4$	Yes	No	Expand	39.16
[4]	No	Yes	Gray	Yes	Yes	-	-	-	Small	-
[28]	Yes	Yes	Gray	No	Yes	$m \times n/4$	Yes	No	Expand	51.94
[22]	Yes	Yes	Gray	No	Yes	$m \times n/4$	Yes	No	Expand	45
[43]	No	No	Gray	No	Yes	-	-	-	Small	-
[10]	Yes	Yes	Gray	No	Yes	$m \times n/4$	Yes	No	Expand	44.62
[8]	Yes	Yes	Gray	Yes	Yes	$m \times n/4$	Yes	No	Expand	40.97
[73]	No	Yes	Color	No	Yes	-	-	-	Small	-
[11]	Yes	Yes	Gray	Yes	Yes	$(m \times n \times (t-3))/4$	Yes	Yes	Expand	43.36
[23]	Yes	Yes	Gray	No	Yes	$m \times n/4$	Yes	No	Expand	46.36
[27]	No	Yes	Gray	Yes	Yes	-	-	-	Small	-
[31]	No	No	Gray	Yes	Yes	-	-	-	Small	-
[6]	No	No	Gray	Yes	Yes	-	-	-	small	-

TABLE II
COMPARISONS OF VARIOUS AUTHENTICATION BITS

Ref	Authentication Bits
[3],[9],[72]	1
28	1 to 3
8	4

TABLE III
COMPARISONS OF STEGANOGRAPHY ALGORITHM

Ref	Steganography Algorithm
[23]	LSB OPAP
[11]	Modular operator
[10],[8],[28],[3],[9]	LSB

The analysis reveals that the ideal access structure with intersection number equal to one coincides with the vector space, and there is no member of this family whose optimal information rate lies between $2/3$ and 1 [45] proved that a secret can be shared among a set of participants in a way that it can be recovered only by a qualified set of participants. Problems like ideal access structure and search for bounds on optimal information rate for 3-homogeneous access structures are addressed; that is, the minimal qualified subsets have exactly three participants [46] proposed a hierarchical threshold secret

image sharing scheme, where the secret image is partitioned into several levels and the threshold is set according to the access structure. The shadows are embedded into a cover image using modular operation. Using the Birkoff interpolation, the secret image is retrieved based on the satisfaction of the threshold property.

III. Secret Sharing Using Random Grids

Kafri, O., et. al., proposed a (2,2) scheme for the secret sharing of two dimensional patterns and shapes based on random grids [47], where the secret is reconstructed by the superimposition of the random grids.

This scheme overcomes the limitations of visual cryptography, such as pixel expansion and code book generation. Secret sharing using random grids was extended to greyscale and color images in [48], [49] where shares do not leak any information. This is based on the access structure where only a qualified set reconstructs the secret, whereas the forbidden set does not. The shares generated are meaningless. [50] were the first to enhance the sharing of random grids for threshold structure; i.e., (2, n) and (n, n). The correctness of this scheme was studied, verified and extended to gray and color images; This scheme generates meaningless shares.

Shyu, D. J., encrypted a secret image into n random grids and by superimposing [51] all the n grids the secret could be recovered. [52] enhanced the random grid secret sharing for threshold based sharing, for both binary and color images. The applications are visual authentication, digital watermarking, and image hiding. Generally, the generated secret shares are meaningless. User friendly, i.e., meaningful shares were generated by [7] where meaningful shares created in visual cryptography have a size larger than the secret.

A novel method was introduced to hide the share into a cover image to generate meaningful shares. This scheme supports both gray and binary images. [53] modifies the (2, n) scheme to a threshold based (t, n), using an access structure with no pixel expansion. Small white regions are constructed as black, thus leading to the loss of the secret. Improving the contrast is left as an open problem for future research.

Cheating happens when dishonest participants collude to cheat others by giving wrong secret information generated by the former. [54] provided the insight into the collusion attacks of random grids. A Cheating prevention scheme is a good topic for future work. [55] proposed the Random grid based visual secret sharing of greyscale and color images for general access structure, and also provided an authentication mechanism to prevent cheating by dishonest participants. A Cheating immune method using the genetic algorithm is provided to prevent the cheating of the participants [56]. [57] proposed a fixed number of shares using the (2, n) scheme, and adjustable number of shares using the (2, ∞) scheme where the shares can be extended at any time.

[58] encrypts multiple secret images into two circular grids, and decrypts one secret by stacking as it is and gradually rotating the other at a fixed degree to disclose the other secret. Four secrets were encrypted into 2 random grids without any pixel expansion by [59]. They reconstructed the first secret by directly stacking it and the other three secrets by rotating at different angles of 90°, 180°, or 270°. All these proposed research works had no pixel expansion and no code book was required to generate the shares during encryption.

TABLE IV
FUNCTIONALITY COMPARISON OF RECENT RESEARCH WORKS

Ref	Threshold	Pixel Expansion	Codebook Necessary	Verification	Meaningful Shares
[47], [48], [51]	No	No	No	No	No
[50]	Yes	No	No	No	No
[52]	Yes	No	No	No	No
[49]	No	No	No	No	Yes
[53]	Yes	No	No	No	No
[55]	No	No	No	Yes	No

IV. Secret Sharing Using Visual Cryptography

Visual Cryptography encrypts a secret into n shares (transparencies) and the secret is retrieved by stacking these shares. When visual cryptography is applied on a binary image, the size of the reconstructed image is increased, which leads to a poor contrast level.

The shares are meaningless too [60] were the first to introduce visual cryptography for binary images in the year 1995 [61] shared a binary secret using halftone visual cryptography. Meaningful halftone shares were generated based on the error diffusion technique.

This scheme was extended to color images in [62] by the same authors, using Visual Information Pixel synchronization and error diffusion. Multiple secrets were framed [63] from a single image by dividing the image into more number of regions, which are further shared among a group of n participants. The secret can be reconstructed only if all the shares are pooled together. [64] proposed a secret sharing for binary images where each share is divided into sub shares such that each Sub share will be of a different size. The sub shares should be expanded before stacking.

The sub shares are stacked recursively to retrieve the shares. [65] developed a scheme that shares multiple secrets among a pair of participants, by converting the secrets into many bit planes and embeds, based on the LSB substitution technique of steganography. A pair of participants can therefore reconstruct the secret without destroying its secrecy by the simple computation and stacking of two stego images. The PSNR values of the stego-images are always above 42 dB. The only limitation of the proposed scheme is that a minor computation is required before stacking the stego-images. The required computation is simple and efficient, and effectively resolves the unstackability problem of

visual cryptography. [9] proposed a new (t, n) probabilistic visual secret sharing scheme based model with non expansion. The proposed method achieves the same contrast level as that of the conventional visual secret sharing schemes, which also proved that the conventional visual secret sharing scheme can be transferred to Probabilistic visual secret sharing by using a transfer function. The frequency of the white pixel is used to show the contrast of the recovered image.

To avoid the distortion in aspect ratio of the recovered images, an aspect ratio invariant visual secret sharing scheme is proposed by [66] For circular images, the pixel expansion is not a square value, and hence, the circle will change in to an ellipse after applying the visual secret sharing technique; therefore, there will be a loss of information. To avoid the changing of the aspect ratio, dummy pixels are added to avoid the distortion of images. To construct an aspect ratio invariant visual secret sharing scheme, a method is proposed to significantly reduce the number of extra sub pixels. Color visual cryptography was analysed in [67]-[75] where [76] extends the secret sharing by hiding in to a cover image in extended Visual cryptography.

Wu, H C., et.al., developed a $(2, 2)$ visual secret sharing scheme, in which two sets of confidential messages are embedded at different angles [77], and the secrets are reconstructed by rotating the share to a fixed angle. Since circular shares are used in this scheme, it overcomes the limitations of rectangular shares which have angular restrictions. In this scheme every two degrees is considered as one unit. The first share is created randomly and the second share according to the first share and it embeds the confidential messages. By stacking the two shares the content of the first confidential message can be obtained, and then by rotating one of the shares to a certain degree the second confidential message can be obtained.

The proposed technique can be used to embed two times the number of confidential messages when compared to traditional visual cryptography. A two-in-one visual cryptography scheme [78] is presented, that not only shares a confidential image in two noisy transparencies, but also hides a confidential text file describing the image in these two transparencies. None of the transparencies alone can reveal anything about the image or text. The secret image can be viewed by simply stacking the two transparencies; and with simple computations, the more confidential text data can also be extracted [79] explains how the sub pixels in the share are arranged to retain the spatial location and direction of the sub pixel, in order to avoid the distortion of the recovered image.

A more efficient k -out-of- n visual secret sharing scheme with reduced pixel expansion was developed by [80]. A new color model is defined which is feasible and more realistic for the human visual system. The proposed visual secret sharing scheme is based on the new color model and existing binary scheme, with a pixel expansion of $\lceil \log_2 c \rceil \times m$ where m is the pixel expansion

of the exploited binary scheme. The final decryption process is done by the human visual system which avoids complex computations, and hence, the scheme is handy and cost-effective. [81] and [82] schemes have the pixel expansion of about $c \times m$, whereas [80] has a pixel expansion of about $\lceil \log_2 c \rceil \times m$ which is more advanced and considerable when c , the number of colors in the secret image becomes large. Multiple secrets are encrypted into two circular shares [83], where every secret is obtained by rotating to various angles [84] proposed a friendly progressive secret sharing which generates meaningful shares, where all the shares are progressively stacked to retrieve the original secret.

[66] proposed the $(2, 2)$ verifiable secret sharing scheme to share secret images like the binary, greyscale or color images using three techniques, such as Error diffusion, half toning transform, and image clustering, that do not need any pixel expansion. Using an embedded halftone logo, the verification by participants can be done without the need of any additional information. Based on the mean square error between the extracted halftone logo and the original halftone logo, the proposed scheme allows truthful participants to verify their liability of their constructed secret image. If the *MSE* value is not equal to zero, the constructed image is considered as fake, even if the reconstructed image is meaningful.

The secret image is recovered based on the Boolean operator XOR using the look up table. The reconstructed secret image by the scheme gives quality, in terms of the PSNR ranging from 32.63 to 34.37 dB regardless of whether the images are greyscale or color. The same authors extended the work [85] for a (t, n) threshold scheme, where t shares reconstruct the secrets and less than that cannot. Two secrets are shared into two meaningless share images by Tsung Lih Lin et.al., [86] with no Pixel expansion.

This does not leak any information, so security is confirmed. During reconstruction the two shares are stacked, to retrieve the secret without any computation.

Vinodhini, A., et.al., used CAPTCHA as a secret message [87], which is an authentication mechanism, and the secret is reconstructed without any loss. [88] analyzed various Halftoning techniques that can be used in visual cryptography to generate meaningful shares.

The authors found that the Floyd and Steinberg scheme gives good performance in Visual Cryptography. [89] securely shared a binary image using visual cryptography, and their shares were embedded into a host image using the LSB technique. This requires a small processing during decryption in order to retrieve the shares, which are then stacked to reconstruct the secret.

Lin, D J., et.al., encoded two secrets [90] where the first secret is retrieved by stacking as it is, and the other is recovered by flipping and the stacking. Optimal contrast is achieved without pixel expansion and is perfect. In [91] Pixel expansion can be set by the user; if the limit is 1 then the shares do not suffer pixel expansion. The quality of secret images is equivalent to

the existing deterministic visual secret sharing. [92] proposed a novel secret sharing, where a secret and a confidential secret are also hidden into meaningful cover images with no pixel expansion. The secret is decrypted back by stacking the shares and by fixing one share image and flipping the other the confidential secret is retrieved without any computation. Authentication is provided in this scheme, whereas it is not dealt with in [78] scheme. Multiple secrets such as two or four or eight are encrypted into two shares [93], where turning or flipping around reconstructs all the secrets. [94] developed a new (t, ∞) scheme where the number of shares are not fixed and can be increased dynamically without disturbing the original shares. [95] proposed an authentication based cheating prevention mechanism, to identify dishonest participants for the Noar and Shamir method.

TABLE V
FUNCTIONALITY COMPARISON OF RELATED SCHEMES
IN VISUAL CRYPTOGRAPHY

Ref	Meaningful Shares	Pixel Expansion	Verification	Number of Secrets
[86]	No	No	No	Multiple
[90]	No	No	No	Multiple
[92]	Yes	No	Yes	Multiple
[93]	No	Yes	No	Multiple
[94]	No	Yes	No	Single
[95]	No	Yes	No	Single
[72]	No	No	Yes	Single
[85]	No	No	Yes	Single
[61]	Yes	Yes	No	Single
[62]	Yes	Yes	No	Single
[64]	No	Yes	No	Single
[65]	Yes	No	No	Single
[66]	No	Yes	No	Single
[77]	No	Yes	No	Multiple
[78]	No	Yes	No	Multiple
[79]	No	No	No	Single
[80]	No	Yes	No	single

The size invariant visual cryptography scheme reduces pixel expansions, but has two other draw backs, i.e. it suits only a coarse secret image and has the thin line problem [96] addressed this problem which improves visual quality by reducing the variance of the darkness level. This avoids the thin line problem also.

V. Secret Sharing Using Other Methods

[97] and [98] dealt with secret sharing based on the Chinese remainder theorem. [97] deals with the lattice based theory to facilitate the threshold-changeability of secret sharing, whereas [98] demonstrates Function sharing with the Asmuth bloom secret sharing scheme. Robustness, Pro-activity, and the efficiency improvement are the major problems left open.

The secret sharing based on the Cellular Automata proposed in [99], [29] is highly effective against differential attacks. The computational complexity of such schemes is linear and they reconstruct lossless secrets. The generated shares are of the same size as that

of the secret. This scheme utilizes the neighborhood configurations of the general linear 2-D cellular automata in secret image sharing [100]-[102] deals with multiparty quantum secret sharing which are insecure towards cheating. [103] proposed a self renewable hash chain based on the Ito-Saito-Nishizeki secret sharing scheme.

The Hash chain is used in a variety of encryption applications and services [104] proposed a new multiplicative linear ramp secret sharing scheme based on error correcting codes, and their strong multiplicative properties are analyzed. [105] used a multi party computation to solve the leakage of some secrets during the process of reconstruction due to the published shares.

The Multi party computation protocol is used to solve linear multi secret sharing [106] proposed a coding theory which reduces the computational efficiency because a huge amount of data is present in the image. This method can be extended to audio and video tracks, and has many practical applications. This uses a coding technique which is similar to the RS eraser code, where the secret image is revealed using coding and decoding.

[107] demonstrated two secure verifiable multi secret sharing schemes, based on non homogeneous linear recursion and elliptic curves. The security factors are based on the ECDLP. It does not need any secure channel for secret distribution. The security of this system is based on the ECRSA cryptosystem and EDCLP. [108] also securely shares the secret using homogenous linear recursion [109], [110] deal with the Vamos Matroid, which includes two non isomorphic access structures having information rates of atleast $3/4$.

The problem related to pixel reconstruction is analyzed, which results in better contrast.

VI. Summary of the Literature Survey

Secret Sharing for a single secret is performed using the Shamir polynomial, where evaluation is done using a prime field GF(p). The coefficient of the polynomial is a secret pixel and random numbers in the Shamir model, modified by Thien and Lin where all the coordinates are the secret pixel, so that the size is reduced to $1/t$ of that of the secret. Wang R Z et.al, [4] reduced the shadow size further to about 40% than Thien and Lin's using the Huffman encoding. Thien and Lin' used the $GF(2^t)$, which allows the reconstruction of the secret without loss; instead, if GF(p) is used always the prime chosen is 251, which leads to the loss of data above that. Meaningful shares are created by embedding them into a host image using the LSB [28],[22],[10],[8],[23] or the modular operator [11], where a better PSNR is achieved in all the schemes whereas [10] is better. Most of the shares get expanded, when they are embedded into a host image. Scalable secret sharing and threshold secret sharing are presented, where scalable sharing reconstructs the secret as the shares pooled get scaled, whereas in the threshold scheme t or more shares are pooled to reconstruct the secret. Verification of the shares is done by embedding the authentication bits into

the share image. The embedding capacity of [9],[28],[22],[10],[8],[23],[11] is $m \times n/4$ where as in Lin et.al,[11] it is $m \times n(t-3)/4$.

Sharing of multiple secrets was proposed by YCH, [33] based on the Shamir model. This YCH scheme [75], [36]-[40] further improves the verification of both the participant and the dealer. This is performed using the one way function or discrete logarithm or RSA, to avoid cheating by both. Multiple secrets are shared using the Monotone span program [111]-[118] which has the same security as that of Shamir's secret sharing, and is a perfect scheme where meaningful shares are generated and the secrets are reconstructed without loss. Sharing can be done based on the access structure [41], [43], [44], [105] where secrets can be reconstructed only if the qualified shares are pooled, not if the forbidden set is pooled.

Random grids [39], [47], [50]-[56], [59], make use of binary values where no pixel expansion and code book generation is necessary for share creation. Kafi and Karen developed a (2,2) sharing scheme, where a random grid is generated, based on a probability of 0.5 for binary values. Based on this random grid, and using the equivalent or complement rule, the other share is generated. Staking the shares the secret can be reconstructed with low contrast. Later, this scheme was extended to n shares where n refers to the participants, and further extended to threshold based sharing, where t participants can reconstruct the secrets. Meaningful shares are generated and verification is also introduced to avoid cheating.

Visual Cryptography is applied on a binary image, and the size of the reconstructed secret is increased, which leads to a poor contrast level. The generated shares are meaningless. When the cryptographic technique is extended on the gray scale image, first it gets converted into a binary image and preceded. This scheme suffers from pixel expansion and poor contrast level, and the shares are meaningless.

In extended visual cryptography, the meaningless shares are embedded in the host images to make them meaningful, to avoid the suspicion of intruders. To improve the quality of meaningful shares, a halftoning process is required.

This also suffers from pixel expansion. Though it has a better contrast level when compared to the meaningless shares, still the contrast is poor. When cryptography is applied on true color image reconstruction, it considers only three color components like Red, Blue, Green or Cyan, Yellow, Magenta. The pixel expansion depends upon the number of colors used in the original image.

Reconstruction of colors is found difficult due to the color addition phenomenon. So this also suffers loss of quality.

Multiple secrets are framed from a single image, by dividing it into a number of regions and multiple shares are created from those regions. Pixel expansion depends upon the number of secrets. The main drawback in visual cryptography is that it requires a codebook for share

generation and the created shares suffer from pixel expansion and the reconstructed secret has low contrast.

VII. Scope for Future Work

Future scope for the sharing using Shamir model lies in the creation of meaningful shadows without pixel expansion. In visual cryptography the generation of code book has to be modified in such a way that it leads to the creation of shadows without pixel expansion, which should also improve the contrast of the reconstructed secret.

In Random grid secret sharing an improvement in the contrast of the reconstructed secret can be dealt with seriously, since it does not suffer from pixel expansion and code book generation. A lot of scope is there for creating meaningful shares with better PSNR to avoid the suspicion of intruders. Future work can focus on the reconstruction of colors when color images are shared. Better cheating prevention schemes for both the participants and the dealer can improve the authenticity of shares.

VIII. Conclusion

Recent technological advances in the digital world and the security of digital images have become an inseparable issue since the communication is over an open network. In this paper, the existing research works related to secret sharing have been surveyed and analyzed.

The performance of secret sharing provides high security. The functional comparison of various techniques is presented, analyzed and evaluated.

References

- [1] Shamir, A. "How to share a secret," *Communications of the ACM*, (22:11), 1979, 612-613.
- [2] Thien, C. C., Lin, J. C. "Secret image sharing," *Computers & Graphics*, (26:5), 2002, 765-770.
- [3] Lin, C. C., Tsai, W. H., "Secret image sharing with steganography and authentication," *The Journal of Systems and Software*, (73:3), 2004, 405-414.
- [4] Wang, R. Z., Su, C. H. "Secret image sharing with smaller shadow images," *Pattern Recognition Letters*, (27:6), 2006, 551-555.
- [5] Wang, R. Z., Shyu, S. J. "Scalable secret image sharing," *Signal Processing Image Communication*, (22), 2007, 363-373.
- [6] Lin, Y. Y., Wang, R. Z., "Scalable Secret Image Sharing With Smaller Shadow Images," *IEEE Signal Processing*, (17:3), 2010, 316-319.
- [7] Wang, R. Z., Chien, Y. F., Lin, Y. Y. "Scalable user friendly image sharing," *Journal of Visual Communication and Image Representation*, (21:7), 2010, 751-761.
- [8] Chang, C. C., Hsieh, Y. P., Lin, C. H. "Sharing secrets in stego images with authentication," *Pattern Recognition*, (41:10), 2008, 3130-3137.
- [9] Yang, C. N., Chen, T. S., Yu, K. H., Wang, C. C. "Improvements of image sharing with steganography and authentication," *Journal of Systems Software*, (80:7), 2007, 1070-1076.
- [10] Yang, C., Chen, T., "Extended visual secret sharing schemes: improving the shadow image quality," *International Journal of Pattern Recognition and Artificial Intelligence*, (21:5), 2007, 879-898.

- [11] Lin, P Y., Lee J S., Chang, C C. "Distortion-free secret image sharing mechanism using modulus operator," *Pattern Recognition*, (42:5),2009, 886 – 895
- [12] Li,S., Zheng,X. "On the security of an image encryption method," *Proceedings of the 2002 IEEE International Conference on Image Processing*, (2) , 2002, 925–928.
- [13] Thien,C C., Lin,J C. "A simple and high-hiding capacity method for hiding digit- by-digit data in images based on modulus function," *Pattern Recognition*, (36 :12), 2003, 2875–2881.
- [14] Li,S., Li,S., Lo,K T., Chen,G. "Cryptanalysis of an image encryption scheme," *J. Electron. Imaging*, (15:4) , 2006, 043012–043113.
- [15] Li,C., Li,S., Alvarez,G., Chen,G., Lo,K T. "Cryptanalysis of a chaotic block cipher with external key and its improved version," *Chaos Solitons Fractals*, (37:7), 2008, 299–307.
- [16] Yang,C N., Lai, C S. "New colored visual secret sharing schemes," *Des. Codes Cryptogr.* (20:3), 2000, 325–335.
- [17] Tripathy, P.K., Biswal, D., Multiple server indirect security authentication protocol for mobile networks using elliptic curve cryptography (ECC), (2013) *International Review on Computers and Software (IRECOS)*, 8 (7), pp. 1571-1577.
- [18] Stinson, D.R. "Cryptography—Theory and Practice," CRC Press, New York, USA, 2002.
- [19] Rivest, R.L., Shamir,A., Adleman,L. "A method for obtaining digital signatures and public-key cryptosystems," *Commun. ACM*, (21:2) , 1977 , 120–126.
- [20] Stallings,W. "Cryptography and Network Security–Principles and Practices," Pearson Education Inc., New Jersey, USA, 2006 , 238–241.
- [21] Tsai,D S., Horng,G., Chen,T H., Huang,Y T. "A novel secret image sharing scheme for true- color images with size constraint," *Information Sciences*, (179:19), 2009. 3247–3254.
- [22] Chang,C C., Chen,Y H., Wang,H C. "Meaningful secret sharing technique with authentication and remedy abilities," *Information Science*, (181:14), 2011, 3073–3084.
- [23] Ulutas,M., Ulutas,G., Nabiyeve, V V. "Medical image security and EPR hiding using Shamir's secret sharing scheme," *The Journal of Systems and Software*, (84:12),2011, 341–353.
- [24] Anbarasi, L J., Kannan, S. "Secured Secret Color Image Sharing With Steganography," *IEEE- International Conference on Recent Trends in Information Technology*, 2012, 44 - 48.
- [25] Blakley, G.R., "Safeguarding cryptography keys," *Proc. of the AFIPS, National Computer Conference*, (48), 1979, 313-317.
- [26] Tso, H.K., "Sharing secret images using Blakley's concept," *Optical Engineering*,(47:7) , 2008.
- [27] Ulutas,G., Ulutas,M., Nabiyeve, V V., "Distortion free geometry based secret image sharing," *Procedia Computer Science*, (3), 2011, 721–726.
- [28] Eslami,Z., Ahmadiabadi,J Z. "Secret image sharing with authentication-chaining and dynamic embedding," *The Journal of Systems and Software*,(84:5) , 2011, 803–809.
- [29] Jin,J., Wu,C H. "A secret image sharing based on neighborhood configurations of 2-D cellular automata," *Optics & Laser Technology*,(44:3), 2012, 538–548.
- [30] Yang,C., Chen, T., "Extended visual secret sharing schemes: improving the shadow image quality," *International Journal of Pattern Recognition and Artificial Intelligence*, (21:5), 2007, 879–898.
- [31] Yang,C N., Chu,Y Y. "A general (k, n) scalable secret image sharing scheme with the smooth scalability," *The Journal of Systems and Software*, (84:10), 2011, 1726– 1733.
- [32] Elsheh,E., Hamza,A B., "Secret sharing approaches for 3D object encryption," *Expert Systems with Applications*,(38:11), 2011, 13906–13911.
- [33] Yang,C C.,Chang,T Y., Hwang,M S "A (t, n) multi-secret sharing scheme," *Applied Mathematics and Computation*, (151:2), 2004, 483–490.
- [34] Chien, H Y., Jan, J K., Tseng, Y M. "A practical (t, n) multi-secret sharing scheme," *IEICE Transactions on Fundamentals* (83:12) , 2000, 262–2765.
- [35] Shao,J., Cao,Z. "A new efficient (t, n) verifiable multi-secret sharing (VMSS) based on YCH scheme," *Applied Mathematics and Computation*,(168:1),2005, 135–140.
- [36] Harn,L. "Efficient sharing (broadcasting) of multiple secrets," *IEE Proc. Comput. Digit. Tech.* (142:3), 1995, 237–240.
- [37] Chang,T Y., Hwang,M S.,Yang,W P. "An improvement on the Lin–Wu (t,n) threshold verifiable multi-secret sharing scheme," *Applied Mathematics and Computation*, (163:1), 2005,169–178.
- [38] Lin,T Y., Wu, T.C. "(t, n) threshold verifiable multisecret sharing scheme based on factorization intractability and discrete logarithm modulo a composite problems," *IEEE Proc. Comput. Digit. Tech.*, (146 :5) ,1999, 264–268.
- [39] Zhao J., Zhang,J., Zhao,R. "A practical verifiable multi-secret sharing scheme," *Computer Standards & Interfaces*,(29:1), 2007, 138–1
- [40] Hu,C., Liao,X., Cheng,X. "Verifiable multi-secret sharing based on LFSR sequences," *Theoretical Computer Science*, (445), 2012, 52–62.
- [41] Fredkin, E F. "A new multi- secret image sharing scheme using Lagrange's interpolation," *System Software Digital Mechanics*, (76:8), 2005) 327-329.
- [42] De Santis, A., Masucci, B. "New results on non-perfect sharing of multiple secrets," *The Journal of Systems and Software*, (80:2), 2007, 216–223.
- [43] Chan, C W., Chang,C C. "A scheme for threshold multi-secret sharing," *Applied Mathematics and Computation*, (166:1) , 2005, 1–14.
- [44] Farre,J M., Padro,C. "Secret sharing schemes on access structures with intersection number equal to one," *Discrete Applied Mathematics*, (154:3), 2006, 552 – 563.
- [45] Farre, J M. "A note on secret sharing schemes with three homogeneous access structure," *Information Processing Letters*, (102:4), 2007, 133–137.
- [46] Guo, C., Chang,CC., Qin,C. "A hierarchical threshold secret image sharing," *Pattern Recognition Letters*, (33:1), 2012, 83–91.
- [47] Kafri,O., Keren,E., "Encryption of pictures and shapes by random grids," *Optic Letters*, (12:6), 1987, 377–379.
- [48] Shyu,D J. "Image encryption by random grids," *Pattern Recognition*,(40:3), 2007, 1014 –1031.
- [49] Chen,T H., Tsao,K H. "User-Friendly Random-Grid-Based Visual Secret Sharing," *IEEE Transactions on Circuits and Systems for Video Technology*, (21:11), 2011. 1693 – 1703.
- [50] Chen,T H., Tsao,K H. "Visual secret sharing by random grids revisited," *Pattern Recognition*, (42:9), 2009, 2203 – 2217.
- [51] Shyu,D J., "Image encryption by multiple random grids," *Pattern Recognition*,(42:7),2009,1582- 1596.
- [52] Tzung,H C., Her. C., Tsao, KH., "Threshold visual secret sharing by random grids," *The Journal of Systems and Software*, (84:7), 2011, 1197–1208.
- [53] Shyu,S J., "Visual Cryptograms of Random Grids for General Access Structures," *IEEE Transactions on Circuits and Systems for Video Technology* ,(23:3), 2013, 414 - 424.
- [54] Lee,Y S., Chen,T H. "Insight into collusion attacks in random-grid-based visual secret sharing," *Signal Processing*, (92:4),2012, 727–736.
- [55] Wu,X., Sun,W. "Random grid-based visual secret sharing for general access structures with cheat preventing ability," *The Journal of Systems and Software*,(85:5), 2012, 1119– 1134.
- [56] Chen,T H., Tsao,K H. "Threshold visual secret sharing by random grids," *The Journal of Systems and Software*, (84:7),2011, 1197–1208.
- [57] Chen,S K., Lin,S J. "Optimal (2, n) and (2, ∞) visual secret sharing by generalized random grids," *Journal of Visual Communication and Image Representation*, (23:4) ,2012, 677–684.
- [58] Chen,T H., Li,K C. "Multi-image encryption by circular random grids," *Information Sciences*, (189), 2012, 255–265.
- [59] Chen ,T H., Tsao,K H., Lee,Y S. "Yet another multiple-image encryption by rotating random grids," *Signal Processing*, (92:4), 2012, 2229–2237.
- [60] Naor,M.,Shamir.,A. "Visual cryptography," *Lecture Notes in Computer Science*, (950),199 1–12.
- [61] Wang,Z.,Arce, G R., and CrescenzoG D., "Halftone Visual Cryptography via Error diffusion," *IEEE Transaction on Information Forensics and Security*, (4:3), 2009, 383– 396.
- [62] Kang,I., Arce,GR., Lee,H K., "Color Extended Visual Cryptography Using Error Diffusion," *IEEE Transactions on*

- Image Processing*, (20:1), 2011, 132 - 145.
- [63] Wang,R Z., "Region Incrementing Visual Cryptography," *IEEE Signal Processing*, (16 :8) , 2009, 659 - 662.
- [64] Liu,F., Wu,S., and Lin,X. "Step Construction of Visual Cryptography Schemes," *IEEE Transaction on Information Forensics and Security*, (5:1), 2010 ,27 - 38.
- [65] Tsai,C S., Chang,C C., Chen,T S. "Sharing multiple secrets in digital images," *The Journal of Systems and Software*, (64:2), 2002, 163–170.
- [66] Yang,C N., Chen,T S. "Aspect ratio invariant visual secret sharing schemes with minimum pixel expansion," *Pattern Recognition Letters*, (26:2) , 2005, 193– 206.
- [67] Liu,F., Wu,C K., Lin,XJ., "Color Visual Cryptography Schemes," *IEEE Transaction on Information Forensics and Security*, (2:4), 2008, 151 - 165
- [68] Leung,B W., Ng,F Y., Wong,D S. "On The Security of a Visual Cryptography Scheme for Color Images," *Pattern Recognition*, (42:5), 2009, 929-940
- [69] Liu,F., Wu,C K., Lin ,X. "A New Definition of the Contrast of Visual Cryptography Scheme," *Information Processing Letters*, (110:7), 2010, 241–246.
- [70] Blundo,C., DeSantis, A., Stinson, D A. "Improved schemes for visual cryptography," *Des. Codes Cryptogr.*, (24:3), 2001, 255– 278.
- [71] Anbarasi, L J., Vincent J., Mala, G C A. "A Novel Visual Secret Sharing Scheme for Multiple Secrets via Error Diffusion in Halftone Visual Cryptography," *IEEE- International Conference on Recent Trends in Information Technology*, 2011, pp 3-5.
- [72] Chang,C C., Lin,C C., Le,H N., Le,H B. "Sharing a verifiable secret image using two shadows," *Pattern Recognition*,(42:1), 2009, 3097 – 3114.
- [73] Chang,C C., Lin,C C., Lin,C H., Chen,Y H. "A novel secret image sharing scheme in color images using small shadow images," *Information Sciences*, (178:11), 2008, 2433–2447.
- [74] Chen,T H., Huang,J C., "A novel user-participating authentication scheme," *Journal of Systems and Software*, (83:5) , 2010, 861–867.
- [75] Hajiabollahsani,H., Cheraghi,A., "Bounds for Visual Cryptography Schemes," *Discrete Applied Mathematics*,(158:6), 2010, 659-665.
- [76] Wang,D., Yi,F., Li,X. "On General Construction for Extended Visual Cryptography Schemes," *Pattern Recognition*, (42:11) , 2009, 3071-3082.
- [77] Wu,H C., Chang,C C. "Sharing visual multi-secrets using circle shares," *Computer Standards & Interfaces*, (28), 2005, 123–135.
- [78] Fang,W P., Lin,J V. "Visual cryptography with extra ability of hiding confidential data," *Journal of Electronic Imaging*, (15:2),2006), 0230201–0230207.
- [79] Yang,C N., Chen,T S. "Reduce shadow size in aspect ratio invariant visual secret sharing schemes using a square block-wise operation," *Pattern Recognition*, (39:7) ,2006,1300 – 1314.
- [80] Shyu,D J. "Efficient visual secret sharing scheme for color images," *Pattern Recognition*, (39:5) 2006, 866-880.
- [81] Liu,F., Wu ,C K ., Lin,X J. "Color Visual Cryptography Schemes," *IEEE Transaction on Information Forensics and Security*, (2:4), 2008, 151 – 165.
- [82] Yang,C N., Chen,T S. "Colored Visual Cryptography Scheme Based on Additive Color Mixing," *Pattern Recognition*, (41:10), 2008, 3114-3129.
- [83] Shyu,D J., Huang, S Y., Lee,Y K., Wang,R Z., Chen,K. "Sharing multiple secrets in visual cryptography," *Pattern Recognition*, (40:12), 2007, 3633 – 3651.
- [84] Fang,W P., "Friendly progressive visual secret sharing," *Pattern Recognition*, (41:4), 2008, 1410 – 1414.
- [85] Chang,C .Lin,C C.,Le,H N.,Le,H B. "Self-Verifying visual secret sharing using error diffusion and Interpolation Techniques," *IEEE Transactions on Information Forensics and Security*, (4), 2009 , 790 – 801.
- [86] Lin,T L., Horng,S J., Lee,K H., Chiu,P L., Kao,T W., Chen,Y H.,Run,R D., Lai,J L., Chen,R J. "A novel visual secret sharing scheme for multiple secrets without pixel expansion," *Expert Systems with Applications*, (37) , 2010, 7858–7869.
- [87] Vinodhini,A., Anbarasi,L J. "Visual Cryptography for Authentication Using CAPTCHA," *International Journal of Computer and Internet Security*, (2:1), 2010, 67-76.
- [88] Alex, N S., Anbarasi, L J., "Enhanced Image Secret Sharing via Error Diffusion in Halftone Visual Cryptography," *International Conference on Electronics Computer Technology*, (2), 2011, pp.393 – 397.
- [89] Prasanna, D. R. L., Anbarasi,L J., Vincent J. "A Novel Approach for Secret Data Transfer using Image Steganography and Visual Cryptography," *ICCCS'11*, 2011, 12-14.
- [90] Lin,D J., Chen,S K., Lin,J C. "Flip visual cryptography (FVC) with perfect security, conditionally-optimal contrast, and no expansion," *Journal of Visual Communication and Image Representation*, (21), 2010, 900–916.
- [91] Wang,D., Yi,F., Li,X. "Probabilistic visual secret sharing schemes for grey-scale images and color images," *Information Sciences*, (181:11), 2011, 2189–2208.
- [92] Lou,D C., Chen,H H., Wu,H C., Tsai,C S. "A novel authenticatable color visual secret sharing scheme using non-expanded meaningful shares," *Displays*, (32:3), 2011, 118–134.
- [93] Shyu,S J., Chen,K. "Visual multiple secret sharing based upon turning and flipping," *Information Sciences*, (181:15), 2011, 3246–3266.
- [94] Lin,S J.,Chung,W H. "A Probabilistic Model of (t, n) Visual Cryptography scheme with dynamic Group,"*IEEE Transactions On Information Forensics And Security*,(7:1),2012,197– 207.
- [95] Chen,Y C.,Tsai,D S.,Horng,G. "A new authentication based cheating prevention scheme in Naor shamir's visual cryptography," *Journal of Visual Communication and Image Representation* 23:8), 2012, 1225–1233.
- [96] Liu,F., Guo,T., Wu,C., Qian,L. "Improving the visual quality of size invariant visual cryptography scheme," *Journal of V Comm and Image Representation*,(23:2),2012, 331–342.
- [97] Steinfeld,R., Pieprzyk,J., Wang,H. "Lattice-based threshold-changeability for standard CRT secret-sharing schemes," *Finite Fields and their Applications*, (12:4),2006, 653 – 680.
- [98] Kaya,K., Selcuk,A A., "Threshold cryptography based on Asmuth–Bloom secret sharing," *information Sciences*,(177:19), 2007, 4148–4160.
- [99] Wu,X., Ou,D., Liang,Q., Sun,W. "A user-friendly secret image sharing scheme with reversible steganography based on cellular automata," *The Journal of Systems and Software*, (85:8), 2012, 1852– 1863.
- [100] Zhang,Z J., Gao,G., Wang,X., Han, L F., Shi,S H. "Multiparty quantum secret sharing based on the improved Boström–Felbinger protocol," *Optical Communication*,(269:2), 2007, 418–422.
- [101] Lin,S.,Gao,F.,Wen,Q .,Zhu,F C. "Improving the security of multiparty quantum secret sharing based on the improved BoströmFelbinger protocol," *Optical Comm*, (281:17),2008, 4553-4554.
- [102] Gao,G. "Eavesdropping on the improved three-party quantum secret sharing protocol," *Optics Communications*, (284:3), 2011, 902–904.
- [103] Ting,D., Ping,H H., Chuan,W R., Xing,P X. "Novel self-renewal Hash chain based on Ito-Saito- Nishizeki secret sharing scheme," *The Journal of China Universities of Posts and Telecommunications*, (19:2), 2012, 122–127.
- [104] Chen,Q., Pei,D., Tang,C., Yuea,Q., Ji,T. "A note on ramp secret sharing schemes from error- correcting codes," *Mathematical and Computer Modelling*, ,July (2011).
- [105] Liu,M., Xiab,L., Zhang,Z. "Linear multi-secret sharing schemes based on multi-party computation," *Finite Fields and their Applications*, (12:4) , 2006, 704-713.
- [106] Tang, D. "The Research of Secret Image Sharing Based on RS Erasure Code," *Procedia Engineering*,(29:1),2012, 27 – 32.
- [107] Dehkordi,M H., Mashhadi,S. "Verifiable secret sharing schemes based on non-homogeneous linear recursions and elliptic curves," *Computer Communications*, (31:9), 2008, 1777–1784.
- [108] Dehkordi,M H., Mashhadi,S. "New efficient and practical verifiable multi-secret sharing schemes," *Information Sciences*, (178:9), 2008, 2262–2274.
- [109] Metcalf-Burton, J R . "Improved upper bounds for the information rates of the secret sharing schemes induced by the Vámos matroid," *Discrete Mathematics*, (311:8-9), 2011, 651– 662.

- [110] Li,P.,Ma,PJ.,Su,X H.,Yang,C N. "Improvements of a two-in-one image secret sharing scheme based on gray mixing model," *Journal of Visual Communication and Image Representation*, (23:3), 2012, 441–453.
- [111] Zhang,Z., Liu,M., Xiao,L. "Rearrangements of access structures and their realizations in secret sharing schemes," *Discrete Mathematics*, (308:21), 2008 , 4882–4891.
- [112] Hsu,C F., Cheng, Q., Tang,X. Zeng,B. "An ideal multi-secret sharing scheme based on MSP," *Information Sciences*, (181:7), 2011, 1403–1409.
- [113] Guo,C., Chang,C C., Qin,C. "A Multi-threshold Secret Image Sharing Scheme Based on MSP, " *Pattern Recognition Letters*, (33:12), 2012, 1594 -1600.
- [114] Hsu, C.F., Cheng, Q., Tang, X.M., Zeng, B., "An ideal multi-secret sharing scheme based on MSP," *Inf. Sci.* (181 :7) , 2011, 1403–1409.
- [115] Liu,Z., Liu,S., Ahmad,M A. "Image Sharing Scheme Based on Discrete Fractional Random Transform," *Optik - International Journal for Light and Electron Optics*, (121:6), 2010, 495– 499.
- [116] Wu,T C., He,W H. "A geometric approach for sharing secrets," *Computers & Security*, (14) 995,135-145.
- [117] Devi, M., Chenthur Pandian, S., An efficient autonomous key management with verifiable secret sharing schemes for reduced communication/computation costs in MANET, (2014) *International Review on Computers and Software (IRECOS)*, 9 (1), pp. 48-53.
- [118] Yang,C N. "New visual secret sharing schemes using probabilistic method," *Pattern Recognition Letters*, (25:4),2004, 481–494.

Authors' information

¹Sri Ramakrishna Institute of Technology,Coimbatore, India.

²Professor & Head, Department of CSE, Easwari Engineering College, Chennai, India.

Method for Automatic Ontology Building in Costumer Support Expert System for Energy Consumption

A. Stropnik¹, M. Zorman²

Abstract – Today, web portals include much information, mostly in an unstructured form that users easily read and understand. Reading and processing of such information is a rather complicated task for machine or computer. Raw and unemployed information as well as untapped knowledge is very unacceptable due to increased tendency for automated information processing. Therefore, we developed a system for automatic building of a knowledge base through web pages. This paper presents such system using VIPS algorithm. Demonstrated as part of an existing expert system for customer support of energy consumption, the new system was tested by users test group of Slovenian energy provider Elektro energija d.o.o. The results of testing are presented at the end of the article. **Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords: Automatic Building Ontology, Expert System, Energy Consumption, Information Retrieval

I. Introduction

In the last decade, a real boom marked the World Wide Web and it widespread at home and at work. As one of the largest and most frequently exploited such resources, it is most often used to obtain a variety of information, which derives from various portals (e.g. news, TV schedules, prices of electricity, etc.).

These acquired information users most often read (web) or download as file containing this information (e.g., presentation sheets, images, etc.). Regardless of how users gain information, most World Wide Web data is based on the HTML language (Hyper Text Summary Markup Language). HTML enables very good presentation of data and information for end users through appropriate softwares (e.g. Internet Explorer). From the computer point of view information are messy, unstructured and, consequently, more difficult to manage for computer processing. In addition to large amounts of correct information available online, one can also find misleading or incorrect arguments. This is even more challenging, since incorrect information could hide the (new) knowledge.

For successful parting from good, useful data and information from web, garbage and foreground, it is necessary for unstructured data and information to transform in form that computer or manual processing will handle easier.

This paper presents our view on obtaining data and information from the World Wide Web and its recording in format based on modern technologies of Web 3.0. Chapter 2 presents different approaches to data processing and information from the World Wide Web that had a significant influence on the system development.

Chapter 3 presents a proposed approach on obtaining information from the World Wide Web and its record in a structured format suitable for further use. Chapter 4 presents the application of the proposed method as part of the existing expert system of Slovenian Energy Company Elektro energija d.o.o. Further, user perspective and architecture of the entire system is presented.

Chapter 5 describes the results of the experiment (ontology construction from the portal Elektro energija d.o.o.) considered from a technical and content point of view. We continue with the results from the experiment where we included generated ontology into the existing ontology of portal's expert system tested by end users.

Chapter 6 concludes the paper with objective discussion of advantages and disadvantages for proposed method, which also indicates further development of the proposed method.

II. Background and Related Work

Automatic data collecting from the World Wide Web is becoming increasingly important as the different data links (Links) reveal much more information than the average human internet user can detect. Tim Barnes-Lee, one of internet gurus, who introduced the idea of the semantic web, referred to as Web 3.0 [1], saw this potential. However, to create links between online data that is easy to read by computer, existing portals and associated web pages must be properly prepared. World Wide Web processing data could be roughly divided in two groups: namely mining structures of websites (Web Structure Mining), and mining content of web pages (Web Content Mining) [2], [3].

Both share some common methods like the well-known PageRank algorithm and HITS algorithm and search engines to index the content of web pages primarily use them both [2], [4], [5]. There are many methods and tools for building ontologies from web described in the literature. There are also many different approaches for constructing ontologies from web, but all of them base on similar web data mining methods:

- Alani and others present their system for automatic knowledge extraction from web documents. Based on VIPS, algorithm extracts data from web into ontology [6].
- Mo and others present their method of building Domain-Specific ontologies from web. Their method also rests on page segmentation and VIPS algorithm. Method extracts data from web and creates or updates existing ontology [7].
- Haav presents his method of automatic constructing of domain-specific ontologies using Natural Languages Processing (NLP) and Formal Concept Analysis (FCA) [8].
- Sanchez and Moreno also present their method of constructing domain-specific ontology for medical purpose on unsupervised approach [9].
- Karthikeyan and Karthikeyani present their method PROCEOL (Probabilistic Relation of Concept Extraction in Ontology Learning) of constructing ontology [10].

Researchers tackled the development of technologies for gaining and processing of data from the web differently. Some focused solely on data acquisition, while others added semantic connections to the obtained data. There most important of the latter are:

- Toledo-Alvarado and others present a method for automatic construction of ontologies base from a collection of textual documents. Their procedure rests solely on data mining technologies. It is unique in the fact this is the only method where building does not rely on external sources like vocabularies [11].
- Gunasundari and Karthikeyan present their method for getting data from the Internet – based connections. Method detects page content (segment) that is in accordance with the number of separators, the ratio of the number of characters that are not part of the link, and the number of characters that are part of the link [12].
- Gawrysiak and others present a system Text- Onto - Miner (TOM), a system for analysing natural language (Natural Language Testing System). In this context, they developed and implemented a number of algorithms and approaches for semi- automatic construction of ontologies. Their approach uses text mining and natural language processing methods [13].
- Mehta and others in [14] describe an algorithm obtaining semantic structure of web documents, based on VIPS algorithm and Naive Bayes classifier's proprietary. It incorporates all segments belonging to the same class [14].

- John and Shajin Nargunam present their method of knowledge discovery of users' web usage data based on K-means algorithm [15].

III. Approach and Methodology

Our method rests on approaches [14] and [16], upgraded with some additional steps. The basic idea of the proposed algorithm is to build portal's ontology as close as possible to the one made by human expert. Our approach for automatic ontology building consists of four basic steps:

1. web page segmenting with VIPS algorithm,
2. correcting of web page segments with Data Mining Methods,
3. transforming of segments into ontology,
4. joining ontology and existing ontology.

In order to perform those steps, we have to select an initial portal's web page. First three steps are repeated for all web pages of the selected portal, while the fourth step is performed only once. The result is one ontology for the entire portal.

Below is a pseudo-code of the basic algorithm. The first input parameter of the function is ontology class (optional), the second one is instance of the class (optional), the third one is the ontology (also optional) and the last one is the URL of web page. All optional parameters are used only in the recursive calls of the method, especially where we jump to URL of another web page.

Generally, web pages have many URLs leading to web pages, which are not necessary, a part of the initial portal. The content of another portal is not necessary valid. For that reason, our method only is limited to initial portal.

```

AnalysePage(c, cI, Ontology, URL)
begin
    var V=GetVipsSegments(URL); //get segments of web page
    V.Classify(); //classification of segments
    CheckSegmentsTree(V); //arrange segments

    If c!=null And cI != null And Ontology!=null
    begin
        var b=null;
        var bI=null;
        var p=null;

        If GetClass(V.Class, Ontology)==null
            b= CreateClass(V.Class, Ontology);
        else
            b=GetClass(V.Class, Ontology);

        bI=CreateInstance(b, Ontology, V.Title);

        If GetProperty(V.Class.Property, Ontology)==null
            p= CreateProperty(segment.Class.Property, Ontology);
        else
            p= GetProperty(V.Class.Property, Ontology);

        CreatePropertyInstance(p, c, cI, b, bI, Ontology);
    end

    BuildOntology(V, Ontology); //make ontology
end

```

III.1. Web Page Segmentation

The HTML language normally used at web pages does not serve only for content presentation, but it shows details of individual web page structure. The basic structure of the web page is DOM (Document Object Model Summary) structure. The DOM structure is a tree structure, where every HTML tag in the page corresponds to a node in the DOM tree. Some predefined structural tags, containing important information [17], can structure segmented web page. Those tags are paragraph, table, ul, list, H1-H6 (heading), etc.

Unfortunately, retrieving data from the DOM structure of the web page is not a simple task. The HTML language is very flexible, powerful and robust, but some web pages are not built according to W3C HTML standard and therefore have irregular structure.

One of the most known and successful methods in this area is VIPS algorithm (Vision-Based Page Segmentation), developed and introduced by Microsoft researchers [17]. The task of the algorithm is to extract the semantic structure of the website in terms of its design in a human-like way. Figure 1 shows an example of web page segmentation using the VIPS algorithm [17].

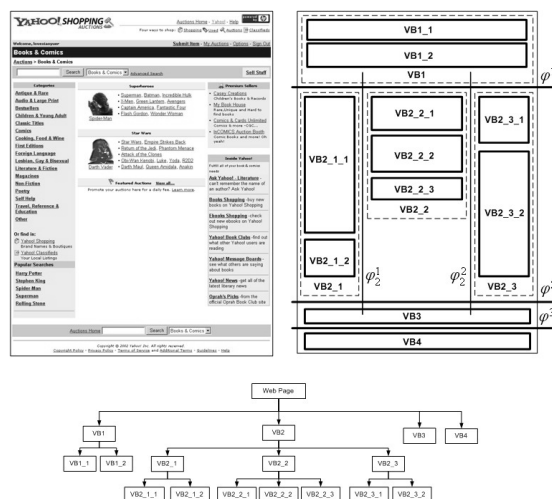


Fig. 1. Web page segmentation using VIPS algorithm

The algorithm is based on the assumption that each web page can be represented as triple, consisting of a plurality of segments, separators (horizontal or vertical lines dividing segments to each other) and relationships between segments. The aforementioned assumption applies to each segment, subsequently treated as a subpage of the website [5], [17].

The algorithm consists of three steps: extracting a segment, extracting separators and making a tree. The procedure is performed recursively for each segment, using a top-down principle. Thus, the algorithm's web page is divided into several major segments in a first step; the result is then added to the tree structure. In the second step, the procedure is recursively repeated in each segment. The process is repeated until no more segment divisions are possible [5], [17].

The result of the algorithm is a tree structure of segments, where each node contains a DOC (Degree of Coherence), which determines the compatibility of the segment based on visual perception.

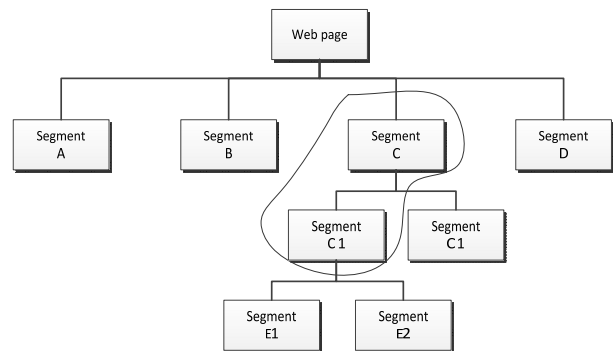


Fig. 2. Tree of segments

III.2. Aggregation and Classification of Web Page Segments

Aggregation and classification of web page segments is the most important step of our method. The Naïve Bayes algorithm classifies all segments, extracted by VIPS algorithm, into several pre-defined classes:

- *Header* – header of web page,
- *Menu* – horizontal or vertical navigation menu,
- *Footer* – footer of web page,
- *Content* – area with content of web page,
- *News* – blocks of news,
- *FAQ* – blocks of question and answers,
- *Module* – specific content not classified in News or FAQ class.

The prospect of segment classification must be at least 60% or more. Otherwise, the segment is marked as unclassified. Unclassified segments are excluded from tree structure therefore; knowledge worker is notified for classification by the system.

It is very important to reduce segments and join the same segments together for the final tree of segments. Reducing segments is not an easy task and few rules need to be considered during the integration:

- only segments that are directly related to the tree and are not at the same level should be joined,
- the content (title) of joined segments must be the same,
- the sub-content has no title.

The content of the segments should extract titles of segments and bold text; records are then compared with each other.

If the parent segment has title or bold text and it's sub-segment does not have a title or bold text, then the content is considered the same and segments can be united. Otherwise, the segments are not joined. When joined segment is dependent, it is moved one level higher. Fig. 3 shows the joining of two segments (Area C and Area C1).

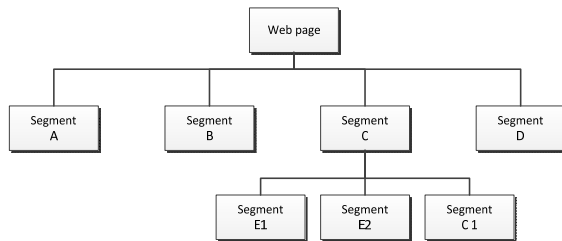


Fig. 3. Final segments tree

The algorithm for combining the two segments is as follows:

CheckSegmentsTree(V)

```

begin
  If V == null
    return;
  else
    begin
      For each segment  $\in V.Segments$ 
        begin
          MergeSegments(V, segment);
          CheckTree(segment);
        end
      end
    end
  end
end

```

s1 – parent segment

s2 – segment, which is subordinate to the segment *s1*

MergeSegments(s1, s2)

```

begin
  If  $s1.Class == s2.Class$  And  $IsEqualContent(s1, s2)$ 
    begin
      For each segment  $\in s2.Segments$ 
        begin
           $s1.AddSegment(segment)$ ;
        end
      end
       $s1.RemoveSegment(s2)$ ;
    end
  end
end

```

III.3. Transformation of Constructed Web Structure into Ontology

Formerly described step uses data mining techniques for properly mapping tree segments into ontology.

Method creates an ontology based on segments of web pages, presented in a tree structure and properly classified and minimized (joined segments). Ontology construction is rather straightforward: for each classified segment, we create a class in the ontology, which is the same as classification. Every candidate for ontology class is verified if it already exists in ontology. If so, it is not created again.

The instance of class is a content of segment. Every instance has data properties of Dublin Core metadata standard. Dublin Core standard consists of eight properties of and their content is extracted from the content of the segment.

Data property "Title" is filled by the title of segment, data property "Description" is reserved for the content segment, while in the data property "Type" is filled by the value of the name of the class ontology.

TABLE I
DUBLIN CORE METADATA ELEMENTS

Metadata elements	Description
Title	The name given to the resource.
Subject	The topic of the content of the resource.
Description	An account of the content of the resource.
Type	The nature or genre of the content of the resource.
Source	A Reference to a resource from which the present resource is derived.
Language	A language of the intellectual content of the resource.
Creator	An entity primarily responsible for making the content of the resource.
Date	A date associated with an event in the life cycle of the resource.

Data property "Subject" is filled by keywords (key words are extracted from news, keywords of the page ...).

If keywords exist and segment is classified as "News", then subclasses are created with the name of the key words.

Following these appropriate instances, classes are added into subclass. For example, news titled "Nove storitve za nove odjemalce" contains the keyword "Vklopi prihranek". According to the previously presented procedure news will be classified in Class "news". But because news contains the keyword "Vklopi prihranek", a subclass "Vklopi prihranek" will be created as a subclass of class "News". An instance of this news will be added to class "Vklopi prihranek". Links between classes or their instances are created when instances of each class in the ontology are added. They are determined according to the tree structure of segments, specifically the segment - sub segment. The whole process of adding instances to the ontology is implemented recursively.

There is also one exception if content of segments contains hyperlinks which have URLs of pages of this portal. In that case hyperlinks are extracted and all steps of this method are performed for each hyperlink. When the ontology construction is complete, it is saved in the OWL language. Created ontology is included in the existing ontology with relevant links created between the classes of existing ontologies, and in the ontology, created automatically. Links between classes are created according to the names of the classes, which easily compare different titles. If the titles of two classes match, a link between classes is created.

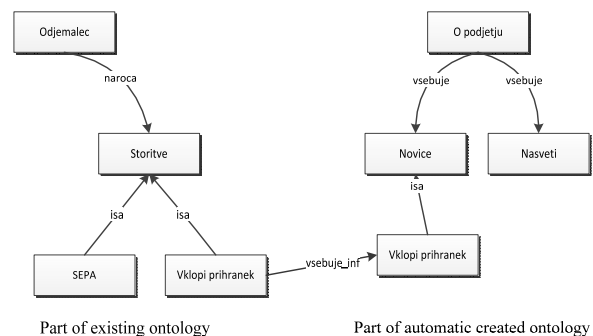


Fig. 4. Linking two ontologies

IV. Practical Application

The proposed method is implemented as part of an existing expert system for “Customer Energy Consumption” based on semantic web technologies and included in self-service portal “Moja energija” of Slovenian energy Provider Company Elektro energija d.o.o. Main task of the expert system is to provide easily understandable and user friendly information to users. Key parameters used when preparing an advice for user are as follows:

- Type of Electric Meter (one-tariff, dual-tariff),
- Ordered services (e.g. dynamic-tariff),
- Report of electricity consumption on a monthly basis,
- Supply of electricity (e.g. costumer electricity supply),
- The number of children and number of adults in the household,
- How to use electricity (basic needs, domestic heating, and hot water).

Parameters to the relevant class execute the classification of users (very economical user, economical user, average user, wasteful user and very wasteful user).

The same process of user classification also executes the comparison of electricity consumption with other users with similar parameters. There are two results of that process. The first one is information about user classification and the second one is proposal for reducing costs and consumption of electricity.

For example if user belongs to the “wasteful user” classification class, links to the articles about saving electricity appear. Fig. 5 shows expert system in practice.

IV.1. Implementation Details

Process described in section 3 is implemented using C# .NET programming language as a programming library (dll). Few additional APIs we also implemented:

- API of VIPS algorithm – PageAnalyzer.dll [18]
- API of Data Mining methods – WEKA [19]
- API of Jena.NET – for working with ontologies and reasoning [20].

Implemented programming libraries were included as a part of existing expert system and run as a scheduler job of portal platform of Elektro energija d.o.o. company portal. Fig. 6 shows architecture of whole system.

The architecture of system is classical 3-tier system architecture. Implemented as DotNetNuke portal module, there is presentation logic on the user interface layer containing user controls and other logic for user interface.

Business logic with classes for different operations such as calculations and so on is very important for correct operation of whole system and is therefore incorporated on its second level.

Expert System library and automatic building ontology library (implementation of proposed method) which using Jena.NET library for operation is included on the same level.

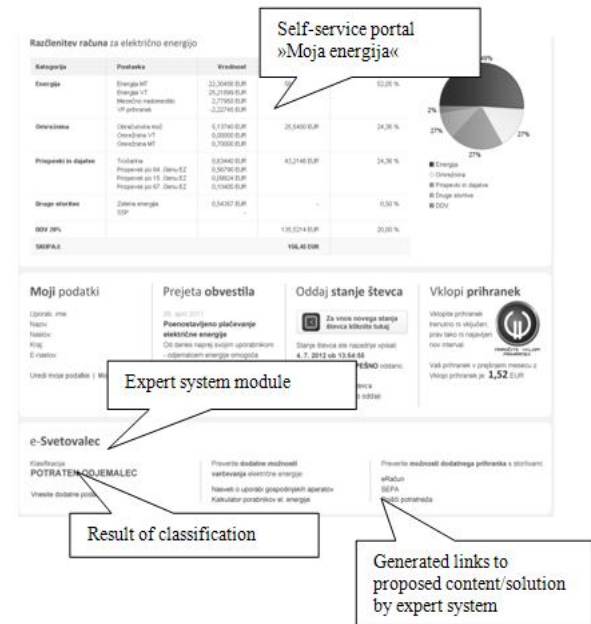


Fig. 5. Expert system in practice

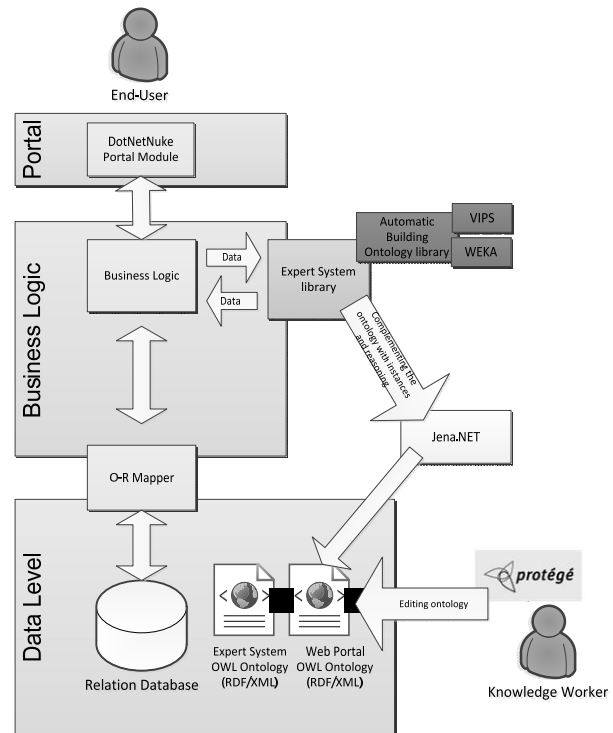


Fig. 6. Architecture of system

On the bottom there is a data layer containing relation databases of portal, ontologies of portal generated by our method and ontology of expert system. Knowledge worker using the Protégé software [21] can edit those ontologies.

IV.2. Experiment

We conducted the experiment of the proposed method two parts. The first one was an automatic building

ontology of Elektro energija d.o.o. company portal, which contains 112 different public web pages in Slovenian and English language. For specific reasons (e.g., users of portal are Slovenian, ontology of existing expert system is for Slovenian users) we focused on Slovenian version of portal. The first goal was to create the technically and semantically (meaning and content) correct ontology of the portal. The second part was including portal ontology into existing ontology of expert system for “Customer Energy Consumption”. Goal was to verify the correct integration of created ontology with existing ontology, which is technically and semantically correct. In both cases, the meaning (content) was extremely important.

V. Results

The experiment resulted in constructed ontology of the entire portal that should be included and linked into the existing ontology of the expert system. Fig. 7 presents a part of automatically created ontology from portal Elektro energija d.o.o.

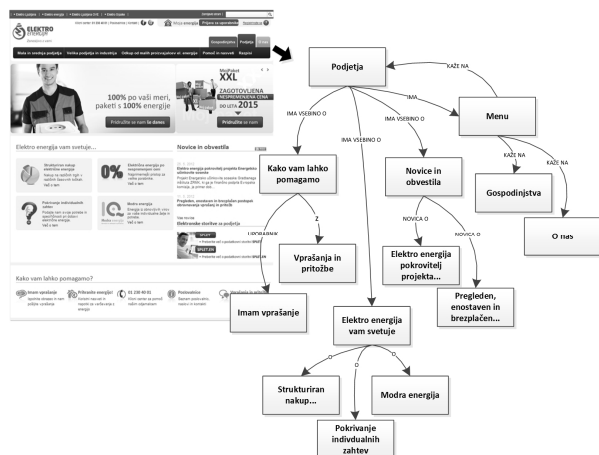


Fig. 7. Part of automatic constructed ontology

To authenticate the results we first performed a manual check of automatically constructed ontology using the PROTÉGÉ tools and confirmed the accordance of constructed ontology with the predefined algorithm based on the structure of each web page. In the next step, we examined the applicability of constructed ontology with questions obtained from the FAQ (Frequently Asked Questions Summary) business websites Elektro energija d.o.o. (Fig. 8). FAQ questions not only describe most common users' issues but also reply with already known answers. Thus, FAQ is an important piece of information for verifying the ontology's accuracy. Questions were converted to SPARQL query language (Simple Protocol and RDF Query Language) [22] which inquiries and collects information from the ontology. Results and answers at the FAQ website were then closely compared.

Where the answer is part of the question in the news or article, the results were positive. Accurately, a link to an article with answer is result of the question.



Fig. 8. Web page of FAQ of Elektro energija d.o.o.

When system does not find the answer to the question, the result is “answer not found”.

TABLE II
RESULTS OF EXPERIMENT IN NUMBERS – PART ONE

Experiment	No. of all items	No. of correct items	No. of not correct items
Web page ontology constructions	112	89	23
FAQ questions	46	39	7

The second part of the verification focused on the integration of constructed ontologies into an existing ontology expert system.

We focused on the proper integration as well as on the convenience in the expert system where the main task of constructed ontology is consulting the customers. Twenty-five (25) employees of Elektro energija d.o.o. tested the expert system with an extended ontology. Seventeen (17) users obtained useful information; four (4) users got adequate information, while the rest of the information obtained did not satisfy the users.

TABLE III
RESULTS OF EXPERIMENT IN NUMBERS – PART TWO

Number of users who got expected results	NUMBER OF WHO GOT ADEQUATE INFORMATION	Number of who were not satisfied with results
17	4	4

Users also gave feedback of employability of the expert system (see Table IV). Since testing users are experts in the energy field, it is to conclude the inclusion of constructed ontology and its application is successful.

VI. Discussion

In the first part of the experiment we focused on utility verification of the developed ontologies using the platform and indicated some deficiencies.

TABLE IV
GROUPED USERS FEEDBACK

Advantages	Disadvantages
Knowledge base (ontology) is very similar to knowledge base made by knowledge worker.	User interface is not very useful.
Results of expert system are useful; users get answers on their questions.	In case of insufficient information and classification of customer, results could not be correct.
Other applications could use the system (e.g. help center, CRM).	Knowledge worker should control knowledge base.

Despite rather successful experiment and many correct answers to most questions, they were only of average quality. For example, question “*How do we find the average daily consumption?*” was answered through the ontology with “*The answer to your question is written in the article covering the individual requirements*”, that is comparable with results of conventional search engine (e.g. Google).

One could find reason for such answers in the very construction of the ontology that focuses only on building segments of a single web page. The content of individual segments, possibly one of the most important carriers of knowledge, remains insufficiently treated with semantic technologies. Further development of the platform focuses towards the creation of ontology for the entire portal, semantic processing of individual segments and ultimately the inclusion of pilot solutions in the context of e-services that are available to the users of the portal Elektro energija d.o.o.

In the second part of experiment, we focused on combining two ontologies. Linking classes of two different ontologies could not be executed every time (for example due to missing class in both ontologies with the same name). However, it is always possible to remove mistakes manually by a knowledge worker.

Due to great amount of falsely distorted information on the web, we focused on the ontology constructed from a chosen portal by human and not from computer e.g. hyperlinks in content directed to another portal.

However, there are some limits of the proposed approach:

- The proposed method is restricted to the predefined domain.
- In the process of VIPS segmentation, there is no knowledge extraction from block content. Only block classification is implemented with data mining method.
- Files are not included in the segmentation process.
- There is no multi-language support.
- There is no separation between correct and wrong information on the web.
- Including constructed ontology into existing ontology is based on the class name matching method.

VII. Conclusion

In this paper, we introduced a method of automatic

building of ontology from the selected portal based on different technologies such as VIPS, data mining and semantic web technologies. The process was performed and analysed for the specific case of building the ontology at the Elektro energija d.o.o. portal. Ontology created was analysed from a technical and content point of view, and included into an existing expert system. As a result, we confirmed the proposed approach of automatic building ontology as correct. We noted deficiencies with automatic inclusion into the existing ontology expert system.

In the future work we will address further limits of the proposed method:

- knowledge extraction from blocks content at VIPS segmentation process,
- knowledge extraction from files,
- multi-language support,
- incorporating constructed ontology into the existing ontology considering meanings of ontology classes and block content.

References

- [1] T. Berners-Lee, Putting the Web back into Semantic Web, *ISWC2005 Keynote*, <http://www.w3.org/2005/Talks/1110-iswc-tbl>, Accessed on 19th Februar 2013.
- [2] B. Liu, *Web Data Mining Exploring Hyperlinks, Context and usage data*, Second edition, (Springer 2011).
- [3] Ambika, M., Latha, K., Web mining: The demystification of multifarious aspects, (2014) *International Review on Computers and Software (IRECOS)*, 9 (1), pp. 135-141.
- [4] Vijayadeepa, V., Ghosh, D.K., Sem-rank: A page rank algorithm based on semantic relevancy for efficient web search, (2013) *International Review on Computers and Software (IRECOS)*, 8 (11), pp. 2642-2647.
- [5] J. Han, M. Kamber, *Data Mining Concept and Techniques*, Second Edition, (Elsevier 2006).
- [6] H. Alani, S. Kim, D. E. Millard, M. J. Weal, P. H. Lewis, W. Hall, N. Shadbolt, Automatic Ontology-based Knowledge Extraction from web documents, *IEEE Intelligent Systems*, vol. 18 n. 1., January/February 2003, pp. 14 - 21.
- [7] W. Mo, P. Wang, H. Song, J. Zhao, X. Zhang, Learning Domain-Specific Ontologies from the Web, *Linked Data and Knowledge Graph Communications in Computer and Information Science*, vol. 406, 2013, pp 132 - 146.
- [8] H.-M. Haav, Learning Ontologies for Domain-Specific Information Retrieval, *Knowledge-Based Information Retrieval and Filtering from the Web, The Springer International Series in Engineering and Computer Science*, vol. 746, 2003, pp 285 - 300.
- [9] D. Sánchez, A. Moreno, Learning Medical Ontologies from the Web, *Knowledge Management for Health Care Procedures, Lecture Notes in Computer Science*, vol. 4924, 2008, pp 32 - 45.
- [10] Karthikeyan, K., Karthikeyani, V., PROCEOL: Probabilistic relational of concept extraction in ontology learning, (2014) *International Review on Computers and Software (IRECOS)*, 9 (4), pp. 716-726.
- [11] J. I. Toledo-Alvarado, A. Guzman-Arenas, G. L. Martinez-Luna, Automatic building of an ontology from a corpus of text documents using data mining tools, *Journal of applied research and technology*, vol 10 n. 3, Mexico dic. 2012.
- [12] R. Gunasundari, S. Karthikeyan, A Study of content Extraction From Web Pages Based on link, *International Journal of Data Mining & Knowledge Management Process (IJDMP)*, vol. 2 n. 3, May 2012.
- [13] P. Gawrysiak, G. Protaziuk, H. Rybinski, Experiments with semi automated ontology building using text onto miner, *Proceedings of the International IIS 08 Conference*, June 16-18, 2008, Zakopane, Poland.

- [14] R. R. Mehta, P. Mitra, H. Karnick, Extracting Semantic Structure of Web Documents Using Content and Visual Information, *The 14th International Conference on World Wide Web (WWW 2005)*, May 10-14, 2005, Chiba, Japan.
- [15] John, J.M., Shajin Nargunam, A., Similarity distance based clustering framework for aggregation of web usage data, (2013) *International Review on Computers and Software (IRECOS)*, 8 (1), pp. 287-295.
- [16] N. Khanaswneh, O. Samarah, S. Al-Omari, S. Conrad, Vision-based Presentation Modeling of Web Application: A reverse engineering approach, *Journal of emerging technologies in web intelligence*, vol. 4 Issue 2, May 2012, pp 134.
- [17] D. Cai, S. Yu, J.-R. Wen, W.-Y. Ma, VIPS: a Vision-based Page Segmentation Algorithm, *Microsoft Research Technical Report, MSR-TR-2003-79*, November 2003.
- [18] VIPS: a Vision based Page Segmentation Algorithm, <http://www.cad.zju.edu.cn/home/dengcai/VIPS/VIPS.html> Accessed on 18th Februar 2013.
- [19] WEKA 3 - Data Mining with Open Source Machine Learning Software in Java, <http://www.cs.waikato.ac.nz/ml/weka> Accessed on 19th Februar 2013.
- [20] Jena.NET - Flexible .NET port of the Jena semantic web toolkit, http://semanticweb.org/wiki/Jena_.NET Accessed on 19th Februar 2013.
- [21] Protege: A free, open-source ontology editor and framework for building intelligent systems, <http://protege.stanford.edu> Accessed on 19th Februar 2013.
- [22] SPARQL Query Language for RDF, <http://www.w3.org/TR/rdf-sparql-query> Accessed on 19th Februar 2013.

Authors' information

¹Kivi Com d.o.o., Kidričeva 3a, SI-2380 Slovenj Gradec.
E-mail: ambroz@kivi.si

²Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova 17, SI-2000 Maribor.
E-mail: milan.zorman@um.si



Ambrož Stropnik, born in Slovenia in 1982. He graduated from Computer Science in 2006 at the University of Maribor, Slovenia. From 2006 – 2010 he worked at the Slovenian white goods company Gorenje Group as an IT developer. Ambrož Stropnik is currently employed as head of research and development in the Kivi Com d.o.o. His interest areas of research include expert systems, semantic technologies and information retrieval from web and usage those technologies in industry. He is also a Ph.D. student at University of Maribor.



Milan Zorman, born in Slovenj Gradec, Slovenija in 1971, obtained his B.Sc. degree in 1995, M.Sc. degree in 1999 and Ph.D. degree in 2001, all from computer science and all at the University of Maribor, Slovenia. He is a professor of Computer Science at the Faculty of Electrical Engineering and Computer Science, Faculty of Health Sciences and Faculty of Medicine at the University of Maribor. As a researcher he is active in the fields of artificial intelligence, machine learning, medical and nursing informatics, data mining, knowledge extraction, hybrid intelligent systems and complex systems. Prof. Zorman participated in numerous international projects with companies, universities and other institutions from all over the world, mostly in the fields of medical and nursing informatics, complex systems, data mining, knowledge extraction, and technology transfer.

Proof of Retrievability Using Elliptic Curve Digital Signature in Cloud Computing

Sumathi D.¹, S. Kathik²

Abstract – Cloud computing is a general term that includes everything which is delivered as a service. Cloud computing can be identified as a technology which leverages economic growth and improvisation of usage of resources on demand. In the dream of computing world, service providers turn their infrastructure world into an environment where organizations can do their business by running their applications over the internet. Apart from that the cloud storage services enables the end users to enjoy their facility by moving their entire data onto the cloud. This facilitates the technology to focus on various security issues and challenges that has to be resolved.

This paper discusses in detail about deploying a protocol by using Elliptical Curve digital signature algorithm. This proposed protocol is established to enforce a security proof by checking for the data integrity. The protocol ensures the security against third party auditor and security against the untrusted server. This paper proves its integrity of the data through a formal analysis. **Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords: Cloud Computing, Elliptical Curve Cryptography, Galois Field, Authentication, Provisioning

I. Introduction

Cloud computing provides customers the view of computing resource which are available anywhere, anytime on request. It is a natural outcome of the latest trend. Cloud computing which belongs to the world of 5th generation technology provides a highly scalable distributed computing platform which provides computing resources on demand ([1]-[21]).

The advantage of cloud computing is that the enterprises pave a way in cutting costs without reworking on infrastructures, software. The idea of cloud computing is to provide all kinds of services to the customers in leveraging capital investments paid for resources. It is defined as a style of computing which consists of information services, a stack of software services, centralized infrastructure that is provided for users over the internet.

Many challenges arises in this cutting edge technology is being deployed on a large enterprise scale. Underlying challenges in the cloud computing are:

1. Monitoring
2. Resource scheduling
3. Security
4. Storage
5. Multi tenancy

This paper discusses about one of the main concerns of cloud computing.

Storing data in the cloud has become a recent trend. Consumers avail the facility of cloud technology by moving their data onto the cloud.

II. Technical Preliminaries

Moving data on to the cloud facilitates the end users by freeing up hard disk space, eliminates the cost of maintenance because the data centre handles it. Security for data storage is considered to be the most important quality of service. Storing data in cloud provider's place relieves the burden of the user since it provides comparably low cost scalable independent platform.

Threats may occur when the data is uploaded by the user. Security considerations have to be concentrated for the data when it is in transit and as well as in rest in cloud storage providers place. Cloud service provider must be responsible for the client's data. Security for the client data can be referred as Data confidentiality, data integrity and availability.

The client's data must be stored confidentially so that any unauthorized user must not access the data and the data must not be known even to the cloud service provider. Hence a strong encryption algorithm must be enforced to secure the data. To ensure the data integrity certain traditional methods can be applied. Cloud service providers take responsible for the consumer's data backup, data storage, data replication and purchasing of additional storage devices. Users must focus on security and privacy risks in spite of the benefits of cloud storage services provided by the service providers. Users (Data Owners) do not retain the local copy of the data uploaded by them. Hence the control of the data is being considered as the risk. The data might be tampered either by the untrusted server or by any intruder.

To avoid all these certain proactive measures must be taken to ensure the security of the data. The biggest hurdle is concern over the data integrity.

II.1. Related Work

Traditional cryptographic methods for data integrity based on hash functions cannot work on the outsourced data. It is not a practical solution for data validation by downloading the entire content of the file since it leads to time consumption and expensive transaction. Various traditional approaches like challenge response protocol have been done to check for data validation in cloud storage. Certain researchers have focused on remote data possession checking schemes to prove the data integrity through public auditability.

Remote data possession schemes can be categorized in two types namely provable data possession (PDP) and proof of retrievability (POR). The difference between PDP and POR is that POR checks the possession of data and it can recover data in case of a failure and it detects the integrity of the data if it is tampered even below a threshold level. In [1] Ateniese et al., developed a scheme for provable data possession (PDP) model which utilizes the RSA-based homomorphic authenticators for auditing outsourced data and the scheme implies sampling a few random blocks of the file.

This scheme does not support public auditability and the number of audits is limited to a bound. Yan Xiangtao, Li Yifa [2] proposed a new remote data integrity checking scheme for cloud storage which uses a "proof of retrievability" (POR) model and give a more meticulous proof of their scheme. To ensure both possession and retrievability of remote data files they have used spot-checking and (ECC) error correcting codes. This scheme lacks the support of dynamic updates and public audit ability. Erway et al [3] proposed the first scheme in dynamic PDP scheme. They used skip list data structure to facilitate data possession with dynamic support. The efficiency is considered to be in question since the search time to insert a block and finding a particular block is longer than in trees.

In [4] Wang et al. implements a dynamic architecture for public checking. The challenge-response protocol is used to determine the data correctness and possible errors are located.

However, the inefficient performance greatly affects the practical application of their scheme. In [5] Zhuo Hao, Sheng Zhong and Nenghai Yu recommend Privacy-Preserving Remote Data Integrity Checking Protocol with Data Dynamics and Public Verifiability.

The drawback is that there is no clear mapping relationship between the data and the tags. Data dynamics is supported only at the block level. Hence when the user tries to change a piece of data cost in updating goes high since it is block level representation.

Data integrity must be verified by the data owner frequently through the auditing task. Frequent auditing task leads to time consumption and expensive.

Various researchers suggest the solutions by introducing the third party auditor (TPA). As a result of this, the cloud storage providers and data owners may choose third party auditor (TPA) for periodically auditing the data outsourced by the data owner. Consider a storage system which consists of cloud service provider that operates cloud server, a client who uploads a file onto the cloud and a third party auditor (TPA) who computes and verifies for data integrity. The client stores their data in the cloud server without taking a copy of it.

If the client wants to check for the data integrity, it is of critical importance that the server must ensure for data integrity. If the cloud server modifies any piece of data, the client must be able to discern it. Data must be kept private against the third party verifier.

Ari Juels and Burton S. Kaliski [6] proposed a scheme by introducing special scheme blocks called sentinels among the data blocks for proof of retrievability. To prove for data integrity verification the sentinels have to be verified by the verifier. Maintaining the sentinels at the data owner side leads to the storage overhead when the thin clients are used. At the server side along with the sentinels error correcting codes are also stored. The drawback in this scheme is that the cost of storage is high. Sravan Kumar R and Ashutosh Saxena [7] encrypt only a few bits of data per data blocks thus reducing the computational cost.

They introduce a meta data verification scheme for integrity verification. This might not be suitable for all applications and files of large size. Zhuo Hao, Sheng Zhong, Member, IEEE, and Nenghai Yu, Member, IEEE [5] devised a new remote data integrity checking protocol which involves data dynamics at block level and it supports public verifiability without the help of third party auditor. Even if any third party auditor is used the privacy of data is maintained. Information about the data is not leaked to the third party auditor. Mihir R. Gohel and Bhavesh N. Gohil [13] investigated a data integrity checking protocol that supports public verifiability ensures that the storage at the client side is minimal and it is beneficial for thin clients.

III. Problem Statement

A cloud storage system consists of a data owner (DO) and a cloud server (CS) under the control of cloud service provider (CSP). We use digital signatures to verify the data integrity. Data Owner (DO) an entity who uploads a file or an archive on to the cloud. Cloud server (CS) an entity who stores the data. CSP controls the cloud server. In this cloud computing paradigm the data owner (DO) stores their file in the cloud server without retaining a copy of it. The critical importance in outsourcing data is that the client has to verify the data for integrity verification. When the server modifies the data the client must be able to detect it. We propose a remote data integrity checking protocol by using digital signature algorithm. Data integrity is checked either by downloading the entire content of the file or meta data of

the file.

Accessing the entire file and checking for integrity verification leads to I/O cost and time constraints. Hence to overcome these drawbacks we devise a scheme to verify data integrity by computing the digest value (M_d) for the whole data. The digest value sent by the server at the client side is compared with the original digest value created by the data owner before uploading the file. If both the values are same it reports that the data has not been modified by the server or any intruder.

III.1. Problem Formulation

We consider a file 'f' of size 'm' is divided into blocks 'b' of equal lengths 'l' where $f = f_1, f_2, f_3, \dots, f_b$ and $b = |f|/l$. We propose a remote data integrity checking protocol. This protocol ensures the privacy of data against the third party verifiers. Computations can be either done by third party verifiers or by the data owner.

The following functions are implemented in this protocol:

1. Key generation process
2. Signing process by the data owner (DO)
3. Signature verification process by the verifier. Verifier may be neither the third party auditor (TPA) nor the data owner (DO).

Data owner (DO) before uploading the file performs the following functions. Both the private and public key is generated by the data owner.

III.2. ECDSA

Our paper defines methods for signature generation and signature verification using Elliptical Curve Digital Signature Algorithm (ECDSA). ECDSA is a variant of RSA and DSA that operates on elliptic curves.

Key pair is related with a particular set of domain parameters $D = (q, FR, a, b, G, n, h)$. Data owner must ensure that the domain parameters are valid. Data owner does the following process to generate key pair.

Key generation process

Select an elliptic curve defined over a field representation. The field can be either finite field F_p or a base point whose elements are represented with respect to a polynomial. Domain parameters consists of:

1. Field size $q = p$ an odd prime.
2. An indication (FR) to represent the field elements F_q
3. Two field elements 'a' and 'b' which define the equation for the elliptic curve E in F_q
Case 1 $p = 2 \quad y^2 + xy = x^3 + ax^2 + b$
Case 2 $p > 2 \quad y^2 = x^3 + ax^2 + b$
4. A bit string of length at least 160 bits called seed value which is an optional if an elliptic curve generated is verifiably at random.
5. A point G (x_g, y_g) of prime order.
6. The order 'n' of point 'G' with $n > 2^{160}$ and $n > 4 \sqrt{q}$
7. Cofactor $h = \# E(F_q) / n$

Data Owner Key pair generation:

1. Select a point 'G' on the elliptic curve.
2. A random unique unpredictable integer 'i' $\in [1, n-1]$ is selected.
3. Do the computation as $Q = i * G$
4. The private key is denoted as 'i' and public key is denoted as 'Q'.

III.3. Signature Generation Process

The signatory (Data Owner) has to sign the message. To sign the message data owner has to do the following procedure. Data Owner is the signatory of the message.

To sign a message data owner must first create a message digest (M_d) with the help of hash function. The domain parameters $D = (q, FR, a, b, G, n, h)$ and the associated public key, private key pair (Q, i) is used:

1. An unique random integer 'k' is selected such that $1 \leq k \leq n-1$.
2. Calculate $K * G = (x_1, x_2)$ and $u = x_1 \mod n$
3. If $u = 0$ then go to step 1.
4. else compute $k^{-1} \mod n$.
5. Compute $M_d = \text{SHA-1}(f)$. where secured hash algorithm is used.
6. Compute $v = k^{-1} (M_d + i * u) \mod n$
7. If $v = 0$ then go to step 1
8. Data Owner signature is denoted as (u,v).

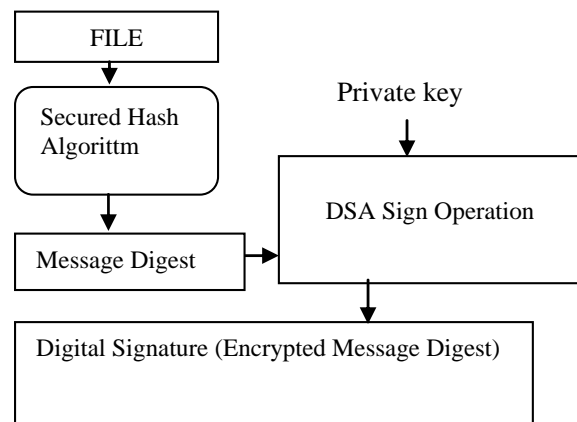


Fig. 1. Signing process by the data owner

A First a message digest (MD) is created at the sender side. A message digest is a short summary of the message that is going to be transmitted from the data owner side to the cloud server side. Hashing algorithm is used to create a message digest. The hashing algorithm ensures the data integrity by generating a hash value when any piece of the data is changed. The Message digest is encrypted with the private key created by the data owner and the encrypted message digest is formed, which is called as digital signature (DS). The digital signature which is formed is sent to the cloud server.

The data owner encrypts the file and uploads onto the cloud provider storage space along with the digital signature.

III.4. Signature Verification by the Verifier

To ensure the correctness of the data in the cloud server the data owner issues a request to send the signature along with the timestamp computed by the server. It is easy to prove the data integrity by completely downloading the entire data from the remote server.

But downloading the large amounts of data just for checking data integrity is a waste of communication bandwidth. Hence the data owner issues a request (R_i) for the digital signature (u_i, v_i) for the data stored by the owner. We propose a verification process that can be either done by the third party auditor (TPA) or by the data owner.

When the verifier is the third party auditor (TPA) this protocol make certain that no private information about the data is leaked. Based on the request issued by the data owner (DO) or third party auditor (TPA) the message digest (M_d) is computed and it is given to the corresponding entity who has issued the request. The cloud service provider computes the message digest (M_d) and digital signature (u, v) for the corresponding file and send to the verifier to check for data integrity. The verifier can be either the data owner (DO) or third party auditor (TPA). If the verifier is the data owner the cloud service provider (CSP) sends the message digest value (M_d) of the corresponding data.

Case 1: Signature verification by the data owner (DO)

Data Owner on request to the cloud service provider (CSP) gets the message digest value for data integrity verification. Data Owner's computation time is very less when compared with the third party auditor (TPA). From the signature sent by the data owner, DO must extract 'u' and 'v' separately.

Data owner must compute:

$$y = k^{-1} (M_d + i * u) \bmod n$$

If $y = u$ then it is proved that data has not been modified. Hence the data integrity verification can be done by the data owner at the minimal cost.

If the verifier is the third party auditor (TPA) the cloud service provider (CSP) must send the digital signature (u_i, v_i) and the message digest value (M_d) for the corresponding data. Data owner initiates the process of signature verification by the third party auditor by issuing the public key (Q) to the TPA.

Case 2: Signature verification by the third party auditor (TPA)

1. Data owner has to send the Public key (Q) to the third party auditor (TPA)
2. On request by the third party auditor (TPA) the cloud service provider (CSP) computes the message digest (M_d) and digital signature (u_i, v_i) of the corresponding data.
3. Third party auditor (TPA) computes the following computations

- i. Compute $z = v^{-1} \bmod n$.
- ii. Compute $r_1 = M_d * z \bmod n$
- iii. Calculate $r_2 = u * z \bmod n$
- iv. Construct a third point on the elliptic curve
- v. by addition operator
- vi. $r_1 * P + r_2 * Q = (x_0, y_0)$
- vii. Find $s = x_0 \bmod n$
- viii. If $s = u$ then it shows that the data has not been modified and it proves for data integrity.

IV. Theoretical and Experimental Analysis

In this section, theoretical analysis of the modified version of ECDSA is been done.

Theorem 1: Proof for signature verification

The signature received by the third party auditor (TPA) is denoted as (u, v) on a data. If a signature (u, v) on a message was certainly generated by an entity data owner (DO), then compute:

$$v = k^{-1} (M_d + i * u) \bmod n$$

Rearranging gives:

$$\begin{aligned} kG &= v^{-1} (M_d + i * u) G \bmod n = \\ &= v^{-1} M_d G + i * u * v^{-1} G \bmod n = \\ &= M_d * zG + z * Q * v \bmod n = \\ &= u * G + v * Q \bmod n \end{aligned}$$

Thus $u * G + v * Q = (u + v * i) G = kG$, and so $s = u$ as required.

Security Analysis in this protocol

The proposed protocol ensures data integrity. Also this protocol proves its correctness in the sense that the server and third party verifier can pass the verification of the data as long as both of them are honest. The server can pass the verification process successfully only when the server has not modified any piece of the data. The third party verifier can prove himself when the verifier reports any malfunction that is found.

The proposed protocol is also proved to be private against the third party verifier. The proposed protocol deals with the two security requirements: security against the server and security and privacy against the third party auditor. First we prove the security against the server. Consider two entities in this scenario.

One entity is the data owner and the other entity being an adversary that stands for the untrusted server.

Definition 1: Security against the server.

Consider a scenario in which the data owner has stored the data in the server and needs to check for its

integrity by posting a query for the corresponding message digest (M_d) of the file (F).

Consider a game between the antagonist and the challenger. The antagonist is considered to be the one who might steal the data and the challenger is the prover.

Here the prover is considered to be the server. The following functions are carried by the entities:

1. $Gen(k, G) \rightarrow (Q, i)$ This function is executed by the challenger.
2. $Sign(F, d, k, n) \rightarrow (u, v)$ This function is done by the challenger.
3. The antagonist might compute the message digest (M_d) of the corresponding file.
4. The challenger must compute a value:

$$y = k^{-1} (M_d + i * u) \bmod n$$

5. The challenger must run the function $verify(y, u)$. If it returns FALSE then the antagonist won the game and thus prove the security against the server.

Definition 2: Privacy against Third-Party Verifier

The third party verifier does not get any other information except the message digest (M_d). This protocol is proved strongly private against the third party verifier.

Proof:

Assumptions:

Simulator (S) for the view of the verifier is constructed. The input to the verifier includes $t = (Pk, Md, (r, s))$ where Pk is public key, M_d is the Message digest and r, s is digital signature.

The output of the simulator is computationally indistinguishable with the view of the verifier:

$$Adv_{X, Y(A)} = |Pr[A(X) = 1] - Pr[A(Y) = 1]|$$

The output of A can be arbitrary, but we deduce the output as {1} indicating that the adversary A is satisfied with the client and if any other output indicates that the adversary's test results are not satisfied. Under the semi honest model the output of this function will return the value as {TRUE}.

The simulator consists of the following steps:

Step 1: The simulator computes the value:

$$z = s^{-1} \bmod n$$

Step 2: Using the Message digest value the simulator computes $r1 = M_d * z \bmod n$ and $r2 = u * z \bmod n$

Step 3: Using a third point on the elliptic curve (p, q) and an addition operator $r1 * p + r2 * q = (x_0, y_0)$

Step 4: Now compute $s = X_0 \bmod n$.

Step 5: The verifier output is denoted as $\{t, (s)\}$

The verifier's view during the protocol execution is denoted as P_v :

$$View_{P_v} = \{t, (s)\} \quad (1)$$

The output of the simulator is denoted as $\{t, s\}$. This output is identically distributed with (u, v) . If the output obtained by the simulator gets matched and identical with the view's output. The simulator's output and View P_v output have identical distributions which are computationally indistinguishable.

Here the verifier cannot get any information from the messages received except its input and output.

Complexity Analysis and Experimental Results

Communication, Computation and storage cost:

The communication, computation and storage cost of the data owner, third party verifier and cloud service provider are analyzed in this paper. The cost of ECDSA signature generation requires one elliptic curve point multiplication (ECPM), one integer inversion, two integer multiplications and SHA-1 invocation on the message. But the verification cost of signature done by the data owner is less compared to traditional ECDSA signature verification. Traditional ECDSA signature verification [10] requires an additional elliptic curve point multiplication (ECPM). In our proposed auditing activity the verification cost is still very less.

Verification cost[14] requires one integer inversion, one integer multiplication and hash function of message.

T_{ecpm} -> Time complexity for executing the elliptic curve point multiplication.

T_{inv} -> Time complexity for executing integer inversion.

T_{mul} -> Time complexity for executing integer multiplication.

Scenario 1: Communication between cloud service provider and data owner without the third party verifier.

The cloud service provider sends the Message digest value of the requested data (M_d) for verification to the data owner. Both the public key and private key is known to the data owner, the cost of the proving for data integrity is very less. Data owner computes a value by using the received Message digest value and the part of the signature created for that corresponding digest value.

That value is compared with the remaining part of the signature. If the function returns Success then it proves that data has not been modified and it proves for data integrity. The time complexity calculated in this function is $T_{inv} + T_{mul} + T_{h(x)}$. The communication cost between the data owner and the cloud service provider can be calculated as follows:

Step 1: The data owner sends a query to the service provider.

Step 2: The cloud service provider sends the Message digest (M_d) which is of 'l' bits to the data owner.

Therefore the communication cost is '1' bits.

Scenario 2: Communication between the data owner, server and the third party verifier (TPA). The proposed protocol analyzes the time complexity, communication cost of client, server and third party verifier separately since this protocol supports public audit ability.

Data owner calculates and send the public key to TPA.

The time complexity is $2T_{inv} + 2T_{ecpm} + T_{h(x)}$

The communication cost can be calculated as follows:

Step 1: The data owner sends the public key to the third party auditor (TPA) for which the data integrity to be proved. Here the length of the public key can be 'k' bits.

Step 2: The third party auditor receives the request for proof of data integrity. Corresponding to that, the data owner sends the public key to the TPA. The TPA in turn issues the request for the Message digest to the service provider. On receiving the request the service provider computes the digest value and sends to the TPA.

Therefore the communication cost is calculated as:

$$|k| + |n| + c \quad (2)$$

k - length of public key in bits

n - length of the digest value in bits

c - refers to the Boolean value {1,0}. one bit.

The storage cost of data owner, cloud service provider and third party verifier are given below (Table I).

TABLE I
STORAGE COST OF DATA OWNER

Data Owner	Cloud service provider	TPA
$ i + n + (u,v) $	$ n $	$ k $

From the above table we can see that the storage cost of data owner is $i+n$ where 'i' refers to the private key, 'n' refers to the message digest and cost of storing digital signature.

In the case of TPA, the cost of storing public key is ignored since it is used for temporary computation and there is no need to store any values.

V. Conclusion

This paper briefly describes the proof of data integrity by devising a protocol using digital signature algorithm.

The security and correctness of the data is proved. This protocol proves the security against the untrusted servers and privacy against the third party verifiers. We have analyzed the time complexity, computation and storage cost of cloud service provider, data owner and third party verifier and through theoretical analysis it is proved to be an efficient one.

References

[1] Ateniese, G., Burns, R.C., Curtmola, R., Herring, J., Kissner, L.,

Peterson, Z.N.J., Song, D.X., 2007. Provable data possession at untrusted stores. In: Proceedings of the 2007 ACM Conference on Computer and Communications Security, CCS 2007, pp. 598–609.

[2] Yan Xiangtao, Li Yifa Information engineering university, Zhengzhou, China, taoexcellen@163.com. "A new remote data integrity checking scheme for cloud storage".

[3] Erway, C.C., Küpc, ü, A., Papamanthou, C., Tamassia, R., 2009. Dynamic provable data possession. In: Proceedings of the 2009 ACM Conference on Computer and Communications Security, CCS 2009, pp. 213–222.

[4] C.Wang, Q. Wang, K. Ren, and W. Lou, "Ensuring Data Storage Security in Cloud Computing," Proc. 17th Int'l Workshop Quality of Service (IWQoS '09), 2009.

[5] Zhuo Hao, Sheng Zhong, Member, IEEE, and Nenghai Yu, Member, IEEE, "A Privacy-Preserving Remote Data Integrity Checking Protocol with Data Dynamics and Public Verifiability" IEEE transactions On Knowledge and Data Engineering, Vol. 23, No. 9, September 2011".

[6] A. Juels and B.S. Kaliski Jr., "Pors: Proofs of Retrievability for Large Files," Proc. 14th ACM Conf. Computer and Comm. Security (CCS '07), pp. 584-597, 2007.

[7] Sravan Kumar, R.; Saxena, A., "Data integrity proofs in cloud storage," *Communication Systems and Networks (COMSNETS)*, 2011 *Third International Conference on*, vol., no., pp.1,4, 4-8 Jan. 2011. doi: 10.1109/COMSNETS.2011.5716422

[8] Hung-Zih Liao, Yuan-Yuan Shen, "On the Elliptic Curve Digital Signature Algorithm", Tunghai, Science Vol. 8: 109126 July, 2006

[9] Z. Hao, S. Zhong, and N. Yu, "A Privacy-Preserving Remote Data Integrity Checking Protocol with Data Dynamics and Public Verifiability," Technical Report 2010-11, SUNY Buffalo CSE Dept., <http://www.cse.buffalo.edu/tech-reports/2010-11.pdf>, 2010.

[10] Reza R. Farashahi, Hongfeng Wu, Chang-An Zhao, Efficient Arithmetic on Elliptic Curves over Fields of Characteristic Three, Lecture Notes in Computer Science Volume 7707, 2013, pp 135-148.

[11] Francesc Sebe´, Josep Domingo-Ferrer, Senior Member, IEEE, Antoni Martí´nez-Balleste´, Yves Deswarte, Member, IEEE, and Jean-Jacques Quisquater, Member, IEEE, "Efficient Remote Data Possession Checking in Critical Information Infrastructures" IEEE transactions On Knowledge and Data Engineering, Vol. 20, no. 8, August 2008.

[12] G. Ateniese, R. Burns, R. Curtmola, J. Herring, O. Khan, L. Kissner, Z. Peterson, and D. Song, "Remote Data Checking Using Provable Data Possession", ACM Transactions on Information and System Security, Vol. 14, No. 1, Article 12, Publication date: May 2011.

[13] Mihir R. Gohel, Bhavesh N. Gohil "A New Data Integrity Checking Protocol with Public Verifiability in Cloud Storage" Trust Management VI IFIP Advances in Information and Communication Technology Volume 374, 2012, pp 240-246

[14] A. Jurisic, A. Menezes, "Elliptic Curves and Cryptography", 2003, <http://www.certicom.com/whitepapers>.

[15] Jena, R.K., Mahanti, P.K., Computing in the cloud: Concept and trends, (2011) *International Review on Computers and Software (IRECOS)*, 6 (1), pp. 1-10.

[16] Jena, R.K., Green cloud computing: Need of the hour, (2012) *International Review on Computers and Software (IRECOS)*, 7 (1), pp. 45-52.

[17] Narayanan, G.G., Raja Ranganathan, S., Karthik, S., An efficient user revocation and encryption methods for secure multi-owner data sharing to dynamic groups in the cloud, (2014) *International Review on Computers and Software (IRECOS)*, 9 (5), pp. 825-831.

[18] Mercy Gnana Rani, A., Marimuthu, A., Kavitha, A., Artificial fish swarm load balancing and job migration task with overloading detection in cloud computing environments, (2014) *International Review on Computers and Software (IRECOS)*, 9 (4), pp. 727-734.

[19] Jawabrah, M., Aboud, M., Enhancing performance of partitioned database in cloud computing environment, (2014) *International Review on Computers and Software (IRECOS)*, 9 (2), pp. 355-364.

- [20] Priyadharshini, M., Baskaran, R., Balaji, N., Saleem Basha, M.S., Analysis on countering XML-based attacks in web services, (2013) *International Review on Computers and Software (IRECOS)*, 8 (9), pp. 2197-2204.
- [21] Alnouri, M., El-Koutly, R., Traffic protection as a service in MPLS cloud network, (2013) *International Review on Computers and Software (IRECOS)*, 8 (7), pp. 1644-1649.

Authors' information

¹PPG Institute of Technology, Coimbatore.

²SNS College of Technology, Coimbatore.



Mrs. **D. Sumathi** is Presently Associate Professor, Department of Computer Science and Engineering, PPG Institute of Technology, Affiliated to Anna University Chennai, Tamil Nadu. She received the B.E degree from Bharathiar University in 1994 and M.E degree from Sathyabama University in 2006, Chennai and pursuing her Ph.D degree in Anna

University, Chennai. Her Research interests include Cloud computing, Network Security and Theoretical Foundations. Mrs.D.Sumathi published papers in international journal and conference papers and has been involved many international conferences as Technical Chair and tutorial presenter. She is a active member of ISTE.



Dr. **S. Karthik** is presently Professor & Dean, Department of Computer Science & Engineering, SNS College of Technology, affiliated to Anna University- Chennai, Tamilnadu, India. . M.E Degree and Ph.D Degree from Anna University Chennai. His research interests include network security, web services and wireless systems. In particular, he

is currently working in a research group developing new Internet security architectures and active defense systems against DDoS attacks. Dr.S.Karthik published papers in international journal and conference papers and has been involved many international conferences as Technical Chair and tutorial presenter. He is a active member of IEEE, ISTE and Indian Computer Society.

Hybridization of ABC and PSO for Optimal Rule Extraction from Knowledge Discovery Database

K. Jayavani, G. M. Kadhar Nawaz

Abstract – Knowledge discovery in database (KDD) has provided a large interest in statistics, machine learning, and artificial intelligence (AI). It is a challenging task for mining the comprehensive and informative knowledge in such complex data by using the existing methods. The challenges come from many aspects, for instance, the traditional methods usually discover homogeneous features from a single source of data while it is not effective to mine for patterns combining components from multiple data sources. It is often very costly and sometimes impossible to join multiple data sources into a single data set for pattern mining. In order to extract knowledge from different datasets, we will propose a hybrid mining technique. The knowledge extraction can be done by association rule mining with the combination of Artificial Bee Colony optimization algorithm (ABC) and Particle swarm optimization (PSO). The main aim of this hybridization is to extract the optimal rules from the association rules for further classification. The accuracy will be checked in terms of optimal rule obtained from the hybridization. The best position for moving the particle will be updated by using ABC algorithm. **Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords: KDD, PSO, ABC, AI

Nomenclature

A, B	Frequent item sets
r_i	Association rules
FIT _i	Fitness values
k	Solution of i
Φ	Random number of range [-1, 1]
$s_{i,j}$	Solution obtained from the employed bee phase
V []	particle velocity
Present []	current particle
Rand ()	random number between (0, 1)
C ₁ , C ₂	learning factors
Cycle	denotes an iterative value
P _i	Probability values for the solutions

I. Introduction

Data Mining and Knowledge Discovery in Databases (KDD) are rapidly evolving areas of research that are at the intersection of several disciplines, including statistics, databases, pattern recognition/AI, visualization, and high-performance and parallel computing [1]-[21]. In this paper, which is intended to be strictly a companion reference to the invited talk, and not a presentation of new technical contributions, we outline the basic notions in this area and show how data mining techniques can play an important role in the analysis of scientific data sets [12]. The majority of the available data mining approaches search for patterns in a single data table [21].

One of the important tasks in data mining is classification. In classification, there is a target variable which is partitioned into predefined groups or classes.

The classification system takes labeled data instances and generates a model that determines the target variable of new data instances [13]. The discovered knowledge is usually represented in the form of if-then prediction rules, which have the advantage of being a high level, symbolic knowledge representation, contributing to the comprehensibility of the discovered knowledge [16]. The discovered rules can be evaluated according to several criteria, such as the degree of confidence in the prediction, classification accuracy rate on unknown-class instances, and interpretability. Accuracy and interpretability are two important criteria in data mining [11] [8].

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both [14]. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified [15]. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases [10] [17]. Classification plays an important role in data mining and the need for building classifiers across multiple databases is driven by applications from various domains [3].

Classification is about grouping data items into classes (categories) according to their properties (attribute values).

Classification is also called supervised classification, as opposed to the unsupervised classification (clustering). “Supervised” classification needs a training dataset to train (or configure) the classification model, a validation dataset to validate (or optimize) the configuration, and a test dataset to evaluate the performance of the trained model. Classification methods include, for example, decision trees, artificial neural networks (ANN), maximum likelihood estimation (MLE), linear discriminant function (LDF), support vector machines (SVM), nearest neighbor methods and case-based reasoning (CBR). Data mining adopted its techniques from many research areas including statistics, machine learning, association rules, neural networks, and so on [9] [18].

- (1) Association rules. Association rule generators are a powerful data mining technique used to search through an entire data set for rules revealing the nature and frequency of relationships or associations between data entities. The resulting associations can be used to filter the information for human analysis and possibly to define a prediction model based on observed behavior.
- (2) Artificial Neural Networks are recognized in the automatic learning framework as universal approximations, with massively parallel computing character and good generalization capabilities, but also as black boxes due to the difficulty to obtain insight into the relationship learned.
- (3) Statistical Techniques. These include linear regression, discriminate analysis, or statistical summarization.
- (4) Machine learning (ML) is the center of the data mining concept, due to its capability to gain physical insight into a problem, and participates directly in data selection and model search steps. To address problems like classification (crisp and fuzzy decision trees), regression (regression trees), time-dependent prediction (temporal trees), and the ML field is basically concerned with the automatic design if then rules similar to those used by human experts. Decision tree induction: the best known ML framework was found to be able to handle large-scale problems due to its computational efficiency, to provide interpretable results, and, in particular, able to identify the most representative attributes for a given task [19].

Successful Knowledge Discovery and Data Mining applications play an important role in data that have clearly grown to surpass raw human processing abilities.

The challenges facing advances in this field are formidable.

Some of these challenges include as follows: Develop new mining algorithms for classification, clustering, dependency analysis, and change and deviation detection that scale to large databases [20].

II. Literature Review

Chin-Ang Wua *et al.* [1] we presented an intelligent data warehouse mining approach incorporated with schema ontology, schema constraint ontology, domain ontology and user preference ontology, to improve the overall data warehouse mining process. The structures of these ontologies are illustrated and how they benefit the mining process is also demonstrated by examples utilizing rule mining. The paper also present a prototype multidimensional association mining system, which with intelligent assistance through the support of the ontologies, can help users build useful data mining models, prevent ineffective pattern generation, discover concept extended rules, and provide an active knowledge re-discovering mechanism.

Sumana Sharma *et al.* [2] discussed that Knowledge Discovery and Data Mining (KDDM) process models provide prescriptive guidance towards the execution of the end-to-end knowledge discovery process, i.e. such models prescribe how exactly each one of the tasks in a Data Mining project can be implemented. The paper has also presented the results of the rigorous evaluation of the Integrated Knowledge Discovery and Data Mining (IKDDM) process model and compares it to the CRISP-DM process model. Results of statistical tests showed that the IKDDM leads to more effective and efficient implementation of the knowledge discovery process.

Tahar Mehenni and Abdelouahab Moussaoui [3] proposed a classification approach across multiple heterogeneous relational databases for data mining. The paper has used a regression model for predicting the most useful links that will be connected to build a multi-relational decision tree, into a given a set of inter-related databases. Experiments performed on different real and synthetic databases were very satisfactory compared with previous classification approaches in multiple databases.

Haitao Gan *et al.* [4] proposed a semi-supervised learning framework which combines clustering and classification. Paper discussed that clustering analysis is a powerful knowledge-discovery tool and it may reveal the underlying data space structure from unlabeled data. Semi-supervised clustering is integrated into Self-training classification to help train a better classifier. The semi-supervised fuzzy c-means algorithm and support vector machines have used for clustering and classification, respectively. Experimental results on artificial and real data sets showed the advantages of the proposed framework.

Ya-Wen Chang Chien and Yen-Liang Chen [5] proposes a GA-based algorithm used to build an associative classifier that can discover trading rules from the numerical indicators. The experiment results showed that the proposed approach is an effective classification technique with high prediction accuracy and is highly competitive when compared with the data distribution method.

Rashedur M. Rahman and Fazle Rabbi Md. Hasan [6] used decision tree induction algorithm on Hospital Surveillance data to classify admitted patients according

to their critical condition. Three class labels, low, medium and high, are used to distinguish the criticality of the admitted patients. Several decision tree models are developed, evaluated, and compared with different performance metrics.

Finally an efficient classifier is developed to classify records and make decision/predictions on some input parameters. The models developed in this research could be helpful during epidemic when huge number of patients arrive daily.

Bilal Alatas [7] a novel computational method, which is robust and have less parameters than that of used in the literature, is intended to be developed inspiring from types and occurring of chemical reactions. The proposed method is named as artificial chemical reaction optimization algorithm, ACROA. In this study, one of the first applications of this method has been performed in classification rule discovery field of data mining and efficiency has been demonstrated.

III. Problem Identification and Proposed Methodology

Knowledge discovery in database (KDD) is an active area of research that resolves the non-trivial process of identifying valid, potentially useful, and ultimately understandable patterns in data. In simple terms, a pattern is actionable if the user can act upon it to her advantage.

Furthermore, actionable patterns can not only afford important grounds to business decision-makers for performing appropriate actions, but also deliver, expected outcomes to business. Association rule mining is a main method to produce patterns. However, as large numbers of association rules are often produced by association mining algorithm, sometimes it can be very difficult for decision makers to not only understand such rules, but also find them a useful source of knowledge to apply to the business processes. In other words, association rules can only provide limited knowledge for potential actions.

Therefore, there is a strong and challenging need to mine for more informative and comprehensive knowledge for decision-making in the real world.

A comprehensive and general approach for discovering informative knowledge in complex data is suggested. It is challenging to mine for comprehensive and informative knowledge in such complex data suited to real-life decision needs by using the existing methods.

The challenges come from many aspects, for instance, the traditional methods usually discover homogeneous features from a single source of data while it is not effective to mine for patterns combining components from multiple data sources. It is often very costly and sometimes impossible to join multiple data sources into a single data set for pattern mining.

In order to extract knowledge from different datasets, we will propose a hybrid mining technique. The knowledge extraction can be done by association rule mining with the combination of Artificial Bee Colony optimization algorithm (ABC) and Particle swarm

optimization (PSO). The main aim of this hybridization is to extract the optimal rules from the association rules for further classification. The accuracy will be checked in terms of optimal rule obtained from the hybridization.

The best position for moving the particle will be updated by using ABC algorithm.

IV. Knowledge Extraction Using Association Rule Mining

Apriori is a classic algorithm for learning association rules. Apriori algorithm is designed to operate on databases containing transactions. The Apriori Algorithm is an influential algorithm for mining frequent item sets for Boolean association rules. It was developed by Developed by Agrawal and Srikant 1994. It is an innovative way to find association rules on large scale, allowing implication outcomes that consist of more than one item. In association rule mining the input given is the database. There are two main steps in association rule mining. First the frequent item sets are generated using the minimum support value assigned. Second, the association rules are generated using the frequency item sets generated and the minimum confidence value assigned. A database is given as input to the association rule mining process. From the input database the number of transactions is calculated. For extracting association rules minimum values of support and confidence should be assigned. The item sets are extracted from the database. Each item is the member of set of candidate. The support values are calculated for the item sets separately. The support value is calculated using the formula:

$$Support(A \rightarrow B) = P(A \cup B) \quad (1)$$

where A and B = Frequent item sets.

The support values calculated for the separate item sets are compared with the minimum support value. The item sets with support value less than the minimum support value is eliminated. The remaining item sets are selected. Then the selected item sets were combined with the same item sets. Again the support value is calculated for the item sets and they are eliminated based on the support value. By the elimination and the pruning step the item set which is for generating association rules are found out. The confidence value can be found out by using the formula:

$$Confidence(A \rightarrow B) = \frac{P(A \cup B)}{P(A)} \quad (2)$$

where A and B = Frequent item sets.

IV.1. Pseudo Code for Association Rule Mining

C_k : Candidate item set of size k

L_k : frequent item set of size k
 $L_k = \{\text{frequent items}\};$
for ($k = 1; L_k \neq \emptyset; k++$) **do begin**
 C_{k+1} = candidates generated from L_k ;
 for each transaction t in database **do**
 increment the count of all candidates in C_{k+1}
 that are contained in t
 L_{k+1} = candidates in C_{k+1} with min_support
 end
return $C_k L_k$;

The frequent item set generated finally and the minimum confidence value is used to generate association rules. All the item sets generated were taken.

The item sets with the items in the frequent item set is identified. The confidence values for the selected item sets were calculated using formula (2). The item sets with confidence value greater than the minimum confidence value assigned are selected. The remaining item sets are rejected. The selected item sets are the association rules.

The association rules generated is given as input to the hybrid ABC and PSO algorithm for optimization.

IV.2. Optimization Using ABC and PSO Algorithm

ABC is an algorithm which is explained by Dervis Karaboga in 2005, inspired by the smart behavior of honey bees. The colony of artificial bees has three set of bees in ABC algorithm; they are employed bees, onlookers and scouts. A bee which is waiting on the dance area for making a choice to pick a association rule is called onlooker and a bee which goes to the association rules that is selected by the onlooker is called employed bee. The other type of bee is scout bee that carries out unsystematic search for discovering novel sources.

The position of the association rules denotes a realistic solution to the optimization issue and the value of a association rules related to the quality (fitness) of the associated solution, estimated by:

$$FIT_i = \frac{1}{1 + r_i} \quad (3)$$

where:

i = number of association rules

r = association rules

The collective intelligence of honey bee swarms consists of three components. They are Employed bees, Onlooker bees and Scout bees. There are two major behaviors.

(a) Association rules:

To select the association rules forager bee evaluates various properties. For simplicity one quality can be considered.

(b) Employed bees:

The employed bee is employed on a particular association rule. It shares the information about the particular association rules with other bees in the hive.

The information which is carried by the bee includes direction, Profitability and the distance.

(c) Unemployed bees:

The unemployed bees include both onlooker bees and the scout bees. The onlooker bee searches the association rules with the information given by the employed bees.

The scout bee searches the association rules randomly from the environment.

The main steps of ABC algorithm are:

C_k : Candidate item set of size k

L_k : frequent item set of size k

$L_k = \{\text{frequent items}\};$

for ($k = 1; L_k \neq \emptyset; k++$) **do begin**

C_{k+1} = candidates generated from L_k ;

for each transaction t in database **do**

 increment the count of all candidates in C_{k+1}
 that are contained in t

L_{k+1} = candidates in C_{k+1} with min_support

end

return $C_k L_k$;

In the ABC algorithm each cycle consists of three steps. Initial step involves sending the employed bee to find out the Association rules to evaluate their values then the Association rules were selected by the onlooker based on the information given by the employed bee then the scout bees was send to find the new Association rules. At the initialization stage a number of Association rules were determined by the bees and their values were calculated.

At the first step of the cycle the employed bees come in to the hive and share the information about the Association rules and their value information. The information is shared by the employed bees with the bees waiting in the dance area. The onlooker bees take the information about the Association rules. Then the employed bees travels to their respective Association rules which they have already visited and finds the neighboring Association rules in comparison through visual information.

At the second step of the cycle the onlooker bee selects the Association rules depending on the information given by the employed bees. If the optimization increases the probability of the association rules chosen also increases. When the onlooker bee arrives in the area as per the information given by the employed bee it chooses the neighboring association rules by comparing the values by visual information same as in employed bees. The new association rules were found by the bees on comparison of values based on the visual information.

At the third step when the association rules was taken by the bees' new association rules was found out by the bees. A scout bee randomly selects the new association rules and replaces the old association rules with the new one. The bees which has the fitness values as good enough is the result of this fitness. The detailed explanation of the ABC algorithm is as follows:

- Initialize the association rules of the solutions $s_{i,j}$.
- Calculate the population.
- Set $cycle = 1$; the cycle denotes an iterative value.

- Create a solution $u_{i,j}$ in the neighborhood of $s_{i,j}$ using the following formula:

$$u_{i,j} = s_{i,j} + \Phi_{i,j} (s_{i,j} - s_{k,j}) \quad (4)$$

where:

$k \rightarrow$ Solution of i

$\Phi \rightarrow$ Random number of range $[-1,1]$.

- Apply the greedy selection process amid $u_{i,j}$ and $s_{i,j}$ based on the fitness.
- Calculate the probability values P_i for the solutions $s_{i,j}$ using their fitness values based on the following formula:

$$P_i = \frac{FIT_i}{\sum_{i=1}^{SN} FIT_i} \quad (5)$$

- In order to estimate the fitness values of the solution we have used the following formula:

$$FIT_i = \begin{cases} \frac{1}{1+r_i}, & \text{if } r_i \geq 0 \\ 1 + \text{abs}(r_i), & \text{if } r_i \leq 0 \end{cases} \quad (6)$$

- Normalize the P_i values into $[0, 1]$.
- Create the novel solutions $u_{i,j}$ for the onlookers from the solutions s_i depending on P_i and calculate them.
- Apply the greedy selection procedure for the onlookers amid s_i and u_i based on fitness.
- Determine the abandoned solution (source), if exist, replace it with a novel unsystematically produced

solution s_i for the scout using the following equation:

$$s_{i,j} = \min_j + \text{rand}(0,1) \times (\max_j - \min_j) \quad (7)$$

- Memorize the optimum association rules position (solution) attained so far.
- Cycle=cycle+1
- Until, cycle=maximum cycle number.

IV.2.1. Pseudo-Code for ABC Algorithm

Require: Max_Cycles, Colony Size and Limit

- Initialize the Association rules
- Evaluate the Association rules
- Cycle=1
- **while** cycle \leq Max_cycles **do**
- Produce new solutions using employed bees
- Evaluate the new solutions and apply greedy selection process
- Calculate the probability values using fitness values
- Produce new solutions using onlooker bees
- Evaluate the new solutions and apply greedy selection process
- Produce new solutions for onlooker bees
- Apply Greedy selection process for onlooker bees
- Determine abandoned solutions and generate new solutions randomly using PSO in scout bee section
- Memorize the best solution found so far
- Cycle = Cycle + 1
- **end while**
- **return** best solution

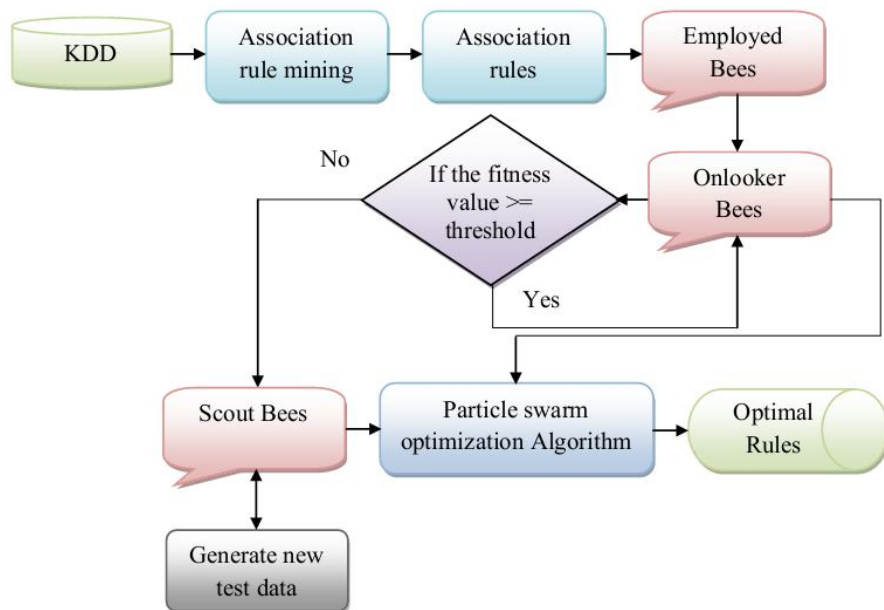


Fig. 1. Architecture of the proposed methodology

The ABC algorithm has several dimensional search spaces in which there are Employed bees and Onlookers bees. Both bees were categorized by their experience in identifying the association rules. The initial population is opted from the employed bee phase. The rules are possessed by the employed bee. The solution of the employed bee is altered in the onlooker bee stage based on the following formula:

$$u_{i,j} = s_{i,j} + \Phi_{i,j} (s_{i,j} - s_{k,j}) \quad (8)$$

where:

$s_{i,j}$ → Solution obtained from the employed bee phase

$\Phi_{i,j}$ → Randomly produced number of range [-1, 1]

k, j → Random indexes in the solution matrix of employed bee

A solution is created based on the formula and the solution is applied in the fitness function to obtain the fitness value. This process would last until the entire employed bee gets processed. The scout bee phase is the eventual stage of the ABC algorithm. This stage is implemented with PSO algorithm in order to find the optimal rule. The scout bee initiates the process by choosing the solution from the onlooker bee phase which poses the lowest fitness value. The onlooker bee phase generates diverse solution based on different $u_{i,j}$ values.

The solution with least fitness value is selected. In the scout bee section PSO optimization algorithm is included.

IV.3. PSO in Scout Bee Phase

The Particle swarm optimization (PSO) is a population based stochastic optimization technique. It was developed by Dr. Eberhart and Dr. Kennedy in 1995.

This technique was inspired by social behavior of bird flocking or fish schooling. PSO model constructed based on three ideas. They are:

- Evaluation
- Comparison
- Imitation

IV.3.1. Pseudo-Code for PSO Algorithm

- **For** each particle
- Initialize particle
- **END**
- **Do**
- **For** each particle
- Calculate fitness value
- If the fitness value is better than the best fitness value (pbest) in history
- Set current value as the new pbest
- **End**
- Choose the particle with the best fitness value of all the particles as the gbest

- **For** each particle
- Calculate particle velocity according Eqn. (9)
- Update particle position according Eqn. (10)
- **End**
- While maximum iterations or minimum error criteria is not attained

In this process the potential solution named as particles fly through the problem space by following the present optimum particles. Each particle maintains the record of its coordinates in the problem space which are related with the fitness of the particle. This value is referred to as pbest. Another best value is calculated which was tracked by the particle swarm optimizer, obtained at a distance of any particle in the neighbors of the particle. This location is referred to as lbest. If a particle considers all the population as its topological neighbors the best value is called global best which is also referred to as gbest. The particle swarm optimization concept consists of steps for changing the velocity towards the pbest and the lbest locations. Acceleration is weighted by a random term, with random numbers which was generated for acceleration towards the pbest and lbest locations. PSO is initialized by a group of random particles. It also searches for optima by updating generations. In Each Iteration, the values are updated using the two values such as pbest and lbest. The first value is the best value it has ever achieved and it is stored in the memory. It was named as pbest. Another best value tracked by the particle swarm optimizer, the location is called the gbest. When a particle takes part of the population as its topological neighbors, the best value is a local best and is called lbest:

$$V[] = V[] + C_1 * rand() * (pbest[] - present[]) + C_2 * rand() * (gbest[] - present[]) \quad (9)$$

$$present[] = present[] + V[] \quad (10)$$

where:

$V[]$ = particle velocity

$Present[]$ = current particle

$rand()$ = random number between (0, 1)

C_1, C_2 = learning factors

Usually $C_1 = C_2 = 2$.

In the scout bee section the rules with values less than threshold were taken and replaced. Here in scout bee section PSO was included and the optimal rule was generated.

V. Results and Discussion

In this section we have given the results of our proposed methodology. A Matlab program is written to extract the optimal rules. The input database used here was KDD cup 99.

KDDCUP'99 is the mostly widely used data set for the evaluation of the intrusion detection in systems.

KDD'99 has been the most widely used data set for the evaluation of anomaly detection methods.

This data set is prepared by Stolfo et al. and is built based on the data captured in DARPA'98 IDS evaluation program. The input database was subjected to association rule mining. Then association rules are generated. The sample association rules were given in the Table I.

TABLE I
TABULAR COLUMN FOR SAMPLE ASSOCIATION RULES

<u>SAMPLE ASSOCIATION RULES</u>
'HH'
'HLH'
'HLHH'
'HLHHH'
'HHHHHHHLHH'

The fitness values calculated in corresponding iterations for ABC, PSO and ABC-PSO were given and compared in Table II.

TABLE II
TABULAR COLUMN FOR COMPARISON OF FITNESS VALUES

ITERATION	FITNESS VALUES		
	ABC-PSO	ABC	PSO
10	55.3306	52.0867	53.1192
20	51.7188	50.2728	43.2605
30	51.5286	51.081	45.4033
40	49.0831	48.9424	44.4308
50	53.1716	51.6944	41.5618

From the above table it is clear that our proposed method has high fitness value when compared to the existing methods. It is proven that our proposed method is efficient than the existing methods. The fitness values in the y-axis with the corresponding iterations in the x-axis were plotted in the following Figs. 2-5. The graphs was plotted separately for ABC, PSO, ABC-PSO and for the comparison of all the three methods. From the graphical plot it was clear that the fitness of our proposed method was higher than the existing methods.

Table III represents the sample optimal rules and their fitness values generated by ABC, PSO and ABC-PSO.

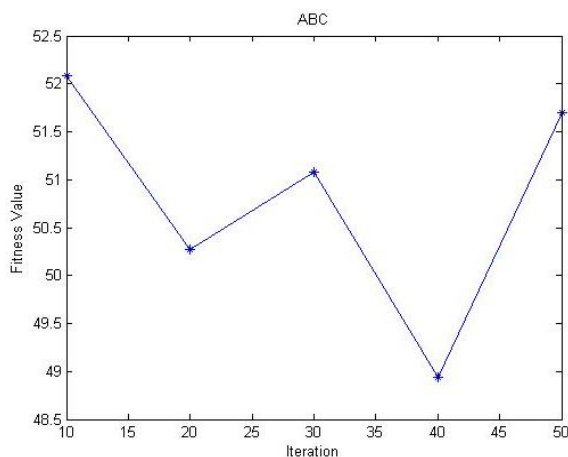


Fig. 2. Graphical plot for Iteration versus fitness value for ABC algorithm

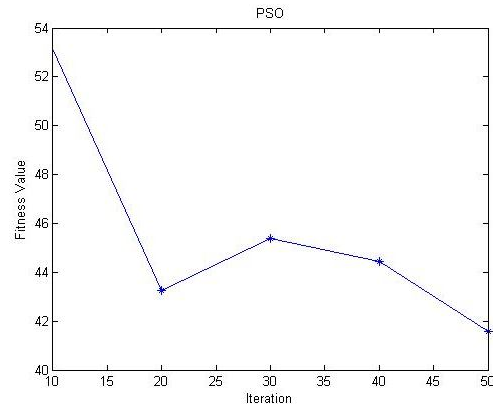


Fig. 3. Graphical plot for Iteration versus fitness value for PSO Algorithm

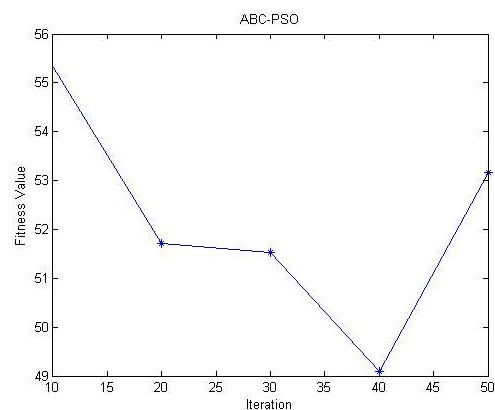


Fig. 4. Graphical plot for Iteration versus fitness value for ABC-PSO algorithm

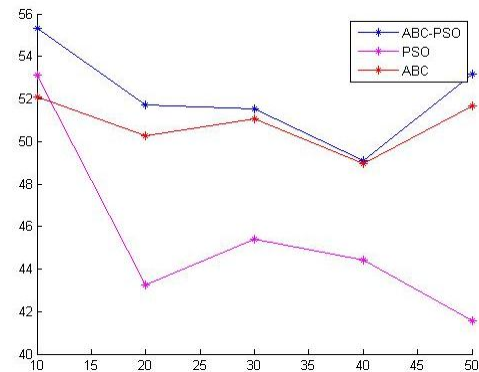


Fig. 5. Graphical plot for the comparison of the three methods

TABLE III
TABULAR COLUMN FOR OPTIMAL RULES GENERATED BY THE THREE METHODS

RULE NO.	OPTIMAL RULES		
	ABC-PSO (Fitness values)	PSO (Fitness values)	ABC (Fitness values)
1	LLHLHHHL (4.796452)	LHLL (4.408238)	HLH (4.021155)
2	LLHHHHHHHLHL (4.206341)	HLH (4.021155)	HLHHH (2.466897)
3	HLHLHLHL (5.496611)	LHLHLLLLL (5.211892)	LHLHL (5.050767)

From the table it is clear that the fitness values of the optimal rules generated by the ABC-PSO algorithm is higher than the fitness values of other optimal rules. Thus the rules generated using the ABC-PSO algorithm were highly optimal than the others.

VI. Conclusion

The ABC is a new search algorithm. Many approach like ant colonies have been successfully used in Data Mining. It is a challenging task for mining the comprehensive and informative knowledge in such complex data by using the existing methods. The challenges come from many aspects, for instance, the traditional methods usually discover homogeneous features from a single source of data while it is not effective to mine for patterns combining components from multiple data sources. It is often very costly and sometimes impossible to join multiple data sources into a single data set for pattern mining. To extract knowledge from different datasets, we have used a hybrid mining technique. The knowledge extraction was done by association rule mining with the combination of Artificial Bee Colony optimization algorithm (ABC) and Particle swarm optimization (PSO). Hybridization was used to extract the optimal rules from the association rules for further classification. The accuracy was checked in terms of optimal rule. The best position for moving the particle was updated by using ABC algorithm. In future work, we plan to compare the performance of the proposed ABC and PSO algorithm with other algorithms we can also plan to discover a new effort, by changing the parameter values of ABC and PSO.

References

- [1] Chin-Ang Wu, Wen-Yang Lin, Chang-Long Jiang, Chuan-Chun Wu, Toward intelligent data warehouse mining: An ontology-integrated approach for multi-dimensional association mining, *Expert Systems with Applications*, Volume 38, Issue 9, September 2011, Pages 11011-11023. <http://dx.doi.org/10.1016/j.eswa.2011.02.144>
- [2] Sumana Sharma, Kweku-Muata Osei-Bryson, George M. Kasper, "Evaluation of an integrated Knowledge Discovery and Data Mining process model", *Expert Systems with Applications*, Vol. 39, No. 13, pp. 11335–11348, 2012.
- [3] Tahar Mehenni and Abdelouahab Moussaoui, "Data mining from multiple heterogeneous relational databases using decision tree classification", *Pattern Recognition Letters*, Vol. 33, No. 13, pp. 1768–1775, 2012.
- [4] Haitao Gan, NongSang, RuiHuang, XiaojunTong and ZhipingDan, "Using clustering analysis to improve semi-supervised classification", *Neurocomputing*, Vol. 101, pp. 290–298, 2013.
- [5] Ya-Wen Chang Chien and Yen-Liang Chen, "Mining associative classification rules with stock trading data – A GA-based method", *Knowledge-Based Systems*, Vol. 23, No. 6, pp. 605–614, 2010.
- [6] Rashedur M. Rahman and Fazle Rabbi Md. Hasan, "Using and comparing different decision tree classification techniques for mining ICDDR,B Hospital Surveillance data", *Expert Systems with Applications: An International Journal*, Vol. 38, No. 9, pp. 11421-11436, 2011.
- [7] Bilal Alatas, "A novel chemistry based metaheuristic optimization method for mining of classification rules", *Expert Systems with Applications*, Vol. 39, No. 12, pp. 11080–11088, 2012.
- [8] Diansheng Guo, Jeremy Mennis, "Spatial data mining and geographic knowledge discovery-An introduction", *Computers, Environment and Urban Systems*, Vol. 33, No. 6, pp. 403–408, 2009.
- [9] Yi Feng, Zhaohui Wu., "Enhancing Reliability throughout Knowledge Discovery Process", *Proceedings of ICDM Workshops on Data Mining*, pp.754-758, 2006.
- [10] Qi Luo, "Advancing Knowledge Discovery and Data Mining," *Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on*, vol., no., pp.3.5, 23-24 Jan. 2008. doi: 10.1109/WKDD.2008.153.
- [11] Hamid Mohamadi, Jafar Habibi, Mohammad Saniee Abadeh, and Hamid Saadi, "Data mining with a simulated annealing based fuzzy classification system", *Pattern Recognition*, Vol. 41, No. 5, pp. 1824 – 1833, 2008.
- [12] "Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases", In *Proceedings of Ninth International Conference on Scientific and Statistical Database Management* Microsoft Research, pp. 2-11, 1997.
- [13] M. Pazzani, S. Mani, and W.R. Shankle, "Comprehensible Knowledge-Discovery in Databases," *Proceeding of 19th Annual Conf. Cognitive Science Soc.*1997, pp. 596–601.
- [14] R. Brachman, T. Khabaza, W. Kloesgen, G. Piatetsky-Shapiro, and E. Simoudis, *Industrial "Applications of Data Mining and Knowledge Discovery"*, *Communications of ACM*, vol. 39, no. 11.1996.
- [15] *Communications of The ACM*, special issue on Data Mining, vol. 39, no. 11.
- [16] "Data warehousing and knowledge discovery: a chronological view of research challenges", In *Proceedings of the 7th international conference on Data Warehousing and Knowledge Discovery*, Springer-Verlag Berlin, Heidelberg, pp. 530-535, 2005.
- [17] Hamid R. Nemat, David M. Steiger, Lakshmi S. Iyer, and Richard T. Herschel, "Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing", *Decision Support Systems*, Vol. 33, pp. 143– 161, 2002.
- [18] Hsu-Hao Tsai, "Global data mining: An empirical study of current trends, future forecasts and technology diffusions", *Expert Systems with Applications*, Vol. 39, pp. 8172–8181, 2012.
- [19] Shu-Hsien Liao, Pei-Hui Chu, and Pei-Yuan Hsiao, "Data mining techniques and applications – A decade review from 2000 to 2011", *Expert Systems with Applications*, Vol. 39, pp. 11303–11311, 2012.
- [20] Ming-Syan Chen, "Data mining: an overview from a database perspective", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6, pp. 866-883, 1996.
- [21] Ravi Sankar, M., Prem Chand, P., A technique to mine the multi-relational patterns using relational tree and a tree pattern mining algorithm, (2013) *International Review on Computers and Software (IRECOS)*, 8 (4), pp. 1053-1061.

Authors' information



K. Jayavani obtained her Bachelor's degree in Electronics Science from Bharathiyar University in 1999. Then she obtained her Master's degree in Computer Applications in 2003 and also completed her Master of Philosophy (M.Phil.) in Computer Science in 2007 from Periyar University. Now she is pursuing Ph.D. (Doctor of Philosophy) in Manonmanium Sundaranar University, Tirunelveli. Her research interests are in Data Mining and Warehousing. Specializations include Wireless application protocol, Networking and Operating Systems. Currently, she is working as Assistant Professor at Sri Vijay College Of Arts and Science, Dharmapuri.

{0, 1, 3}-NAF Representation and Algorithms for Lightweight Elliptic Curve Cryptosystem in Lopez Dahab Model

Sharifah M. Y., Rozi Nor Haizan N., Jamilah D., Zaitun M.

Abstract – Elliptic curve scalar multiplications is the most time-consuming and costly operation in elliptic curve cryptosystem. The scalar multiplication involves computation of $Q = kP$ where k is a scalar multiplier, and P and Q are points on an elliptic curve. This computation can be improved by reducing the Hamming weight of the scalar multiplier k . The Hamming weight of k represents the number of nonzero digits in the scalar multiplier. This paper proposes a new scalar representation in non-adjacent form (NAF) using the digits 0, 1 and 3. This paper also proposes an algorithm for converting from a binary to {0,1,3}-NAF representation. Comparative analysis between the proposed NAF and the traditional NAF with digit {-1,0,1} is carried out. At average case, the proposed {0,1,3}-NAF representation has a lower Hamming weight than the traditional NAF. In our analysis, we use the {0,1,3}-NAF representation in the scalar multiplication operation. The average number of point addition operations in the scalar multiplication is considerably reduced compared to the addition-subtraction scalar multiplication algorithm. **Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords: Lightweight Elliptic Curve Cryptosystem, Scalar Multiplication, Hamming Weight, Lopez Dahab Model, Non-Adjacent Form

Nomenclature

A	point addition
D	point doubling
ECC	elliptic curve cryptosystem
ECDLP	elliptic curve discrete logarithm problem
GF	Galois Field
LD	Lopez Dahab
l	length of binary k
PKC	public key cryptosystem
M	field multiplication
NAF	non-adjacent form
RSA	Rivest, Shamir and Adleman
SKC	symmetric key cryptosystem
S	Field Squaring
T	point tripling

I. Introduction

Secure data transmission through Internet technology is vital. Hence, cryptosystem can be used to transmit secure data, where the encryption and decryption of data can be done before transmission through the Internet ([1]-[30]). Currently, there are two different types of cryptosystems: the symmetric key cryptosystem (SKC) and the public key cryptosystem (PKC). In this paper, we focus on the PKC for reason of stronger security features in real-world implementation, specifically an elliptic curve cryptosystem (ECC) ([27]-[30]).

The ECC is a public key cryptosystem proposed by [1] which is based on the arithmetic of the elliptic curve.

The security of the ECC is based on the difficulty of solving the elliptic curve discrete logarithm problem (ECDLP). Additionally, the bigger size of the keys increases the difficulty of the problem [2].

The appropriate key size for implementation depends on a few aspects such as the duration of time in which the information is required to be kept secure, the computational resources that are available to the attackers, and progressive activities in the cryptanalysis area [3].

The ECC has the advantage of having a smaller key size with equivalent security level with the Rivest, Shamir and Adleman (RSA) cryptosystem. The 163-bit ECC has comparable security level with 1024-bit RSA.

The small key size indicates that the ECC needs lower processing power, bandwidth, storage and hardware requirements than the RSA. For the RSA cryptosystem, the bit-length for secure RSA has increased over recent years and has put a heavier processing load on its application. In contrast, the ECC saves a lot of processing time especially for electronic commerce sites that conduct large numbers of secure transactions [4].

Furthermore, the ECC features are most suitable for constrained devices like pagers, PDAs, smartphones and smart cards. In the literature, the ECC research motivation is the search for efficient scalar representation, scalar multiplication algorithms and lightweight implementation.

This paper provides an overview of elliptic curve cryptosystems in Section 2 and reviews the related work in Section 3. Section 4 outlines the proposed methods for

developing a new scalar representation in non-adjacent form (NAF) using the digits 0, 1 and 3. A conversion algorithm for scalar k is also presented. The results are discussed in Section 5 and the conclusion is presented in Section 6.

II. Elliptic Curve Cryptosystem (ECC)

Practically, in cryptography, the variables and coefficients of the elliptic curve are restricted to elements in a finite field [5]. Generally, two families of curves are used in ECC: prime curves and binary curves. This paper focuses on binary curves, particularly, the elliptic curve over binary field $GF(2^n)$. The elliptic curve over a binary field is attractive for hardware implementation [6].

Traditionally, elliptic curve coordinates are affine coordinates and are represented by two variables, x and y .

The scalar multiplication in affine coordinates involves expensive inversion operations. To reduce the cost of the inversion operation, most ECC implementations use projective coordinates such as the Lopez Dahab (LD) coordinates or Jacobian coordinates [7]. In this research, we focus on the LD coordinates. These coordinates were proposed by Lopez and Dahab in 1998 and are in the form (X, Y, Z) where $(Z \neq 0)$. The LD coordinates correspond to the affine point $(x, y) = (X/Z, Y/Z^2)$. The ECC in LD model has been implemented in [6]-[8].

Scalar multiplication is the main operation and the most resource consuming operation in ECC implementation. Given an elliptic curve point P and a large integer k , the scalar multiplication is the computation of $Q = kP$ where P and Q are points on the elliptic curve, Q is a public key and k is a private key. In this research, kP is computed in LD coordinates and the result is converted to affine coordinates before returning the output.

In the literature, several scalar multiplication algorithms have been proposed such as the double-addition approach and addition-subtraction algorithm.

This operation is used in key generation, encryption and decryption modules. In ECC implementation, k is normally a large number and for better security, k must be greater than 163 bits. In [9], 85% of the ECC running time is for scalar multiplication computation.

II.1. Elliptic Curve Arithmetic in Lopez Dahab Model

The LD projective equation is as follows:

$$Y^2 + XYZ = X^3Z + aX^2Z^2 + bZ^4 \quad (1)$$

The triple $(X_1:Y_1:Z_1)$ represents the affine point $(X_1/Z_1, Y_1/Z_1^2)$ when $Z_1 \neq 0$. The point at infinity P_∞ corresponds to $(1:0:0)$, while the negative of $(X_1:Y_1:Z_1)$ is $(X:XZ + Y:Z)$.

Scalar multiplication involves point operations and field operations. Examples of point operations include point addition, doubling, tripling, quadrupling and etc.

The computation of $Q = kP$ requires repeated point addition and doubling operations based on the occurrence of zero and nonzero digits in the scalar k . The digit 1 in k indicates that one point addition and one doubling operation are required, whereas the digit 0 indicates that only one point doubling is required.

Table I presents a summary of the point operation cost for the elliptic curve over the binary field in the Lopez Dahab model. Researchers in [10] proposed a general addition formula which costs $14M + 6S + 8A$ where M , S and A denotes the field multiplication, squaring and addition, respectively.

TABLE I
POINT OPERATION COST FOR ELLIPTIC CURVE OVER BINARY FIELD
IN LOPEZ DAHAB MODEL

Point Operation	Lopez Dahab
Point addition	$14M + 6S + 8A$ as in [10] $13M + 4S + 9A$ as in [11]
Mixed addition	$9M + 5S + 9A$ as in [12]
Point doubling	$5M + 4S + 5A$ as in [13]
Point tripling	$12M + 7S$ [14] as in [14]
(M = Field Multiplication; S = Field Squaring)	

Researchers in [11] improved the LD addition formula so that it costs $13M + 4S + 9A$. Previous researchers in [12] proposed a mixed addition formula using affine and LD coordinates which costs $9M + 5S + 9A$. Since [12] addition formula is the cheapest we use their formula in this research.

The formula in [12] is as follows:

Consider $P = (X_1, Y_1, 1)$ and $Q = (X_2, Y_2, Z_2)$ such that $P \neq \pm Q$, where P is an affine coordinate and Q is an LD coordinate. Then, $P + Q = (X_3, Y_3, Z_3)$ is given by:

$$A = Y_2 + Y_1Z_2^2, \quad B = X_2 + X_1Z_2, \quad C = BZ_2, \quad Z_3 = C^2$$

$$D = X_1Z_3, \quad X_3 = A^2 + C(A + B^2 + aC)$$

$$Y_3 = (D + X_3)(AC + Z_3) + [(Y_1 + X_1)Z_3^2]$$

If $a \in \{0, 1\}$, then only eight general multiplications are required. Researchers in [10] also proposed a doubling formula which costs $5M + 5S + 4A$.

The researcher in [13] improved the LD doubling formula so that it costs $5M + 4S + 5A$. We use Lange's doubling formula since it is the cheapest [13]. The formula is as follows. Consider P is in Lopez Dahab coordinates and $P = (X_1:Y_1:Z_1)$. Then, $2P = (X_2:Y_2:Z_2)$ is given by:

$$S = X_1^2, \quad U = S + Y_1, \quad T = X_1Z_1, \quad R = UT$$

$$Z_2 = T^2, \quad X_2 = U^2 + R + aZ_2, \quad Y_2 = [(Z_2 + R)X_2 + S^2Z_2]$$

III. Proposed Methods

The complexity of scalar multiplication, kP , depends on the Hamming weight of the scalar k [15].

In the literature, the scalar recoding technique is used to recode a scalar multiplier k into different number

representations without changing the magnitude of the scalar.

Normally, the scalar k is recoded first before it is used for the scalar multiplication algorithm. In certain conditions, scalar recoding can also reduce the bit-length of the scalar k .

A binary method is a traditional scalar multiplication algorithm using k in binary representation, where $k = (k_{l-1}, k_{l-2}, \dots, k_1, k_0)$ and $k_i \in (0, 1)$. Given an elliptic curve point P , then $kP = \sum_{i=0}^{l-1} k_i 2^i P$.

The Non-adjacent form (NAF) is a number representation in which there is no two adjacent nonzero digits [15]. Non-adjacency is also described as sparseness or as a canonical property.

In the literature, NAF is a well-known number representation with base 2, where each digit in the NAF, $a_i \in \{-1, 0, 1\}$, must satisfy $a_i \cdot a_{i+1} = 0$ for all $i \geq 0$. The NAF is unique with an average Hamming weight of $n/3$ where n is the bit-length of the NAF [16]. In the literature, the NAF is widely used for efficient elliptic curve scalar multiplication.

III.1. Proposed $\{0,1,3\}$ -NAF Representation of Scalar k

In the literature, the scalar k can be represented in base 2 or bases other than 2 or using a combination of different bases. For base 2, the scalar k can be represented in NAF or window-NAF. For bases other than 2, k can be represented in r-NAF [17] or g-NAF [18]. Otherwise, using a combination of different bases, k can be represented in mixed ternary/binary [19], double-base number system with bases $\{2,3\}$ [20], triple base with bases $\{2,3,5\}$ [21] and multibase-NAF with bases $\{2,3\}$ and $\{2,3,5\}$ [22].

From previous research, the cost of scalar multiplication can be improved by using different representations of scalar k . The present research proposes a new NAF with digits $\{1,0,3\}$. To differentiate the existing NAF and the proposed $\{0,1,3\}$ -NAF, we refer to existing NAF as $\{-1,0,1\}$ -NAF. Researchers in [23] proved that digits $\{0,1,3\}$ can have a non-adjacent representation.

Thus, our proposed $\{0,1,3\}$ -NAF representation of k is described in the following example.

Example: Comparison of $\{-1,0,1\}$ -NAF and $\{0,1,3\}$ -NAF

Consider two binary no. $k = 1100001111000011_2$ and $k = 1011001110110011_2$. We propose scanning binary digit k from left-to-right, and replacing any occurrence of consecutive digits 11 with 03, and digits 111 with 103.

Then:

- $1100001111000011_2 = 10\bar{1}0001000\bar{1}00010\bar{1}_{\{-1,0,1\}\text{-NAF}}$
(Hamming weight: 6; Bit-length: 17)
- $1100001111000011_2 = 3000003030000003_{\{0,1,3\}\text{-NAF}}$
(Hamming weight: 4; Bit-length: 15)
- $1011001110110011_2 = 10\bar{1}0\bar{1}01000\bar{1}0\bar{1}0\bar{1}0\bar{1}_{\{-1,0,1\}\text{-NAF}}$

(Hamming weight: 8; Bit-length: 17)

d) $1011001110110011_2 = 1003001030030003_{\{0,1,3\}\text{-NAF}}$

(Hamming weight: 6; Bit-length: 16)

From the example, the Hamming weight and the bit length of the proposed method are better than the existing NAF.

III.2. Proposed Conversion Algorithm for Scalar k

A scalar recoding can be done either by scanning digits of k from left-to-right (homogeneous) or right-to-left (heterogeneous) [24]. The homogeneous approach is a real-time conversion because the converted k is used straight away for the scalar multiplication algorithm.

This is possible because scanning digits of k for conversion and scalar multiplication is done using the same mode. In the literature, this type of recoding promotes better memory usage and is largely preferred for memory constraint devices [25]. In the heterogeneous approach, the converted k is saved before it is used in the scalar multiplication algorithm.

This is because the scanning of digits of k for conversion and scalar multiplication is initiated from different directions, that is, the conversion mode is right-to-left and the scalar multiplication mode is left-to-right. Generally, this type of recoding needs an additional n -bit RAM for storage, where n is the bit size of the scalar [26]. The proposed algorithm is used for converting scalar k from binary to $\{0, 1, 3\}$ -NAF. Then, it is expected that the scalar multiplication computation has a reduction in the average number of point operations.

Algorithm 1: Proposed $\{0,1,3\}$ -NAF Recoding Algorithm

Input: $k = (k_{m-1}, \dots, k_0)_2$

Output: $k' = (k'_m, k'_{m-1}, \dots, k'_0)_{\{0,1,3\}\text{-NAF}}$

- $b_m \leftarrow 0; k_m \leftarrow 0; k_1 \leftarrow 0; k_2 \leftarrow 0; k'_m \leftarrow 0$
- for i from $m-1$ down to 0 do
- {
- $b_i \leftarrow \lfloor (b_{i+1} + k_i + k_{i-1})/2 \rfloor$
- switch($b_{i+1}, k_{i+1}, k_i, k_{i-1}, b_i$)
- {
- case (0,0,0,0,0): $k'_i \leftarrow 0$
- case (0,0,0,1,0): $k'_i \leftarrow 0$
- case (0,0,1,0,0): $k'_i \leftarrow 1$
- case (0,0,1,1,1):
- flag $\leftarrow 0$; count $\leftarrow 0$; $j \leftarrow m-1$
- while (flag == 0 && $j > 0$)
- {
- if ($k_j == 1$ && $k_{j-1} == 1$)
- then count \leftarrow count + 1
- else if ($k_j == 1$ && $k_{j-1} == 0$ && count >=
- 1)
- then count \leftarrow count + 1; flag $\leftarrow 1$
- $i \leftarrow i-1$
- } //end while
- if (count%2 == 0)
- $k'_i \leftarrow 0$
- else

```

23.       $k'_i \leftarrow 1$ 
24.      case (0,1,0,0,0) :  $k'_i \leftarrow 0$ 
25.      case (0,1,0,1,0) :  $k'_i \leftarrow 0$ 
26.      case (0,1,1,1,1) :  $k'_i \leftarrow 0$ 
27.      case (1,0,1,0,1) :  $k'_i \leftarrow 1$ 
28.      case (1,0,1,1,1) :  $k'_i \leftarrow 1$ 
29.      case (1,1,0,0,0) :  $k'_i \leftarrow 0$ 
30.      case (1,1,0,1,1) :
31.          if ( $k'_{i+1} = 1 \parallel k'_{i+1} = 3$ )  $k'_i \leftarrow 0$ 
32.          if ( $k'_{i+1} = 0$ )  $k'_i \leftarrow 3$ 
33.      case (1,1,1,0,1) :
34.          if ( $k'_{i+1} = 0$ )  $k'_i \leftarrow 3$ 
35.          if ( $k'_{i+1} = 1$ )  $k'_i \leftarrow 0$ 
36.      case (1,1,1,1,1) :
37.          if ( $k'_{i+1} = 1 \parallel k'_{i+1} = 3$ )  $k'_i \leftarrow 0$ 
38.          if ( $k'_{i+1} = 0$ )  $k'_i \leftarrow 3$ 
39.      } // end switch
40.  } // end for
41. return ( $k'_m, k'_{m-1}, \dots, k'_0$ ).

```

III.3. Proposed {0,1,3}-NAF Scalar Multiplication

The {-1,0,1}-NAF of k uses the addition-subtraction algorithm to compute scalar multiplication. The point operations involved in this algorithm are the point addition, point subtraction and point doubling. Point addition and point subtraction are treated as similar operations with the same complexity. Throughout the execution of this algorithm, the point operation performed is in the form of $2P$, $2P + Q$ or $2P - Q$. The average cost of Algorithm 2 is $\frac{l}{3}A + (l-1)D$ where A is the no. of point addition and D is the no. of point doubling. Algorithm 2 is shown below.

Algorithm 2: Addition-Subtraction Algorithm

Input: $NAF(k) = \sum_{i=0}^{l-1} k_i 2^i$ and $P \in E(F_2^n)$

Output: $Q = kP$, where $Q \in E(F_2^n)$

```

1.  $Q \leftarrow \infty$ 
2. for  $i = l-2$  down to 0
3.    $Q = 2Q$ 
4.   if ( $k_i = 1$ ) then  $Q = Q + P$ 
5.   if ( $k_i = -1$ ) then  $Q = Q - P$ 
6. return  $Q$ .

```

Our proposed {0,1,3}-NAF of k requires a new algorithm to compute scalar multiplication. Therefore, we also propose a {0,1,3}-NAF scalar multiplication algorithm as follows:

Algorithm 3: {0,1,3}-NAF Scalar Multiplication Algorithm

Input: k is in {0,1,3}-NAF such that $k = \sum_{i=0}^{m-1} b_i 2^i$ where $b_i \in \{0,1,3\}$, $P \in E(F_2^n)$

Output: $Q = kP$

```

1.  $P := P(x_1, y_1)$ 
2.  $3P := \text{Tripling}(P)$ 
3. if ( $b_{m-1} = 1$ ) then
4.    $Q := P$ 
5. if ( $b_{m-1} = 3$ ) then
6.    $Q := 3P$ 
7. for  $i$  from  $m-2$  down to 0 do
8.    $Q := \text{double}(Q)$ 
9.   if  $b_i = 1$  then
10.     $Q := \text{add}(P, Q)$ 
11.   if  $b_i = 3$  then
12.     $Q := \text{add}(3P, Q)$ 
13. Return ( $Q = kP$ ).

```

In Algorithm 3, line 7 is performed exactly $m-2$ times. The expected running time of Algorithm 3 is as follows:

$$\text{Cost (Algorithm 3)} = (m/3) A + (m-1) D + 1T$$

where m is the bit-length of the signed-digit {0,1,3}-NAF scalar, A , D and T are point addition, doubling and tripling operations, respectively. The point tripling is computed once only and stored in memory. Digit 3 in the scalar k indicates the need for the point tripling operation.

Table A1 (in Appendix) compares the estimation costs of Algorithm 2 and Algorithm 3 in an average case. The estimation costs are given in terms of the number of point operations (i.e., point addition (A), point doubling (D) and point tripling (T)), and the corresponding field operations (i.e., multiplication (M) and squaring (S)). The estimation is sufficiently accurate to permit meaningful comparisons.

IV. Results

Based on the proposed methods, the result is shown in Table A2 and Table A3 (in Appendix).

Table A2 shows that at average case, k in {0,1,3}-NAF achieved the lowest Hamming weight than the k in {-1,0,1}-NAF. Thus, k in {0,1,3}-NAF is better than the k in {-1,0,1}-NAF at average case. Table A3 provides overall cost estimation for addition-subtraction and the proposed {0,1,3}-NAF scalar multiplication algorithms. Significant parameters in the performance of the scalar multiplication algorithm are the values of p , Hamming weight and bit length of k .

For efficient scalar multiplication, Hamming weight and bit length of k must have small values. In the Table A3, bit length of k before conversion is $l = 24$, and this value is fix in order to monitor the performance of the algorithm. At average case, the Hamming weight of k before conversion is $l/2$, which is equal to 12, and this value also fix for the same reason. Table A2 shows that the value of p is increasing from $p = 1$ up to $p = 6$ as we go down the table. The variation of h_1 , h_2 , l_1 and l_2 are observed. Based on Table A3, the values of h_1 are lower than h_2 indicate that the Hamming weight of k in {0,1,3}-

NAF scalar is better than the Hamming weight of k in $\{-1,0,1\}$ -NAF. Also, by observation, the values of h_1 are decreasing as we go down the table and at all cases the values of h_1 are $h_1 < h_2$. Also, the values of l_1 are lower than l_2 indicate that the bit length of k in $\{0,1,3\}$ -NAF is better than the bit length of k in $\{-1,0,1\}$ -NAF. By observation, at most cases $l_1 = l_2$. But, there are some cases where $l_1 < l_2$ i.e. $l_1 = 23$. This case happens in data d), i), j) and k). Further investigation is carried out for data d), i), j), and k).

Thus, the value $l_1 = 23$ (i.e. reduction of bit size) occurs because the leftmost digit of k in $\{0,1,3\}$ -NAF is 3.

It is also observed that the proposed $\{0,1,3\}$ -NAF scalar multiplication algorithms has less number of iteration than the addition-subtraction algorithm. Thus, the percentage of cost reduction is better than in normal cases. Furthermore, the percentage of cost reduction helps to identify significant improvement of the scalar multiplication algorithms.

The highest percentage of cost reduction is 16.2, which happens when $p = 6$. Negative percentage of cost reduction of -1.4, -6, and -6.6 indicate no improvement in the cost of the proposed algorithm. Further investigation is carried out for data a).

The value $p = 1$ may be the reason for the negative value in percentage of cost reduction in data a). Also, data (c) and data g) are analyzed.

Thus, the value $h_1 = h_2$ and $l_1 = l_2$ may be the reason for the negative values in the percentage of cost reduction in data c) and g).

V. Conclusion

In this study, we provide an efficient representation of scalar k together with its conversion algorithm from binary to $\{0,1,3\}$ -NAF.

We also proposed $\{0,1,3\}$ -NAF scalar multiplication algorithm because the proposed $\{0,1,3\}$ -NAF of k cannot operates with the existing scalar multiplication algorithm. In analysis, we prove that the Hamming weight of k in $\{0,1,3\}$ -NAF is better than the Hamming weight of k in $\{-1,0,1\}$ -NAF.

Finally, from analysis we prove that the cost of the new $\{0,1,3\}$ -NAF scalar multiplication algorithm is better than addition-subtraction algorithm at average case. These efficient algorithms will allow low cost of scalar multiplication and will give result to a more efficient elliptic curve cryptosystem.

Acknowledgements

This work was supported by RUGS Grant of Universiti Putra Malaysia, Serdang, Selangor, Malaysia.

Appendix

TABLE A1
AVERAGE CASE ANALYSIS OF ALGORITHM 2 AND ALGORITHM 3 FOR $K=24$ BIT LENGTH

a) $\{-1,0,1\}$ -NAF Scalar Multiplication (Addition-Subtraction) (Algorithm 2)		b) $\{0,1,3\}$ -NAF Scalar Multiplication (Algorithm 3)	
i	$NAF(k) = \sum_{i=0}^{l-1} k_i 2^i ; k_i \in \{-1,0,1\}$ $k = 1010101010101010-10101001$	i	$\{0,1,3\}$ -NAF, $k = \sum_{i=0}^{l-1} b_i 2^i ; b_i \in \{0,1,3\}$ $k = 10101010101010100030101001$
23	P	23	P
22	2P	22	2P
21	2(2P)+P=5P	21	2(2P)+P=5P
20	2(5P)=10P	20	2(5P)=10P
19	2(10P)+P=21P	19	2(10P)+P=21P
18	2(21P)=42P	18	2(21P)=42P
17	2(42P)+P=85P	17	2(42P)+P=85P
16	2(85P)=170P	16	2(85P)=170P
15	2(170P)+P=341P	15	2(170P)+P=341P
14	2(341P)=682P	14	2(341P)=682P
13	2(682P)+P=1365P	13	2(682P)+P=1365P
12	2(1365P)=2730P	12	2(1365P)=2730P
11	2(2730P)+P=5461P	11	2(2730P)+P=5461P
10	2(5461P)=10922P	10	2(5461P)=10922P
9	2(10922P)+P=21845P	9	2(10922P)=21844P
8	2(21845P)=43690P	8	2(21844P)=43688P
7	2(43690P)+P=87379P	7	2(43688P)+3P=87379P
6	2(87379P)=174758P	6	2(87379P)=174758P
5	2(174758P)+P=349517P	5	2(174758P)+P=349517P
4	2(349517P)=699034P	4	2(349517P)=699034P
3	2(699034P)+P=1398069P	3	2(699034P)+P=1398069P
2	2(1398069P)=2796138P	2	2(1398069P)=2796138P
1	2(2796138P)=5592276P	1	2(2796138P)=5592276P
0	2(5592276P)+P=11184553P	0	2(5592276P)+P=11184553P
Cost = 23D + 11A		Cost = 23D + 10A + 1T	

TABLE A2
ANALYSIS OF HAMMING WEIGHT FOR K OF 16 BIT LENGTHS

No	K in Binary	K in {-1,0,1}-NAF	K in {0,1,3}-NAF	H _{binary}	H _{NAF}	H _{{0,1,3}-NAF}
1.	1100001111000011	10100010001000101	300000303000003	8	6	4
2.	1101100011011000	10010100100101000	300300003003000	8	6	4
3.	1001101110011010	1010010010101010	1000301030003010	9	7	6
4.	1010110110101101	10101001001010101	1010030030100301	10	8	7
5.	1011001110110011	10101010001010101	1003001030030003	10	8	6
6.	110100111010011	10101010001010101	301000303010003	10	8	6
7.	1110011011100110	10010100100101010	1030003010300030	10	7	6

TABLE A3
ESTIMATION COST OF THE ADDITION-SUBTRACTION AND THE PROPOSED {0,1,3}-NAF SCALAR MULTIPLICATION ALGORITHMS

	p	h ₁	h ₂	l ₁	l ₂	Estimation Cost (microseconds)		% Cost Reduction
						Addition-Subtraction	Proposed Signed-digit {0,1,3}-NAF	
a)	1	11	12	24	24	49211.3	49894.0	-1.4
b)	2	10	12	24	24	49211.3	47964.8	2.5
c)	3	9	9	24	24	43423.8	46035.6	-6
d)	3	9	10	23	25	46570.0	44818.7	3.8
e)	3	9	12	24	25	50428.3	46035.6	8.7
f)	4	8	12	24	25	50428.3	44106.5	12.5
g)	5	7	7	24	24	39565.5	42177.3	-6.6
h)	5	7	10	24	24	45353.0	42177.3	7
i)	6	6	10	23	25	46570.0	39031.2	16.2
j)	6	6	8	23	25	42711.6	39031.2	8.6
k)	6	6	9	23	25	44640.8	39031.2	12.6

(p = no. of digit 3 in {0,1,3}-NAF of k; h₁ = Hamming weight of k in {0,1,3}-NAF;

h₂ = Hamming weight of k in {-1,0,1}-NAF;

l₁ = bit size of k in {0,1,3}-NAF, l₂ = bit size of k in {-1,0,1}-NAF.)

References

- [1] Miller, V. 1986. *Use of Elliptic Curves in Cryptography*. Advances in Cryptology. Proceedings of CRYPTO'85. LNCS (page: 218 Year of Publication: 1986).
- [2] Kumar, R. and Anil, A. Implementation of elliptical curve cryptography. *International Journal of Computer Sciences Issues (IJCSI)*, Vol. 8, pp. 544-549, 2011.
- [3] Muthukuru, J. and Sathyanarayana, B. Fixed and variable size text based message mapping techniques using ECC, *Global Journal of Computer Science and Technology*, Vol. 12, Issues 3, 2012.
- [4] S. Rao, P. Setty, Efficient Mapping methods for Elliptic Curve Cryptosystems. *International Journal of Engineering Science and Technology*, Vol. 2, Issues 8, pp. 3651-3656, 2010.
- [5] W. Stallings, *Cryptography and Network Security: Principles and Practices*. 3rd Edition (Pearson Education Inc., New Jersey, 2003).
- [6] Yong-Ping, D., Xue-cheng, Z., Zheng-lin, L., Yu, H., and Li-hua, Y. High-Performance Hardware Architecture of Elliptic Curve Cryptography Processor Over GF(2¹⁶³). *Journal of Zhejiang University Science*, Vol. 10, Issue 2, pp. 301-310, 2009.
- [7] Al Khatib, M. and Al-Salem, A. *Efficient Hardware Implementations for Tripling Oriented Elliptic Curve Cryptosystem*. *International review on Computers and Software*. Vol. 9, Issues4, pp.609-617, 2014.
- [8] M. Mosoumi, H. Mahdizadeh, A Novel and Efficient Hardware Implementation of Scalar Point Multiplier. *Iranian Journal of Electrical & Electronic Engineering*, Vol. 8, n. 4, pp. 290-302, 2012.
- [9] N. Gura, A. Patel, A. Wander, H. Eberle, S. Shantz, Comparing Elliptic Curve Cryptography and RSA on 8-bit CPUs. *Proceedings of Workshop on Cryptographic Hardware and Embedded Systems (CHES'04)*, Springer, pp. 119-132, 2004.
- [10] J. Lopez, R. Dahab, Improved Algorithms for Elliptic Curve Arithmetic in GF(2ⁿ). *Tech. Report IC-98-39*, October 1998.
- [11] A. Higuchi, and N. Takagi, A fast addition algorithm for elliptic curve arithmetic in GF(2ⁿ) using projective coordinates. *Information Processing Letter*. Vol. 76. Issues 3, pp. 101-103, 2000.
- [12] E. Al-Daoud, R., Mahmod, M. Rushdan, A. Kilicman, A New Addition Formula for Elliptic Curves Over GF(2ⁿ). *IEEE Transactions on Computers*. Vol. 51, pp.972-975, 2002.
- [13] T. Lange, A Note on Lopez-Dahab Coordinates. <http://eprint.iacr.org/2004/323.pdf>, 2004.
- [14] M. Y. Sharifah, *New Signed-Digit {0,1,3}-NAF Scalar Multiplication Algorithm for Elliptic Curve Binary Field*. Phd Thesis. Faculty Computer Science and Information Technology. University Putra Malaysia, 2011.
- [15] B. Wang, H. Zhang, Z. Wang, Y. Wang, Speeding Up Scalar Multiplication Using a New Signed binary Representation for Integers. *LNCS*, Vol. 4577, pp.277-285, 2007.
- [16] G.W. Reitwiesner, Binary Arithmetic. *Advances in Computers*. Academic Press, Vol. 1, pp. 231-308, 1960.
- [17] T. Takagi, S. Yen, B. Wu, *Radix-r Non-adjacent Form*. International Conference on Information Security (ISC'04). LNCS, Vol. 3225, pp.99-110, 2004.
- [18] M. Joye, S. Yen, New Minimal Modified Radix-r Representation with Applications to Smart Cards. *Public Key Cryptography (PKC 2002)*. LNCS, Vol. 2274, pp. 375-384, 2002.
- [19] M. Ciet, M. Joye, K. Lauter, P. Montgomery, Trading inversions for multiplications in elliptic curve cryptography. *Designs, Codes and Cryptography*. Vol. 39, Issues 2, pp.189-206, 2006.
- [20] V. Dimitrov, L. Imbert, P. Mishra, The Double-base Number System and Its Application to Elliptic Curve Cryptography. *Mathematics of Computation Journal*, Vol. 77, Issues 1075-1104, 2008.
- [21] P. Mishra, V. Dimitrov, Efficient Quintuple Formulas for Elliptic Curves and Efficient Scalar Multiplication Using Multibase Number Representation. *LNCS*, Vol. 4779, Issues 390-406, 2007.
- [22] P. Longa, Accelerating the Scalar Multiplication on Elliptic Curve Cryptosystems over Prime Fields. Master Thesis, University of Ottawa, Canada. 2007.
- [23] J. Muir, D. Stinson, Alternative Digit Sets For

- Nonadjacent Representations. *SIAM Journal. Discrete Math.*, Vol. 19, Issues 1, pp. 165–191, 2005.
- [24] M. Joye, S. Yen, Optimal left-to-right signed-digit recoding. *IEEE Transactions on Computers*, Vol. 49, Issues 7, pp.740-748. 2000.
- [25] M. Khabbazi, T. Gulliver, V. Bhargava, A New Minimal Average Weight Representation for Left-to-right Point Multiplication Methods. *IEEE Transactions on Computers*. Vol. 54, Issues 11, pp. 1454-1459, 2005.
- [26] K. Okeya, K. Schmidt-Samoa, C. Spahn, T. Takagi, *Signed Binary Representations Revisited*. Proceedings of CRYPTO'04. pp.123- 139, 2004.
- [27] Alkhatib, M., Al Salem, A., Efficient hardware implementations for tripling oriented elliptic curve crypto-system, (2014) *International Review on Computers and Software (IRECOS)*, 9 (4), pp. 609-617.
- [28] Tripathy, P.K., Biswal, D., Multiple server indirect security authentication protocol for mobile networks using elliptic curve cryptography (ECC), (2013) *International Review on Computers and Software (IRECOS)*, 8 (7), pp. 1571-1577.
- [29] Muthu Kumar, B., Jeevananthan, S., HELP multiplier based montgomery key generation for Elliptic Curve Cryptography over GF (2^m), (2012) *International Review on Computers and Software (IRECOS)*, 7 (3), pp. 943-949.
- [30] Al-Khatib, M., Jaafar, A., Zukarnain, Z., Rushdan, M., Hardware designs and architectures for projective Montgomery ECC over GF (p) Benefiting from mapping elliptic curve computations to different degrees of parallelism, (2011) *International Review on Computers and Software (IRECOS)*, 6 (6), pp. 1059-1070.

Authors' information



Dr. **Sharifah Md. Yasin** received the bachelor degree BSc. (Hons) in Mathematics and Statistics from University of Bradford, England in 1991. Her master degree is MSc. in Information Technology from the University Kebangsaan Malaysia, in 2002. She graduated her Ph.D. degree in the computer security field from University Putra Malaysia, 2011. From

1999 to 2003, she was a tutor at University Putra Malaysia. From 2004 to 2011, she has been a lecturer at University Putra Malaysia. Currently, she is a senior lecturer at University Putra Malaysia. Her PhD research work related to elliptic curve cryptography (ECC). She developed new formula and algorithms in ECC. Her research interest is in cryptography and computer security. She is a member in Malaysian Security for Cryptology Research (MSCR) since 2006. She is also a member in Information Security Research Group for Faculty of Computer Science and Information Technology, University Putra Malaysia since 2011.



Dr. **Rozi Nor Haizan Nor** has acquired her doctorate from University of Technology Malaysia (UTM) in Computer Science. He is currently working as a senior lecturer in the department of Software Engineering and Information System in Faculty of Computer Science and Information Technology (FSKTM), Universiti Putra Malaysia (UPM) Selangor,

Malaysia. Her expertise is in the area of Information Systems, ICT services, ICT service quality and ICT Governance.



Dr. **Jamilah Din** is a faculty member in the Software Engineering and Information System Department of the Faculty of Computer Science and Information Technology at Universiti Putra Malaysia. She received the B.S. in Computing Science from University of Evansville, Indiana in 1987, MSc in Computer Science from Putra University of Malaysia in 2002, and PhD in

Computer Science from National University of Malaysia in 2009. Her teaching interests are in the area of Object-Oriented Analysis and Design, and Software Engineering.

Hybrid Re-Clustering Algorithm for Enhancement of Network Lifetime in Wireless Sensor Networks

Aby K. Thomas¹, R. Devanathan²

Abstract – In this paper, we propose a re-clustering framework for wireless sensor networks to undergo global re-clustering and local delegation in order to enhance the lifetime of the network. The variation in energy distribution across the CH network can be characterized by a mapping function using a metric, based on energy cost, defined on a metric space. The mapping characterizes the change in link cost in terms of energy distribution. Supremum of the change in link costs is defined as the distortion of the map. Distortion exceeding a given threshold is used to decide on re-clustering / local delegation. The global monitoring of energy distribution in the CH network afforded by the metric space and the distortion concept helps to provide a energy management framework for WSN which is proved to be effective. Simulation carried out in MATLAB on a typical WSN showed favourable comparison in terms of average network life time with respect to established results. **Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords: Distortion, Global Re-clustering, Local Delegation, Metric Space, Network Lifetime

Nomenclature

CH	Cluster Head
CN	Cluster Node
C_i	Cost of i^{th} Node
R_i	Residual Energy of i^{th} Node
d_{ij}	Metric, distance between i^{th} and j^{th} nodes in CH
X,Y	Metric Space
x_1, x_2	Nodes
$dis(f)$	Distortion
T_1	Threshold
L	Packet Size
d	Distance between Transmitter and Receiver Unit
E_{elec}	Energy Dissipated by Transceiver Circuit
ϵ_{fs}	Amplification Constant for free Space
ϵ_{mp}	Energy Dissipated by Transmit Amplifier

I. Introduction

Wireless sensors networks (WSNs) operate unattended in harsh environments. Sensors are expected to be deployed randomly in the area of interest by relatively uncontrolled means, and to collectively form a network in an ad-hoc manner. Designing and operating such large size network requires scalable architectural and management strategies. In addition, sensors in such environments are energy constrained and their batteries cannot be recharged. Therefore, designing energy-aware algorithms becomes an important factor for extending the lifetime of sensors ([1]-[23]).

Depending on the activity level of a battery driven node, its lifetime may only be a few days if no power management schemes are used. Since most systems require much longer lifetime, significant research has been undertaken to increase lifetime while still meeting functional requirements.

At the hardware level it is possible to add solar cells or scavenge energy from motion or wind. Improvements have taken place in the performance of batteries, low power circuit designs and microcontroller designs. Most hardware platforms allow multiple power saving states (off, idle, on) for each component of the device (each sensor, the radio, the microcontroller). In this way, only the components required at a particular time need to be active. At the software level, power management solutions are targeted such as (i) in minimizing communications since transmitting and listening for messages is energy expensive, and (ii) in creating sleep/wake-up schedules for nodes or particular components of nodes [1]. Minimizing the number of messages seems to be a feasible solution. Efficient neighbor discovery, time synchronization, localization, query dissemination and flooding can all reduce the number of messages thereby increasing lifetime. Solutions to schedule sleep/wake-up patterns vary considerably. Cluster head role rotation and duty-cycling of nodes are two effective ways to balance energy consumption. Energy consumption being the most important factor to determine the life of a sensor network, energy awareness needs to be incorporated into every aspect of design and operation of WSN.

In sensor networks, a typical application is the gathering of sensed data to a distant base station (BS). A sensor can communicate directly only with other sensors

within its range. To go beyond, sensors need to form multi-hop links. It is well acknowledged that clustering is an efficient way to save energy for sensor networks.

In multi-hop networks clustering is very effective in reducing communications, i.e., the data gathered by the sensors is combined at the cluster-heads before sending to the BS. Clustering is particularly crucial for scaling the network to hundreds or thousands of nodes. Many clustering algorithms have been proposed [2]-[9] for wireless ad hoc networks. Most of these algorithms are specifically designed for generating stable clusters in mobile networks. But in sensor networks, the locations of nodes are mostly fixed and instability is not important. In sensor networks, clustering is mainly for communication efficiency.

Low Energy Adaptive Clustering Hierarchy (LEACH)[10] is a cluster-based protocol that includes distributed cluster formation. The authors proposed a randomized rotation of the role of cluster head with the objective of reducing energy consumption and to distribute the energy load evenly among the sensors in the network in order to enhance the network lifetime. LEACH uses localized coordination to enable scalability and robustness for dynamic networks and incorporates data fusion into the routing protocol in order to reduce the amount of information that must be transmitted to the base station. However, selective rotation of cluster head (CH) proved more advantageous. This protocol is divided into rounds, each round consists of two phases namely Set-up Phase and Steady Phase. In the Set-up Phase, each node decides, independent of other nodes, whether it will become a CH or not. This decision takes into account whether the node has already served as a CH. The node that has not been a CH for a long time is more likely to elect itself than nodes that have been a CH recently.

Although LEACH protocol acts in a good manner, it suffers from many drawbacks such as i) CH selection is random, that does not take into account energy consumption. ii) It cannot cover a large area. iii) CHs are not uniformly distributed and may be located at the edges of the cluster. Many researches have taken place to make the LEACH protocol perform better. Multi-hop-LEACH protocol [11] selects optimal path between the CH and the BS through other CHs and uses these CHs as relay stations to transmit data through them. Multi-hop -LEACH protocol is almost the same as LEACH protocol, only in that it assumes communication mode from single hop to multi-hop between CHs and BS.

Hybrid Energy Efficient Distributed Clustering Protocol [HEED][12] picks the cluster head from the group of nodes on the basis of their residual energy and other parameters like the node degree or proximity of the nodes to the sink node. HEED is one of the effective data gathering protocols without location support. Repeated clustering introduces communication and processing overhead which taxes the sensor energy. E-LEACH protocol [13] improves on the CH selection procedure. It makes residual energy of node as the main metric which decides whether the nodes turn into CH or not after the

first round similar to LEACH protocol. In E-LEACH every node has the same probability to turn into CH, in the first round, whereas in the next rounds, the residual energy of each node is taken into account for the selection of the CHs.

An enhancement over the LEACH protocol called LEACH-C [14], uses a centralized clustering algorithm which can produce a better performance by dispersing the cluster heads throughout the network. A new version of LEACH called Two-level LEACH was proposed in [15] where CH collects data from other cluster members as the original LEACH, but rather than transferring data to the BS directly, it uses one of the CHs that lies between the CH and the BS as a relay station. A dynamic multi-level hierarchical clustering approach [16] was also proposed where, a group of nodes together forms a cluster and chooses a cluster head from among the member nodes depending on the maximum residual energy, the distance between the nodes and the node degree. The cluster heads of different clusters again form a cluster and choose, its head based upon a similar criteria and so on. So the whole system can be multi-hierarchical in nature. In this paper, in view of the importance of monitoring of the energy levels of CHs, we provide a global approach to the same. We consider the network of CH to form a metric space. A metric is defined between any two CHs in terms of energy spent. After data transmission for a certain period, energy gets depleted in the nodes. This changes the distribution of energy especially among CH. The network of CH after data transmission has taken place can be considered a distorted metric space with the new value of the metric between any two CHs. The maximum distortion between any two nodes can be a global parameter to monitor the rate of a depletion of energy.

The global monitoring combined with a CH rotation/re-clustering (if it is necessitated) is the key strategy proposed to preserve energy and extend the average lifetime of nodes in WSN. The main contribution of the paper is the introduction of the original concept of metric space for the networks of CH to monitor their energy distribution and the use of distortion of function to monitor the network energy depletion rate. The concept is exploited to provide a framework for WSN energy management where CH rotation/re-clustering is taken in the homogenous network. Simulation studies reveal favorable comparison of the results of the proposed approach with that of LEACH. The rest of the paper is organized as follows. Section II discusses the framework of energy management of WSN and the distortion concept [17]. Methodology used and the proposed Hybrid Re-clustering Algorithm are described in Section III. Simulation results are discussed in Section IV and Section V concludes the work with future research directions.

II. Metric Space in WSN

Consider a WSN consisting of clusters each with

cluster Head (CH).

Since CH receives data from all cluster nodes (CN) and relays the data to the base station (BS), the CH is expending energy much faster than CN in homogeneous network. Considering the network of CHs in WSN, one can formulate it as a metric space. We define a metric in terms of energy cost. The set of nodes together with the metric defined can be considered a metric space [17].

When data transmission takes place, the energy cost attribute associated with a node changes, changing the metric between the given two nodes. More correctly, the set of nodes with the associated energy cost of each node corresponds to the set of underlying elements of a set with the metric defined in terms of energy cost.

Consider two CHs, CH_i and CH_j . The metric d_{ij} can be defined as follows:

$$d_{ij} = \max\{C_i, C_j\} \quad (1)$$

where C_i corresponds to the cost associated with the node CH_i . The cost is given by:

$$C_i = 1 - \mathcal{R}_i \quad (2)$$

where \mathcal{R}_i is the residual energy of CH_i . It can be verified easily that d_{ij} is a metric and satisfies the following axioms:

$$\begin{aligned} d_{ij} &\geq 0, i \neq j; & d_{ii} &= 0 \quad (\text{assumed}) \\ d_{ij} &= d_{ji}, i \neq j \\ d_{ij} &\leq d_{ik} + d_{kj} \quad i \neq j \quad d_{ii} = 0 \end{aligned} \quad (3)$$

$$d_{ij} \leq d_{ik} + d_{kj} \quad i \neq j \text{ or } k \quad (4)$$

Assuming an initial energy (normalized) to unity, C_i represents the energy spent by CH_i . All the variables C_i , \mathcal{R}_i and d_{ij} are functions of time since transmission is assumed to take place continuously depleting energy of the nodes. We assume an energy depletion rate for all CHs which does not include higher depletion rate due to data relaying.

Let us call the metric space formed by the CHs initially as X . With data transmission taking place, the cost of nodes and links change causing a change in the value of metric between any two given points. This change may be considered to correspond to a mapping.

$$f: X \longrightarrow Y \text{ such that } x \in X \mapsto f(x) \in Y.$$

In other words, $f(\cdot)$ corresponds to change in energy levels of nodes in the metric space due to data transmission. Considering two nodes x_1 and x_2 in X ,

which are mapped to $f(x_1)$ and $f(x_2)$ points in Y which denotes state of the network after data transmission has taken place. Since the distance metric defined above denoted as ' d ' is a functional taking (i, j) to \mathcal{R} , where \mathcal{R} is the space of real numbers $(0, \infty)$, the distortion caused by $f(\cdot)$ can be defined [17] as:

$$dis(f) = \sup_{x_1, x_2 \in X} \left\{ \left| d(x_1, x_2) - d(f(x_1), f(x_2)) \right| \right\} \quad (5)$$

A decision will be taken to carry out local delegation /re clustering of CH in all clusters when the below condition is met:

$$dis(f) > T_1 > 0 \quad (6)$$

where T_1 is an appropriately chosen threshold.

III. Hybrid Re-Clustering Methodology

The global monitoring combined with a CH rotation/re-clustering is the key strategy proposed to preserve energy and extend the average lifetime of nodes in WSN.

III.1. Clustering & CH Selection

We assume that clustering has been carried out at the beginning by the base station. The clustering process is done using the Hausdorff distance concept as in [18]. We assume homogeneous nodes so that each node can be a CH in turn inherently, Hausdorff distance implies more than a single hop communication from cluster node (CN) to CH. This also means that some CN tend to relay other CN's data to CH with the result that CNs in a cluster lose energy at differential rates.

As per the Hausdorff criteria [18], the lowest transmission power level, with a range R_1 , is used to cover the intra-cluster transmission. The higher power levels are for reaching neighboring cluster heads with a range of R_2 .

In Hausdorff criteria based cluster, every node is at least within a distance R_1 from some node in the cluster. Hence, CH rotation can be done without the need for re-clustering. Initially, only local delegation based on maximum available energy is required. Re-clustering may be needed only when no more eligible CN, with sufficient residual energy is available in the cluster.

Thus, two kinds of phases are possible: CH re-clustering and CH rotation. In CH rotation CH selection will take place within the cluster based on the highest residual energy node in the cluster nominated to be the next CH after certain number of data gathering rounds depending on the threshold set. A setup message is sent to all cluster nodes about the new CH. Also a message of TDMA schedule is sent to all cluster nodes by the CH.

In re-clustering, CH candidate broadcasts a setup message, Potential members receive the message and the members transmit a join message.

Initial clustering followed by data gathering round is carried out for a specified number of times. This is followed by re-clustering/CH rotation with subsequent data gathering a specified number of times until a lifetime limit is reached as specified.

The lifetime limit can be specified as the time until the first cluster node or CH loses all its energy or 50 % of nodes in the networks loose their energy to 20% of their respective maximum energies.

III.2. The Energy Cost Models

We have assumed a simple first order radio model for the analysis. Here the transmitter and the receiver dissipate energy to run the radio electronics but transmitter additionally expends some energy due to its signal amplifier. Moreover, the actual power dissipated depends on the distance between transmitter and receiver.

Path loss can be modeled as proportional to inverse of square of the distance if the distance is small whereas it is taken inversely proportional to the fourth power of the distance if the distance is large. In this way, to transmit an L-bit message across a distance d , the energy expenditure $E_{tr}(L, d)$ can be modelled as

$$E_{tr}(L, d) = E_{tr,elec}(L) + E_{tr,amp}(L, d) \quad (7)$$

where:

$$\begin{aligned} E_{tr,elec}(L) &= LE_{elec} \text{ and} \\ E_{tr,amp}(L, d) &= L\epsilon_{fs}d^2; \text{ for free space} \\ &= L\epsilon_{mp}d^4 = \text{ for multipath} \end{aligned} \quad (8)$$

The values of E_{elec} , ϵ_{fs} , and ϵ_{mp} in the simulation are specified in Table I.

TABLE I
SPECIFICATIONS ADOPTED FOR THE SIMULATED NETWORK

Type	Simulation Parameter	Qty
Network	No. of Nodes	100
	Area	100×100
	Cluster head rotation periodicity	5
	Initial Energy of a Node	1 Joule
Application	Data Packet size	500 bytes
	Broadcast packet size	25 bytes
	Packet header size	25 bytes
	Average Duty Cycle of the network	0.4
	Compression Ratio	0.1
	Distortion Threshold	0.4
Radio model	E_{elec}	50nJ/bit
	ϵ_{fs}	10pJ/bit/m ²
	ϵ_{mp}	0.0013pJ/bit/m ⁴

III.3. Hybrid Re-clustering Algorithm (HRA)

We present below the steps followed in the proposed algorithm:

1. Do initial deployment of nodes randomly.
2. Divide the network geographically into N clusters

with a CH nominated for each cluster.

3. Assign initial energy of unity to all nodes in the network
4. Initiate intra-cluster data transmission followed by data aggregation and data transmission from CHs to BS. For intra-cluster data transmission, routing of data packets through higher energy routes using R1 range is used. All CNs in a cluster are assumed to go into sleep mode as per the TDMA schedule adopted. Average Duty cycle of 40% is considered for simulation. Also, for inter-cluster communication, an energy aware routing criteria is assumed for data transmission to BS from CHs.
5. After each round of data collection and transmission to BS, determine distortion function $dis(f)$ values and check whether they exceed their respective thresholds.
6. If yes in 5, carry out CH local delegation in all clusters by nominating the CN based on, say, the highest residual energy / maximum degree mixed criteria as determined by the outgoing CH and go to step 4. If no more CN is available in any cluster, go to step 8.
7. If no in 5, go to step 4 for the next round of data collection. This is the global re-clustering phase.
8. Collect data on the number of nodes and determine average lifetime.
9. Go to step 1. Exit if enough data is available to determine the expected average lifetime of node.
10. Compare the expected lifetime computed empirically with the known methods.

IV. Simulation Results

The mathematical formulations discussed in Section III have been tested for their practical effectiveness through MATLAB simulations. We simulated HRA and LEACH protocol for a random network of 100 nodes spread over a 100x 100 unit area. Unlike LEACH, HRA uses multi-hop communication and rely on a synergic balance of local delegation and global re-clustering in order to achieve longevity of life of the network. Moreover, nodes are smartly duty-cycled based on U-Connect-C protocol [19][20]. However, for simulation purposes, an average duty cycle of 40 % is considered. Table I shows the different parameters considered and their values adopted for the formation of the network scenario. Fig. 1 shows a graph of average residual energy of a node. The average energy of a node at 1000th second is 0.4775J in LEACH whereas in HRA it is 0.6984J. In LEACH energy is dissipated at a faster rate so that the average node depletes all its energy by the 2250th sec.

The depletion of energy happens in HRA at a much slower pace with all energy depleted at 4638th s. This shows the power retention capability of HRA. Fig. 2 shows how HRA outperforms LEACH in prolonging the life of the network. Network lifetime is quantitatively analyzed by considering the time elapsed before 1st node death and that before 50% node death.

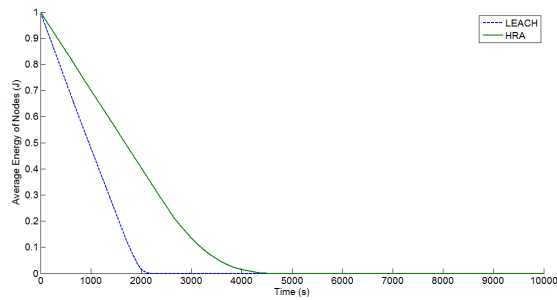


Fig. 1. Average Energy of a Node in a Network

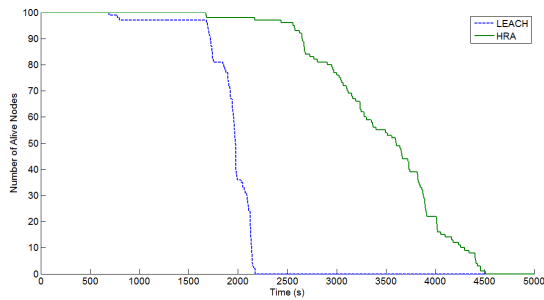


Fig. 2. Network Lifetime

In HRA, the 1st node death and 50% node death happen at 1681 and 3606 s respectively whereas the same in LEACH are at 691 and 1975 sec respectively.

However, Fig. 3 shows the throughput or number of received packets in HRA is moderately lesser compared to that in LEACH. At 1000th sec HRA gives a throughput of 4.99e+04 and LEACH gives 9.956e+04. It is to be noted that while the throughput of HRA is nearly half of LEACH, the 50 % lifetime of HRA is more than double that of the LEACH. Fig. 4 shows the number of data packets received at Base Station before 1st node death and 50 % node death. These values are 6.91e+04 and 1.88e+05 for LEACH and 8.482e+04 and 1.636e+05 for HRA respectively. Fig. 5 finds optimum number of clusters to maximize the lifetime of network as per our algorithm. It is observed that for a node population of 100 in a 100×100 area with the other assumed parameters, the optimum value of number of clusters is 10 for maximum lifetime. Thus network life depends on the number of clusters in a network. Fig. 6 depicts average energy dissipated per rounds for different cluster population in a network.

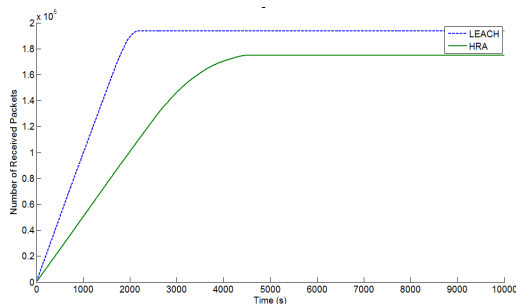


Fig. 3. Total Throughput of the network

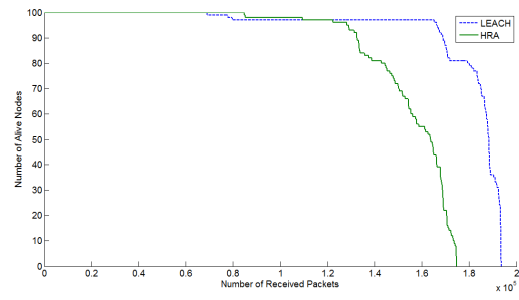


Fig. 4. Throughput of the Network Vs No. of Alive nodes

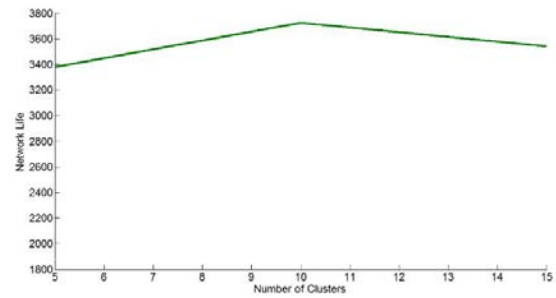


Fig. 5. No. of clusters Vs Network life

It is found that for a node population of 100 in a 100×100 area, the optimum value of number of clusters is 10 for minimum dissipation of energy by a node. Distortion threshold also has an impact on the Network lifetime as per the graph depicted in Fig. 7. It is clear from the graph that an optimum value of 0.4 for distortion threshold yields enhancement in network lifetime.

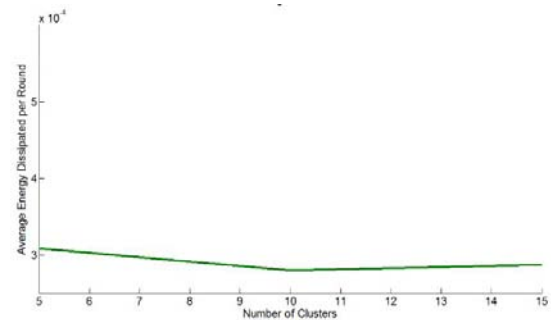


Fig. 6. Effect of cluster population on average energy dissipation of node

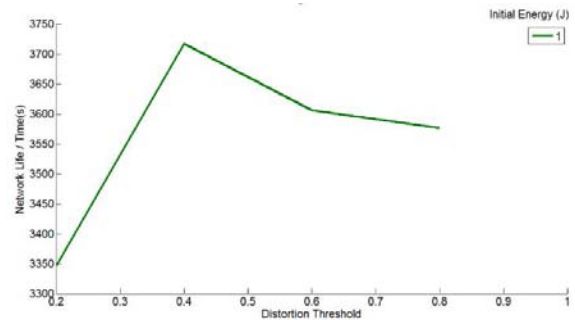


Fig. 7. Effect of Distortion threshold on Network Life time

V. Conclusion

This paper attempted to study the effect of conglomerating local cluster head delegation and global re-clustering in order to enhance the lifetime of the wireless sensor network. Global monitoring is designed based on a function concept in the context of mapping between two metrics space characterized by the state of the CH network before and after data transmission. We could arrive at an optimum value of distortion threshold that significantly enhances the network lifetime. For future work a metric based on Gromov-Hausdorff metric for global monitoring and re-clustering is to be considered.

References

- [1] J.A.Stankovic, Ldouo, *Wireless Sensor Networks*, IEEE Computer Magazine, Vol.41, n.10, 92-95, 2008.
- [2] A. D. Amis and R. Prakash, *Load Balancing Clusters in Wireless Ad Hoc Networks*, in *Proceedings of ASSET*, pp. 25-32, 2000.
- [3] A. D. Amis, R. Prakash, T. H. P. Yuong, and D. T. Huynh, *Max-Min D-Cluster Formation in Wireless Ad Hoc Networks*, in *Proceedings of IEEE INFOCOM*, Vol.1, pp. 32-41, 2000.
- [4] M. Chatterjee, S. K. Das, and D. Turgut, *WCA: A Weighted Clustering Algorithm for Mobile Ad hoc Networks*, *Journal of Cluster Computing*, pp. 193-204, 2002.
- [5] Saravanakumar, R., Mohankumar, N., Raja, J., An optimal cluster head selection technique adopted node activation protocol for lifetime improvement in wireless sensor networks, (2013) *International Review on Computers and Software (IRECOS)*, 8 (6), pp. 1382-1389.
- [6] S. Banerjee and S. Khuller, *A Clustering Scheme for Hierarchical Control in Multi Hop Wireless Networks*, *Proceedings of IEEE INFOCOM*, Vol.2, pp. 22-26, 2001.
- [7] M.Gerla, T. J. Kwon, and G. Pei, *On demand Routing in Large Ad hoc Wireless Networks with Passive Clustering*, in *Proceedings of WCNC*, Vol.1, pp. 23-28, 2000.
- [8] Shankar, T., Shanmugavel, S., Karthikeyan, A., Hybrid approach for energy optimization in wireless sensor networks using PSO, (2013) *International Review on Computers and Software (IRECOS)*, 8 (6), pp. 1454-1459.
- [9] Maizate, A., El Kamoun, N., A new metric based cluster head selection technique for prolonged lifetime in wireless sensor networks, (2013) *International Review on Computers and Software (IRECOS)*, 8 (6), pp. 1346-1355.
- [10] Heinzelman, Wendi B., Anantha P. Chandrakasan, and Hari Balakrishnan. *An application-specific protocol architecture for wireless microsensor networks*. *Wireless Communications, IEEE Transactions on* 1.4 pp. 660-670, 2002.
- [11] R. V. Biradar, S. R. Sawant, R. R. Mudholkar, and V. C. Patil, *Multihop routing in self-organizing wireless sensor networks*, in *Proc. IJCSI International Journal of Computer Science Issues*, Vol. 8, (issue 1), pp. 155-164, January 2011.
- [12] Ossama Younis, Sonia Fahmy, *HEED: A Hybrid, Energy-Efficient Distributed Clustering Approach for Ad Hoc Sensor Networks*, *IEEE Transactions on Mobile Computing*, Vol.3, n.4, pp.366-379, 2004.
- [13] Fan Xiangning; Song Yulin, *Improvement on LEACH Protocol of Wireless Sensor Network*, *Sensor Technologies and Applications, International Conference on Sensor Comm.*, pp.260-264, 2007.
- [14] M. J. Handy, M. Haas, D. Timmermann, *Low Energy Adaptive Clustering Hierarchy with Deterministic Cluster-Head Selection Proc. IEEE Conference on Mobile and Wireless Communications Networks, Stockholm, Erschienen*, September 2002.
- [15] Loscri, V., G. Morabito, and S. Marano. *A two-levels hierarchy for low-energy adaptive clustering hierarchy (TL-LEACH)*. *IEEE Vehicular Technology Conf.*, Vol. 62. no. 3, IEEE; 1999, 2005.
- [16] G. S. Tomar; & Shekhar Verma, *Dynamic multilevel hierarchal clustering approach for wireless sensor networks*, *UKSim 11th International Conference on Computer Modelling and Simulation*, 2009.
- [17] Burago, Dmitri, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*. Vol. 33. (Providence: American Mathematical Society, pp. 241-260, 2001.
- [18] Xiaorong Zhu; Lianfeng Shen; Yum, T.-S.P., *Hausdorff Clustering and Minimum Energy Routing for Wireless Sensor Networks, Vehicular Technology, IEEE Transactions on* Vol.58, n.2, pp.990-997, Feb. 2009.
- [19] Aby K. Thomas, R. Devanathan, *Variable Duty-Cycle Based Efficient Network Discovery in WSN*, *European Journal of Scientific Research*, Vol. 93, N. 2, pp.266-278, December 2012.
- [20] Aby K Thomas, R Devanathan, *Energy Efficient U-Connect-C Protocol for Neighbourhood Discovery in a Clustered WSN Journal of Emerging Technologies Image Processing and Networking* Vol. 8, special issue 1, pp. 210-214.
- [21] Alla, S.B., Ezzati, A., A QoS-guaranteed coverage and connectivity preservation routing protocol for heterogeneous wireless sensor networks, (2012) *International Journal on Communications Antenna and Propagation (IRECAP)*, 2 (6), pp. 363-371.
- [22] Eroglu, A., Design of wireless data acquisition sensor system for health care applications, (2012) *International Journal on Communications Antenna and Propagation (IRECAP)*, 2 (6), pp. 386-391.
- [23] Krief, F., Bennani, Y., Gomes, D., Neuman de Souza, J., LECSOM: A low-energy routing algorithm based on SOM clustering for static and mobile wireless sensor networks, (2011) *International Journal on Communications Antenna and Propagation (IRECAP)*, 1 (1), pp. 55-63.

Authors' information

¹Research Scholar, Faculty of ECE, Sathyabama University, Chennai.
E-mail: abykt2012in@gmail.com

²Professor Emeritus, School of Electrical Sciences, Hindustan University, Chennai.
E-mail: devanathanr@hindustanuniv.ac.in



Aby K. Thomas received his B.E in Electronics and Communication Engineering and M.E degree in Applied Electronics from Madras University, India in 1992 and 2001 respectively. Presently he is working as Professor in the Department of Electronics and Communication Engineering, Hindustan University, Chennai. He is pursuing his doctoral research in Sathyabama University, Chennai, India. He has published many research papers in national and international conferences and journals. He has vast experience in both academia and industry. His research interests are Wireless Sensor Networks and Communication Protocols. He is a fellow of IETE.



R. Devanathan received his Ph.D in Electrical Engineering from Queen's University, Kingston, Ontario, Canada, in 1972. He obtained his M.Sc(Eng) in Electrical Engineering from the same university in 1969. Prior to that, he obtained his B.E and M.E degrees in Electrical Technology and Power Engineering respectively from the Indian Institute of Science, Bangalore, India. He worked in Electronics Commission, Govt. of India in the area of information, planning and analysis during 1973-1978. From 1983 to 2004, he was on the faculty of Nanyang Technological University (NTU), Singapore. Currently he is serving as Professor Emeritus in the Department of Electrical and Electronics Engineering of Hindustan Institute of Technology and Science, Padur, Chennai, India. He has published over 130 papers in International and national conference proceedings and journals. He has been active with the IEEE serving as chapter chair for Control system and Education society chapters. He has chaired and co-chaired International conferences organized under IEEE and NTU.

Smart Sentinel: Monitoring and Prevention System in the Smart Cities

J. M. Sánchez Bernabéu, J. V. Berná Martínez, F. Maciá Pérez

Abstract – Today, faced with the constant rise of the Smart cities around the world, there is an exponential increase of the use and deployment of information technologies in the cities. The intensive use of Information Technology (IT) in these ecosystems facilitates and improves the quality of life of citizens, but in these digital communities coexist individuals whose health is affected developing or increasing diseases such as electromagnetic hypersensitivity. In this paper we present a monitoring, detection and prevention system to help this group, through which it is reported the rates of electromagnetic radiation in certain areas, based on the information that the own Smart City gives us. This work provides a perfect platform for the generation of predictive models for detection of future states of risk for humans. **Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords: Smart City, Electromagnetic Radiation, Sensor System, Electromagnetic Hypersensitivity, Electro-Sensitivity, Monitoring, Detection, Risk, Internet of Things, Human

I. Introduction

Smart Cities appeared by the need to optimize the efforts of the cities in the development of policies that focus on: IT-education, IT-infrastructure, IT-economy and quality of life [1]. The process of incorporating and use in a massive way the IT in the cities, it is called 'smartification' by [2], [3] and the own *Institute of Electrical and Electronics Engineers* (IEEE) [4], and implies that a city needs a broad deployment of systems and techno-logical devices throughout the city, which can help you transform it and adapt to their sustainability needs to better manage their resources and energy sources [5], [6].

In this process aims to achieve an efficient management of all areas of the city: ur-ban planning, infrastructure, transport, education, health, public safety, energy, natural environment and quality of life by meeting their needs and those of its citizens.

These systems make use of *machine to machine infrastructure* (M2M) [16], which al-lows the integration in the platform of devices defined in a Smart City, connecting via gateways M2M or through the interconnection between heterogeneous devices.

Nevertheless, the technology used is capable of generating a complex combination of weak electric and magnetic fields, also called electro pollution, that normally do not affect the majority of the population but there is a collective of people, according to studies by [7], [8], [17], which is affected by this exposure studied by [18]-[20], presenting symptoms such as headaches, memory loss, sleep disorder, blurred vision, nausea, or fatigue.

It has been suggested that exposure to magnetic fields at power frequencies (50/60 Hz), that is to say, *extremely low frequency* (ELF), could lead to an increased incidence of cancer in children and other adverse health effects. The evidence comes mainly from epidemiological studies in residential areas. These studies suggest that there is a partnership between children's exposure to ELF magnetic fields and the increased risk of leukemia [9]. *Electromagnetic fields* (EMF) are characterized by its wavelength or frequency in a radio-active or two categories (Table I illustrates the features of each one of them):

- Non-ionizing: low level of radiation which generally sees the human being but without causing serious injuries.
- Ionizing: can alter the DNA due to their potency.

TABLE I
CLASSIFICATION OF ELECTROMAGNETIC FIELDS (SOURCE [10])

Definition	Forms of radiation	Examples
Non-Ionizing <i>Low to mid frequency, which is generally perceived as harmless due to its lack of potency</i>	Extremely Low Frequency (ELF) Radio Frequency(RF) Microwaves Visual Light	Extremely Low Frequency (ELF) Radio Frequency(RF) Microwaves Visual Light
Ionizing <i>Mid to high frequency radiation which can, under certain circumstances, lead to cellular and DNA damage with prolonged exposure</i>	Ultraviolet(UV) X-Rays Gamma	Ultraviolet Light X-ray range between 30 * 10 ¹⁶ * 30 Hz to 10 ¹⁹ Hz Some Gamma Rays

The concept of Smart City allows us to have access to a large volume of data generated by the various types of technological resources in real time, processing them, and developing new analytical tools for a great value to help this group of persons hypersensitive.

To meet the new challenges in the Smart Cities arises the need for control panels or applications, through which users or citizens, with some type of hypersensitivity to this type of radiation, to have information in real time and deferred of the levels of EMF as the zone where people move, to try to avoid its prolonged exposure. There are proposals in which deals with the measurement of EMF for monitoring such as those of Urbinello [11] or Huss [12], but suffer from the fact that require the installation of new measurement devices or Smart Meters that further enhance the magnitude of the EMF radiation.

In order to give a solution to this problem, this work provides a monitoring and in-formation system which allow us to measure the electromagnetic radiation level in buildings and cities, allowing its use in both real-time and deferred.

This system will allow you to use a smartphone by way of such tool that allows us to define an alert level to inform users of the proximity of the antennas and the levels of radiation that can generate, advising as well to the users of the possibility of being affected by this radiation, this tool will serve as a basis for future lines of development which may be incorporated into the tool to other risk factors based on the monitored data and even the generation of predictive models based on the existing historical data.

II. Proposal: Smart System Sentinel EMF

The system that we present what we have called Smart Sentinel: monitoring of risks system in Smart Cities, and in particular we are going to focus on the Non-ionizing EMF fields generated by the different electrical and electronic devices existing anywhere.

We propose a system for process, analyze, monitor and generate scorecards in real time and deferred, in which the user may display the radiation levels caused by the non-ionizing EMF fields in any area where there is a smart infrastructure. With this approach the users will be able to avoid prolonged exposure in certain areas, as to day of today do not have that knowledge, and alleviate the symptoms produced by electromagnetic hypersensitivity. To achieve the objective of our proposal we relied on Smart City generic architectural view defined by levels as proposed in [9].

Smart Sentinel System acts as a layer to cross all the levels defined in a Smart City architecture (see Figure 1 and Figure 2). This layer is fed by all the levels defined in it, from the level of monitoring in which you specify all the sensors and electronic devices and infrastructure, the level of business where we will store and we will process the data obtained in the previous level, the level of implementation where we will generate value-added

resources such as web applications, standalone and mobile apps, and the level of communications which will allow you to get to know the infrastructure of internetworking, our aim is to be able to have all the necessary information about the technical characteristics of all the devices, the data recorded by each of them, the data and use of radiation emitted at each moment and the geographical positions, in order to process and analyze the information and offer services to users and systems or devices that can consume the type of information that we provide.

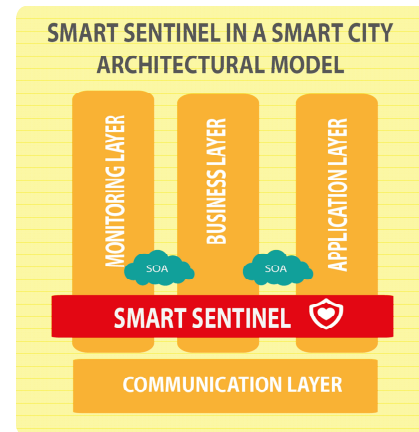


Fig. 1. Smart Sentinel on Smart City architecture

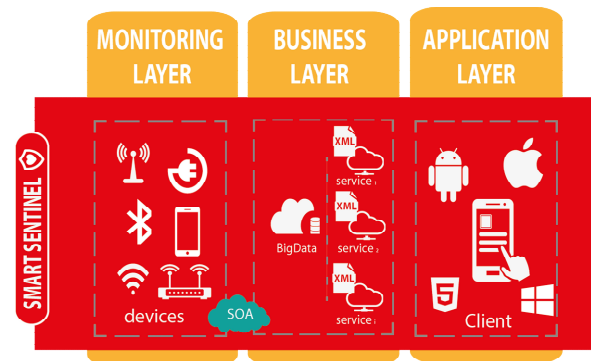


Fig. 2. Smart Sentinel internal level on Smart City architecture

At the internal level Smart Sentinel System makes use of all information generated by all levels of the Smart City architecture making intensive use of the level of business, this level will be responsible for generating the information in web services.

II.1. Monitoring Level

In this level we use sensors, wireless systems, internetworking devices and electrical networks, which receive and transmit information and energy respectively for the environment in which have been integrated.

The following devices are included in the list: smartphones, handhelds, and tablets that are connected via wifi and make use of the network.

These two sets of devices provide us the information by which we are going to perform the implementation of

our system.

The wireless systems employees are able to provide real time information about connections made and the level of data issued by each one of them, have a *Service-oriented architecture* (SOA) Services that feed information to business level for the subsequent analysis of the data proposed by [15].

By the electrical networks we know information on the purchases made in real time, the infrastructure deployed throughout the area and the power handling. As we have explained in section 1, which describes the state of the art, these technologies are basically M2M and services are available to interact with the business layer.

As we have said before in the preceding section, we don't employ an external measurement devices such as Smart Meters, already they would add a higher rate of radiation in the Smart City and therefore harmful to the user.

II.2. Business Level

This level is responsible for collecting, standardizing, process, store, analyze isolated data and massive and transform them into useful information in addition under the paradigm of service, and provide services that provide interoperability, integration, expandability and compatibility between layers.

A part of the data provided by sensors, wireless devices, and internetworking devices, this layer will have other detailed information such as: the technical sheets of all types of devices (density of emitted power, range expansion of the signal, channels, gain, sensitivity of the receptors, operating temperature, among others) and historical information about power consumed.

With all this information and storage systems and analysis can be employed as relational database management system (RDBMS), Data Mining, Data Warehouse, Big Data, we can generate information and offer it through the SOA paradigm to the clients or devices who want to make use of it.

For the distribution of data and services and applications, you can use the cloud as a common platform for data management and distribution as proposed by Armbrust in [13] or Buyya in [14].

II.3. Application Level

Through the development of applications for all types of devices such as smartphones, tablet, and desktop, the user will be able to dispose of information in real-time, historical or deferred about electromagnetic radiation and the areas with the greatest influx according to the need for it, because that will be able to make use of historical data for the generation of more reliable statistics and build predictive models.

The application layer will provide information about the status of a given area and, through a system of reminders, will offer the information to user if you are coming to the spaces with the highest rate of radiation.

This application can work in online mode, but also in offline mode, for those users who are unable to make use of the data connection at all times because of their hypersensitivity, down-loading all the data needed to display the maps EMF.

This layer combines information from different sources to provide a greater level of detail, such as:

- Network Connections of networking devices
- Sheet of technical characteristics of each device
- Network Connections for smartphones, tablets and pc
- Electrical Connections in different locations
- Historical connections and electrical consumption
- Drivers of consumption of the devices themselves

For the distribution of data and services and applications, you can use the cloud as a common platform for data management and distribution.

II.4. Communications Level

As mentioned above, the support of this level to our proposal is crucial, as it will support services transactions that are generated by the services layer. At this level makes use of interconnect systems equipment and internetworking.

III. Prototype

In order to test our system we have developed an application for Android devices. This application makes use of the service in real-time geolocation of the mobile device, or in offline mode if you want to, and the Google API to provide information about the index of radiation to which you are subject, both inside and outside of the area.

In this case a simulation has been made in the facilities of the General Library of the University of Alicante, which has a large amount of wireless devices, servers, routers, and electrical panels in certain areas of the same.

III.1. Characteristics of the information

To provide actual data to users, we have used the information provided by each of the devices on which we rely to generate our maps of EMF radiation in our prototype. We can distinguish two different types of sources of information:

Data streaming

- *Wireless access points*: for these types of devices we get existing connections in every instant in the wireless nodes, MAC, IP, or the name of the connected device. These data are obtained by the level of business and, after its convenient processing, services are offered through the level of application
- *Processing devices*: the services of business level are responsible for monitoring devices that are active on our system, such as routers, application servers, storage servers, information panels and in general all those devices that are part of the infrastructure to

provide an estimate of areas with EMF radiation due to consumption of these devices.

Stored Data

- *Technical Information:* in the level of business will incorporate the technical information on the infrastructure and devices that are part of the facilities to be monitored, as can be routers, antennas, electrical panels, lighting, electrical lines, network wiring and in general any other electronic device.
- *Consumption data:* by the providers of services are obtained information concerning the global electrical consumption, peaks of electrical consumption, consumption by building and consumption of telecommunications network.

It should be noted that the devices connected to the data network can be monitored from the level of business, at least to know if it is active or not. This we know, for example, how many of the computers inventoried in the field of control (for ex-ample the University of Alicante) are consuming resources and issuing EMF radiation

III.2. Viewing the Information

The application is composed of two display modes of the commented previously information. A spatial mode in which we can combine all the information on the different types of data to give a vision to the user of the space that surrounds him and the EMF influence in your environment, and other display in meter mode that provides a simplified view of the EMF level of the current geographic location.

In the spatial mode of display we inform the user of the areas of radiation in the location in which it is situated. Shown here is a map, using the Google API, which highlights the user's current location and placed various elements that produce influence EMF in addition to various levels. One of the most important elements are the nodes of wireless access, in Fig. 3 we can see the representation of the wireless access points around the user that is currently in the general library of the UA campus.



Fig. 3. Spatial Mode

Each of these nodes is represented by an area of influence that takes a color more or less intense depending on the signal from the antenna, the number of Wi-Fi connections that are enduring that node and the activity of the device itself.

On this map are formed by juxtaposing the different elements that produce the risk that we are being monitored, in addition to nodes of Wi-Fi access, also shows the electrical devices, areas of transformers, electrical consumption of a building, and the processing devices such as routers and servers. In Fig. 4(a) we can see a complete representation of all the elements of influence EMF that the application collects, along with the current position of the user, in this way through the greater or lesser intensity of the color that represents the EMF radiation a user can have an idea of which areas are potentially dangerous for and therefore should be avoided, also this figure represents the streaming mode of the application, which displays the data in real time.



Figs. 4. (a) Streaming Mode (b) Offline Mode

In Fig. 5 we can visualize the configuration screen of the application, which allows us to mark items we want to represent on the map.

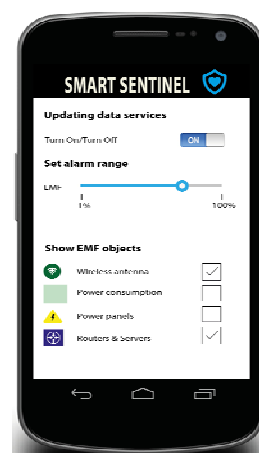
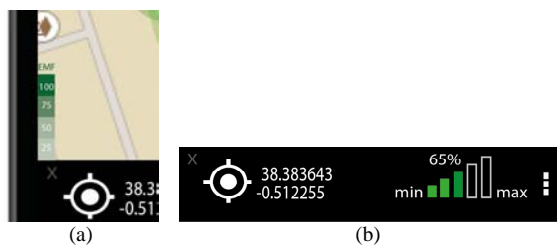


Fig. 5. Setup Mode

The application enables us to operate in offline mode or deferred (see Fig. 4(b)). This mode does not perform any data update and does not receive data from the streaming services, i.e. does not display data of incidence EMF in real time.

When the application is in this mode, the map does not update the elements whose information depend on this service of streaming data, such as the signal strength of the wireless access nodes, or just stop render if we have no information about them. Only shows the data received in your last connection and static data tab as technical and historical data of the devices that emit EMF. In the lower left area of the screen we can see the map legend of EMF radiation, which has a scale of colors ranging from the 1, minor radiation index, to 100, area of greatest impact (see Fig. 6(a)).

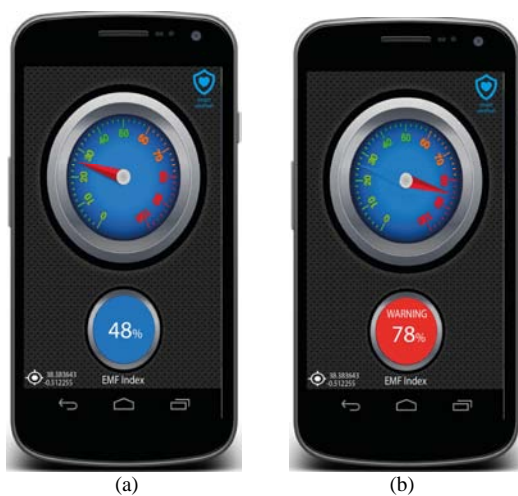


Figs. 6. (a) Legend and (b) radiation index scale

In the lower area of the screen we can see the position of the user in coordinates along with the value of EMF radiation in that position, which has a scale of values ranging from 1 (min), less radiation index, to 100 (max), area of greatest impact (see Fig. 6(b)).

In **Meter mode** the application Smart Sentinel will allow us to inform the user of the index of EMF radiation in its current location.

It has an alert mode, see left side of Fig. 7(a), which prompts you through a red button (see Fig. 7(b)), if you have exceeded the limits indicated in the configuration of smart alarms Sentinel.



Figs. 7. (a) Meter Mode without alarm.
(b) Meter Mode with alarm

The alert level is customizable via the configuration screen that has been seen in Fig. 8, thereby the user to set the maximum level of radiation in which he wants to be contacted depending on your tolerance.

In addition to the two display modes shown, there is also an area of customization and configuration as we have seen in the Fig. 5, and information display (Fig. 8) where are identified the items represented in the application.

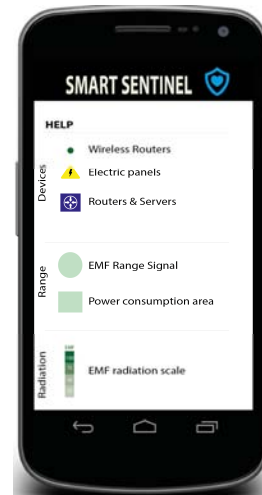


Fig. 8. Help Screen

IV. Conclusions and Future Work

In this paper, a proposal has been presented for the development of a monitoring and information, in real-time or delayed, of the index of electromagnetic radiation of buildings and areas of a Smart City, embodied for our university, and dedicated to users who suffer the pathology of electromagnetic hypersensitivity.

The approach has allowed us to define a prototype capable of taking advantage of existing infrastructures and the historical and technical data of the Smart Cities, without adding new elements such as the *Smart Meterings*, which increases the level of existing radiation. The proposed system has advantages over the actual monitoring system because our system don't require the installation of new measurement devices. We seek to do so through the information supplied by the own devices that form the infrastructure, covering the development under the architectural model of Smart City, in which the initiatives it are built in an integrated manner. Our system makes use of all information provided by the various levels as the basis for the Smart City, and therefore is able to integrate in the proposal any other sensing element or data that you have accommodated within the paradigm Smart City.

In the short term we are considering incorporating in the monitoring system of other risk factors such as could be pollution, pollen, noise or environmental factors. To do this in the proposal it would suffice to integrate other sensing devices and data services to reflect on our map the various risk factors in addition to the EMF.

In the medium term, using all the information stored and with enough baggage, we could generate prediction models (for example as do meteorological models), to facilitate not only monitoring but also being able to predict the occurrence of these risk factors in the areas of interest to users, thus generating preventive tools.

Acknowledgements

This work has been performed within project *Smart University* funded from the Vice President office for Information Technology at the University of Alicante.

References

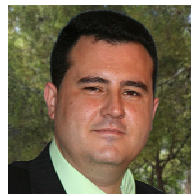
- [1] A. Mahizhnan, Smart Cities: The Singapore Case, *Cities*, Vol. 16, n. 1, pp. 13-18, 1999
- [2] S. Bartolini, B. Milosevic, A. D'Elia, E. Farella, L. Benini, T. S. Cinotti, Natural interaction in Reconfigurable smart environments: Approach and prototype implementation, *Personal and Ubiquitous Computing*. Vol. 16, n. 7, pp. 943-956, 2012.
- [3] S. Ruiz-Romero, A. Colmenar-Santos, F. Mur-Pérez, A. López-Rey, Integration of distributed generation in the power distribution network: The need for smart grid control systems, communication and equipment for a smart city, *Renewable and Sustainable Energy Reviews*, Vol. 38, pp. 223-234, 2014.
- [4] IEEE Smart Cities Initiative [Online], 2014. <http://smartcities.ieee.org/home/ieee-invites-global-municipalities-to-engage-in-new-ieee-smart-cities-initiative.html>
- [5] J. P. Vasseur, A. Dunkels, in Jean-Philippe Vasseur and Adam Dunkels, Morgan Kaufmann (ed.), *Smart Cities and Urban Networks, In Smart Objects with Interconnecting IP*, (Boston 2010, 335-251).
- [6] D. Urbinello, J. Wout, A. Huss, L. Verloocke, J. Beekhuizen, R. Vermeulen, L. Martens, M. Rösli, Radio-frequency electromagnetic field (RF-EMF) exposure levels in different European outdoor urban environments in comparison with regulatory limits, *Environment International*, Vol. 68, pp. 49-54, 2014.
- [7] M. Havas, Electromagnetic hypersensitivity: biological effects of dirty electricity with emphasis on diabetes and multiple sclerosis, *Electromagn Biol Med*, Vol. 25, n. 4, pp. 259-268, 2006.
- [8] A. Caragliu, C. Del Bo, P. Nijkamp, Smarts cities in Europe, *Journal of urban technology*, Vol. 18, n. 2, pp. 65-82, 2011.
- [9] I. Calvente, M.F. Fernandez, J. Villalba, N. Olea, M.I. Núñez, Exposure to electromagnetic fields (non-ionizing radiation) and its relationship with childhood leukemia: A systematic review. *Science of the total environment* Vol. 408, Issue 16, 15, pp. 3062-3069, 2010.
- [10] National Institute of Environmental Health Sciences [online] 2014 <http://www.niehs.nih.gov/health/topics/agents/emf/>
- [11] D. Urbinello, A. Huss, J. Beekhuizen, R. Vermeulen, M. Rösli, Use of portable exposure meters for comparing mobile phone base station radiation in different types of areas in the cities of Basel and Amsterdam. *Science of Total Environment*, Vol. Jan 15, pp. 468-469:1028-1033, 2014.
- [12] A. Huss, D. Urbinello, W. Joseph, L. Verloock, J. Beekhuizen, R. Vermeulen, L. Martens and M. Rösli, Radio-frequency electromagnetic field (RF-EMF) exposure levels in different European outdoor urban environments in comparison with regulatory limits. *Environment International*. Vol. 68, pp. 49-54, 2014.
- [13] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, *A view of cloud computing*. Communications of the ACM. Vol. 53, N.4, pp. 50-58, 2010.
- [14] R. Buyya, J. Broberg and A. Goscinski. *Cloud Computing: Principles and Paradigms*. (John Wiley & Sons), 2011
- [15] K. Barry, D. Dick, D. Managing Change with Incremental SOA Analysis, In *The Savvy Manager Guide* (Ed). *Web Services, Service-oriented Architectures, and Cloud Computing (Second Edition)*. (Boston, Morgan Kaufmann, 2013, 113-127.
- [16] A. Elmangoush, H. Coskun, S. Wahle, T. Magedanz. *Design aspects for a reference M2M communication platform for Smart Cities*. Proceedings of the 9th International Conference on Innovations in Information Technology, IIT 2013 (Page: 204-209, Year of Publication: 2013. ISBN: 978-146736203-0).
- [17] A. Ahlbom, U. Bergqvist, J.H. Bernhardt, J.P. Cesarini. Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz). *Health Physics*. Vol. 74, n. 4, pp. 494-521, 1998.
- [18] A. Bürgi, G. Theis, A. Siegenthaler, M. Rösli. Exposure modeling of high-frequency electromagnetic fields. *Journal of Exposure Science and Environmental Epidemiology*. Vol. 18, n. 2, pp. 183-191, 2008.
- [19] D. Urbinello, M. Rösli. Impact of one's own mobile phone in stand-by mode on personal radiofrequency electromagnetic field exposure. *Journal of Exposure Science and Environmental Epidemiology*. Vol. 23, n. 5, pp. 545-548, 2013.
- [20] M. Khalid, T. Mee, A. Peyman, D. Addison, M. Maslanyj, S. Mann. Exposure to radio frequency electromagnetic fields from wireless computer networks: Duty factors of Wi-Fi devices operating in schools. *Progress in Biophysics and Molecular Biology*. Vol. 107, n. 3, pp. 412-420, 2011.

Authors' information



J. M. Sanchez Bernabeu He was born in Alicante (Spain). He received the Bachelor's Degree in Computer Science from University of Alicante in 2010 and Master in Computer Technologies in 2014. Now is student PhD in Information Technologies. And is part of Middleware Group in Department of Computer Technology in University of Alicante. He's

working with M2M Communications, Smart Cities and Internet of the Things.



Jose Vicente Berna-Martinez was born in Spain in 1978. He received his engineering degree and the Ph.D. degree in Computer Science from the University of Alicante in 2004 and 2011 respectively. From 2006 to 2013, he was an Associate Professor at the University of Alicante, currently he is a Assistant doctor. His research interests are in the area of computer

networks, distributed systems, bio-inspired systems and robotics which are applied to industrial problems.



Francisco Maciá-Pérez was born in Spain in 1968. He received his engineering degree and the Ph.D. degree in Computer Science from the University of Alicante in 1994 and 2001 respectively. He worked as System's Administrator at the University of Alicante form 1996 to 2001. He was an Associate Professor from 1997 to 2001. Since 2001, he is a Professor

and currently he is the Vice President for Information Technologies at the University of Alicante. His research interests are in the area of network management, computer networks, smart sensor networks and distributed systems, which are applied to industrial problems.

Hybrid Fusion Technique Using Dual Tree Complex Wavelet Transform for Satellite Remote Sensor Images

G. Dheepa, S. Sukumaran

Abstract – Image Fusion is the process of merging two or more images to form a single image which has all the details of the individual images. This technique is used in satellite remote sensor images to fuse high resolution panchromatic image with the low resolution multispectral image to form a single high resolution multispectral image. Among the existing fusion techniques, wavelet based methods have proved to produce improved results.

The key objective of this paper is to introduce a new hybrid fusion method to fuse PAN image and MS image by using Dual Tree complex wavelet transform (DTCWT) and IHS technique. Firstly the limitations of classical Discrete Wavelet Transform (DWT) are explained in brief. Secondly the key properties of complex wavelets and its theory are described followed by proposing a new fusion method using complex wavelets.

Finally experimental results are evaluated using quality assessment metrics which shows that the proposed method performs remarkably better than the classical DWT method. **Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords: Complex Wavelet Transform, IHS Method, Image Fusion, Remote Sensing

Nomenclature

σ	Standard deviation
h/l	Ratio of the spatial resolution of original Pan and MS images
$C_R^H(i, j)$	Real part of the high resolution Pan image
$C_I^H(i, j)$	Imaginary part of high resolution Pan image
$C_R^L(i, j)$	Real part of low resolution Intensity image
$C_I^L(i, j)$	Imaginary part of low resolution Intensity image
μ	Mean value
M_{MS}	Mean value of multispectral image
M_F	Mean value of fused image
(i, j)	Pixel index
N	Number of bands
CC	Correlation Coefficient
RMSE	Root Mean Square Error
UIQI	Universal Image Quality Index
ERGAS	Relative global dimensional error

I. Introduction

Most of the earth observation satellites such as Spot, Ikonos, Quickbird, and so on record the image in two different modes, a low-resolution multispectral (MS) and high-resolution panchromatic (PAN) mode. A PAN image gives detailed geometric features, while the MS images contain richer spectral information. The objective of image fusion is to combine the panchromatic and the

multispectral information to form a fused multispectral image that retains the spatial information from the high resolution panchromatic image and the spectral characteristics of the lower resolution multispectral image. Applications for integrated image datasets include environmental/agriculture assessment, urban mapping, and change detection [1].

Before fusing two images, it is necessary to perform a geometric registration and a radiometric adjustment of the images to one another. When images are obtained from sensors of different satellites as in the case of fusion of SPOT or IRS with Landsat, the registration accuracy is very important. But registration is not much of a problem with simultaneously acquired images as in the case of Ikonos/Quickbird PAN and MS images. The PAN images have a different spatial resolution from that of MS images. Therefore, resampling of MS images to the spatial resolution of PAN is an essential step in some fusion methods to bring the MS images to the same size of PAN. Many fusion methods have been proposed and as far as the methods discussed earlier in the literature, wavelet transform based methods are found to give better results. The advantage of wavelet transform is that it can analyze signal in both frequency domain and time domain. The wavelet transform comes in many forms.

The most common of them is Discrete Wavelet Transform (DWT). Stephane Mallat proposed the filter-bank implementation scheme of DWT. The concept of Multi-Resolution Analysis (MRA) is also connected to wavelet theory. Though having its advantage, DWT suffers from some major limitations. A slight shift in input signal leads to major variations in the

amplitude of DWT coefficients at different scales.

This results from the down sampling operation at each level [2].

As the DWT filters are separable and real, it cannot distinguish between opposing diagonal directions which results in poor directional selectivity. The wide spacing of wavelet coefficient samples leads to aliasing. Any wavelet coefficient processing will perturb the subtle balance between the forward and inverse transforms, which leads to artifacts in the reconstructed signal.

To a little extent, a few of these limitations can be overcome by using some of DWT's extensions, such as the undecimated DWT which is translation invariant or using the Discrete Wavelet Packet Transform, which offers a better directional selectivity. But a better way of overcoming these limitations is to use Complex Wavelet Transforms (CWT).

II. Complex Wavelet Transform

Complex wavelets can be used to analyze and represent both real-valued signals and complex valued signals. Complex wavelet Transform (CWT) can be broadly classed into two types [3]; Non-Redundant CWT (NRCWT) and Redundant CWT (RCWT). In NRCWT, if the input signal has N samples, the transformation will provide us N output coefficients.

In RCWT, if the input signal has N samples, then we will obtain M output coefficients, with $M > N$. In this case, the original signal is passed through two parallel wavelet filter-bank trees that contain carefully designed filters of different delays that minimize the aliasing effects due to downsampling [4]. The resulting coefficients being complex; the output values of the first tree have real part and the output values of the second tree have imaginary part.

Complex wavelets have not been used widely in image processing due to the difficulty in designing complex filters which satisfy a perfect reconstruction property. To overcome this Nick Kingsbury [5], proposed a dual-tree implementation of the CWT (DT CWT) which uses two trees of real filters to generate the real and imaginary parts of the wavelet coefficients separately as shown in Fig. 1.

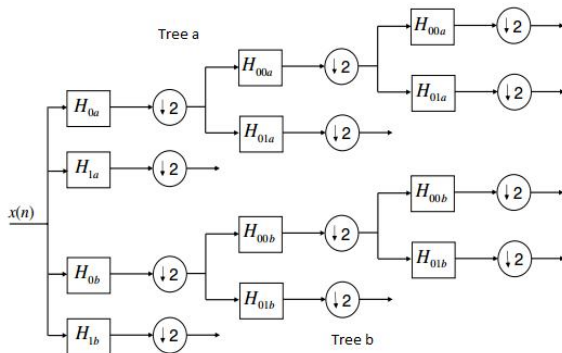


Fig. 1. Three level complex dual tree

DT CWT has the following properties [6]:

- A nearly shift-invariant magnitude with a simple near-linear phase encoding of signal shifts;
- Substantially reduced aliasing;
- Directional wavelets in higher dimensions;
- Moderate redundancy: 2:1 for 1-D (1:2m for m-D);

The redundancy is much less than the $\log_2 N \times$ redundancy of a perfectly shift-invariant DWT, which, moreover, will not offer the desirable magnitude/phase interpretation of the CWT nor the good directional properties in higher dimensions.

III. Dual-Tree Complex Wavelet Transform

The dual-tree CWT employs two real DWTs; the first DWT gives the real part of the transform while the second DWT gives the imaginary part. They use two different sets of filters which are jointly designed so that the overall transform is approximately analytic. A real biorthogonal wavelet transform lacks shift variance but provides perfect reconstruction and no redundancy. Approximate shift invariance can be achieved by doubling the sampling rate at each level of the tree.

The sampling rates can be doubled by eliminating the down-sampling by 2 after the level 1 filters [7].

III.1. 1-D Dual-Tree Complex Wavelet Transform

For one dimensional signal, two parallel wavelet trees are to be computed. There is one sample offset delay between two trees at level 1, which is achieved by doubling all the sample rates.

The shift invariance is perfect at level 1, since the two trees are fully decimated. To get uniform intervals between two trees below level 1, the filters in one tree must provide half a sample delay from those in the other tree. This is able by using odd-length and even-length filters alternatively from level to level in each tree.

❖ *1-D DT CWT Algorithm:*

- At level 1, there is one sample offset between the trees:

$$(a_A^1)_n = (a^0 * h^0)_{2n} \quad (d_A^1)_n = (a^0 * g^0)_{2n} \quad (1)$$

$$(a_B^1)_n = (a^0 * h^0)_{2n+1} \quad (d_B^1)_n = (a^0 * g^0)_{2n+1} \quad (2)$$

- Beyond level 1, there must be half a sample difference between the trees:

$$(a_A^{j+1})_n = (a_A^j * h^e)_{2n} \quad (d_A^{j+1})_n = (a_A^j * g^e)_{2n} \quad (3)$$

$$(a_B^{j+1})_n = (a_A^j * h^e)_{2n+1} \quad (d_B^{j+1})_n = (a_A^j * g^e)_{2n+1} \quad (4)$$

The details d_A and d_B can be interpreted as the real and imaginary parts of a complex process $z = d_A + id_B$.

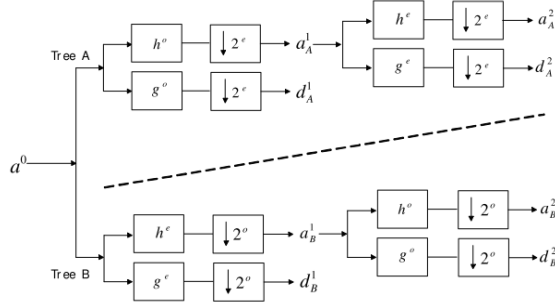


Fig. 2(a). 1-D dual tree complex wavelet transform

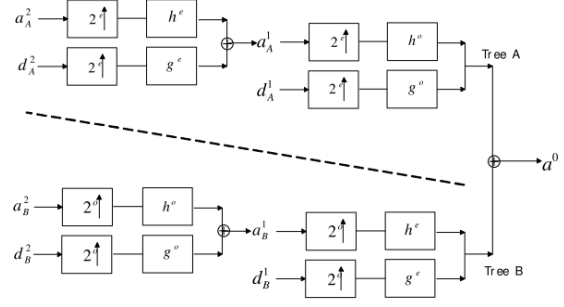


Fig. 2(b). 1-D dual tree inverse complex wavelet transform

The vital property of this transform is that the magnitude of the step response is approximately invariant with the input shift, while only the phase varies rapidly [7].

At level 1, the filter in tree A are odd-length filter, is same to tree B. Beyond level 1, the filters in two trees are different, and they are different between different levels in each tree.

❖ 1-D Inverse DT CWT Algorithm:

To invert the transform, the real part and the imaginary part are each inverted to obtain two real signals.

These two real signals are then averaged to obtain the final output a^0 .

➤ Level $j (j > 0)$:

$$(a_A^j)_n = (\tilde{a}_A^{j+1} * \tilde{h}^e)_n + (\tilde{d}_A^{j+1} * \tilde{g}^e)_n \quad (5)$$

$$(a_B^j)_n = (\tilde{a}_B^{j+1} * \tilde{h}^o)_n + (\tilde{d}_B^{j+1} * \tilde{g}^o)_n \quad (6)$$

➤ At $j = 0$:

$$a_n^0 = \frac{1}{2} \left(\begin{aligned} & \left((\tilde{a}_A^1 * \tilde{h}^o)_n + (\tilde{d}_A^1 * \tilde{g}^o)_n \right) + \\ & \left((\tilde{a}_B^1 * \tilde{h}^e)_n + (\tilde{d}_B^1 * \tilde{g}^e)_n \right) \end{aligned} \right) \quad (7)$$

The 1-D dual tree complex wavelet transform and its inverse is shown in Fig. 2(a) and Fig. 2(b).

III.2. 2-D Dual-Tree Complex Wavelet Transform

To extend the transform to higher-dimensional signals, a filter bank is usually applied separably in all dimensions. In 2-D DT CWT separable filtering is done along columns and then rows.

This operation results in six complex high-pass sub-bands at each level and two complex low-pass sub-bands on which subsequent stages iterate in contrast to three real high-pass and one real low-pass sub-band for the real 2D transform. This shows that the complex transform has a coefficient redundancy of 4:1 or 2m: 1 in m dimensions.

The CWT decomposes an image into a pyramid of complex subimages, with each level containing six oriented subimages resulting from evenly spaced directional filtering and subsampling, such directional filters are not obtainable by a separable DWT using a real filter pair but complex coefficients make this selectivity possible.

The 2-D DWT produces three bandpass subimages at each level, which are corresponding to LH, HH, HL, and oriented at angles of $0^\circ, \pm 45^\circ, 90^\circ$.

The 2-D CWT can provide six subimages in two adjacent spectral quadrants at each level, which are oriented at angles of $\pm 15^\circ, \pm 45^\circ, \pm 75^\circ$ which is shown in Fig 3. The strong orientation occurs because the complex filters separate positive frequencies from negative ones vertically and horizontally. So they won't be aliasing.

Since many of the limitations in DWT are overcome by DTCWT, it is a better choice to use it in the image fusion technique as an alternative to DWT.

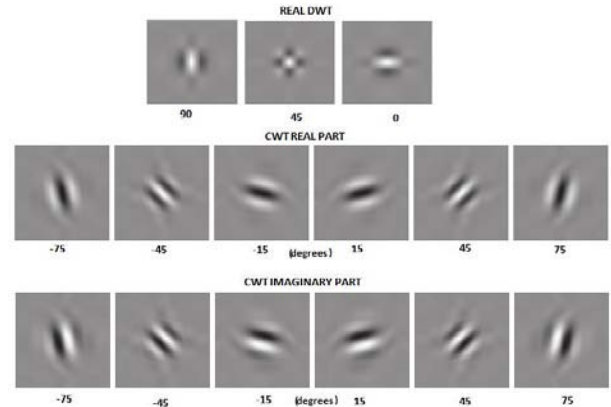


Fig. 3. Filter response showing the orientations of 2D real wavelet filters and complex wavelet filters at level 4

IV. Proposed Fusion Method Using Integration of DT CWT and IHS Technique

The IHS transform is the most commonly used fusion technique because it can effectively separate a standard RGB (Red, Green, Blue) image into spatial (I) and spectral (H, S) information.

But in a stand-alone IHS fusion method the color quality of the fused image strongly depends on the resemblance between the high spatial resolution image (PAN) and the intensity image (I) of the low spatial resolution MS image.

If the grey value distribution of the IHS intensity image is close enough to that of the panchromatic image, the IHS fusion method can well preserve the color information. If there is much difference it will cause a substantial color distortion. For QuickBird and Ikonos images the color distortion is especially significant.

Though wavelet based fusion method can well retain the color information, the spatial detail from a PAN is often different from that of a MS band having the same spatial resolution because of their spectral range difference which introduces some color distortion into the fusion results. Also the integration between color and spatial details appear unnatural. So in order to overcome the limitations and utilize the merits of both methods, we integrate IHS and DT-CWT technique to form a new hybrid fusion method as shown in Fig. 4.

As a preprocessing step, both the PAN and MS images are registered geometrically so that both have the same size.

The fusion procedure is described as follows:

1. Using IHS transform, the MS image is transformed into IHS components.
2. The PAN image is histogram matched to that of the Intensity image (I) to get a new Pan image.
3. The new Pan image and Intensity image (I) are decomposed individually using Dual Tree complex wavelet transform (DTCWT) to obtain complex coefficients with magnitude. The coefficients at point (i, j) of real and imaginary parts in the high resolution Pan image are denoted as $C_R^H(i, j)$ and $C_I^H(i, j)$ respectively. The coefficients at point (i, j) of real and imaginary parts in the low resolution intensity image are denoted as $C_R^L(i, j)$ and $C_I^L(i, j)$ respectively.

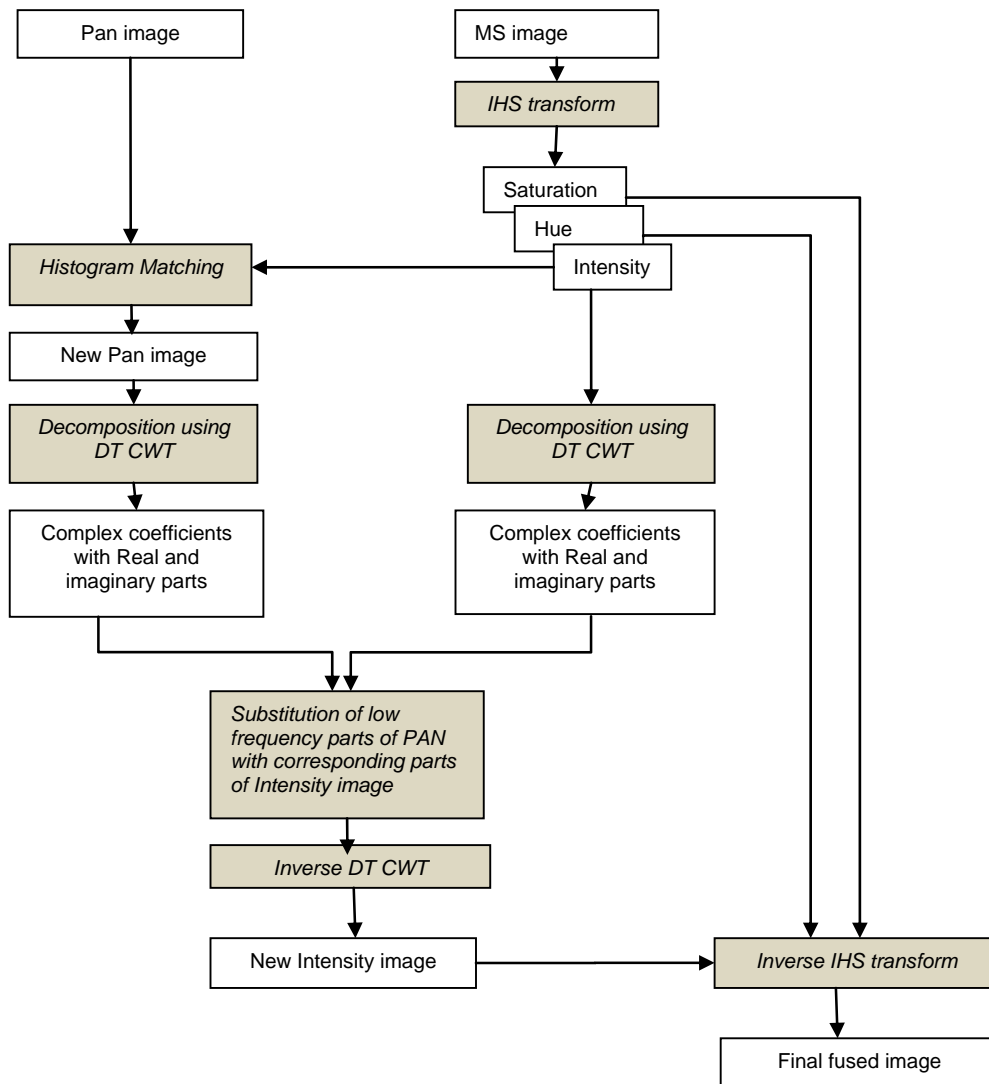


Fig. 4. Flow scheme of proposed DTCWT-IHS hybrid fusion method

4. To inject grey value information of the intensity image into PAN image, the low frequency parts (real and imaginary) of the PAN image are replaced by the corresponding parts of Intensity (I).
5. Inverse DTCWT is applied to obtain a new intensity image which has the spatial detail of PAN image and has a high correlation (similar grey value distribution) with the Intensity (I) image of IHS transform.
6. Using inverse IHS transform, the new intensity together with hue and saturation components are transformed back to RGB space to get the final fused image.

V. Experimental Results

The Ikonos-2 panchromatic band of the 1-m resolution Pan image and the Red, Green, Blue and Near Infrared bands of the 4-m resolution multispectral images are used as experiment data. The 4m resolution multi-spectral bands are resampled to 1-m pixel size before fusion is done. The original panchromatic image and resampled MS image are shown in Fig. 5 and Fig. 6 respectively.

The resampled multispectral and Pan bands are then fused using IHS transform, DWT and the proposed hybrid algorithm. The corresponding fusion results are shown in Fig. 7(a), Fig. 7(b) and Fig. 7(c).

VI. Result Analysis

By comparing the results visually, the IHS method blends the spatial information of PAN and spectral information of MS image.



Fig. 5. Original panchromatic image



Fig. 6. Resampled multispectral image



Fig. 7(a). IHS fusion result



Fig. 7(b). DWT fusion result



Fig. 7(c). Proposed DT CWT-IHS hybrid fusion result

Color distortion appears significantly when compared to the original MS image. The DWT method enhances the spatial information and color information but still a little color distortion exists which is apparently visible.

The proposed hybrid fusion method integrates the spatial information of the PAN and spectral information of the MS image into a single fused image very well. The colors in the fusion result look close to that of the original MS images, and spatial details as detailed as the original PAN image. It well preserves both the spatial and original spectral content.

To evaluate the performance of each fusion method quantitatively, a statistical comparison is done. Mathematical methods were used to judge the quality of fused image in respect to their improvement of spatial resolution while preserving the spectral content of the data. The correlation coefficient is most widely used similarity metric.

Another commonly used assessment metric is the Root Mean Square Error (RMSE). Multimodal statistical indices such as UIQI and ERGAS have also been calculated to compare the fusion results.

➤ *Correlation Coefficient (CC)*

The correlation coefficient measures the closeness or similarity in small size structures between the original and the fused images. It ranges from -1 to +1. Values close to +1 indicates that they are highly similar while the values close to -1 indicate that they are highly dissimilar:

$$CC = \frac{\sum_{i=1}^N \sum_{j=1}^N (MS_{i,j} - \bar{MS})(F_{i,j} - \bar{F})}{\sqrt{\sum_{i=1}^N \sum_{j=1}^N (MS_{i,j} - \bar{MS})^2 \sum_{i=1}^N \sum_{j=1}^N (F_{i,j} - \bar{F})^2}}$$

where CC is the Correlation Coefficient, F is the fused image and i and j are pixels, MS is the multispectral data.

➤ *Root mean square error (RMSE)*

The RMSE was computed from the standard deviation and the mean of the fused and the original image:

$$RMSE = \sqrt{(\sigma_{MS} - \sigma_F)^2 + (M_{MS} - M_F)^2}$$

where σ_{MS} is standard deviation of multispectral image, σ_F is standard deviation of fused image; M_{MS} equals to mean value of multispectral image, and M_F equals to mean value of fused image. The best possible value is zero.

➤ *ERGAS*

ERGAS is the abbreviation of Erreur Relative Globale Adimensionnelle de Synthèse (Relative global dimensional error). It calculates the amount of spectral distortion and the formula is given by:

$$ERGAS = 100 \frac{h}{l} \sqrt{\frac{1}{N} \sum_{i=1}^N \left[\frac{RMSE(B_i)^2}{(M_i)^2} \right]}$$

where N is the number of bands involved in fusion, h/l is the ratio of the spatial resolution of original Pan and MS images. M_i is the mean value for the original spectral image B_i . ERGAS values larger than 3 stand for synthesized images of low quality, while less than 3 represent a satisfactory quality [20].

➤ *Universal Image Quality Index (UIQI)*

The UIQI [13] measures how much of the salient information contained in reference image is transferred to the fused image. UIQI is devised by considering loss of correlation, luminance distortion and contrast distortion.

The range of this metrics varies from -1 to +1 and the best value is 1:

$$UIQI = \frac{\sigma_{AB}}{\sigma_A \sigma_B} \cdot \frac{2\mu_A \mu_B}{\mu_A^2 + \mu_B^2} \cdot \frac{2\sigma_A \sigma_B}{\sigma_A^2 + \sigma_B^2}$$

where σ represents the standard deviation and μ represents the mean value. The first term in RHS is the correlation coefficient, the second term represents the mean luminance and the third measures the contrast distortion. The range of this metrics varies from -1 to +1 and the best value is 1. The quantitative evaluation results are shown in Table I. To simplify the comparison of the different fusion methods, the values of quantitative indicators CC, RMSE, ERGAS and UIQI of the fused images are provided as chart in Fig. 8(a), Fig. 8(b), Fig. 8(c) and Fig. 8(d) respectively.

TABLE I
EVALUATION RESULTS OF FUSION METHODS

Methods / Assessment Indices	IHS	DWT	IHS-DT CWT
CC	0.3847	0.9327	0.9743
RMSE	7.3506	0.9413	0.5426
ERGAS	18.6865	2.9421	2.2468
UIQI	R	0.4687	0.8775
	G	0.4521	0.8213
	B	0.2213	0.7649
	NIR	0.8452	0.8502

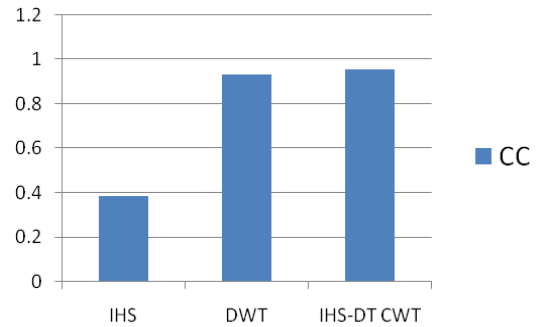


Fig. 8(a). Correlation Coefficient for fusion methods

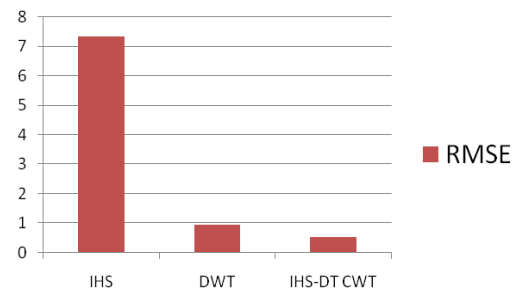


Fig. 8(b). RMSE value for fusion methods

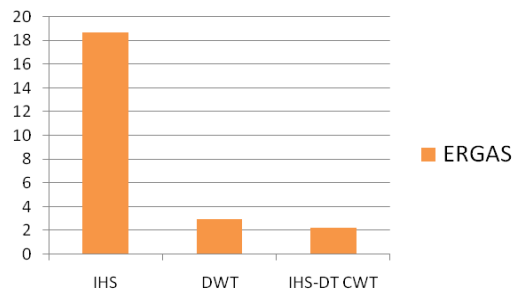


Fig. 8(c). ERGAS value for fusion methods

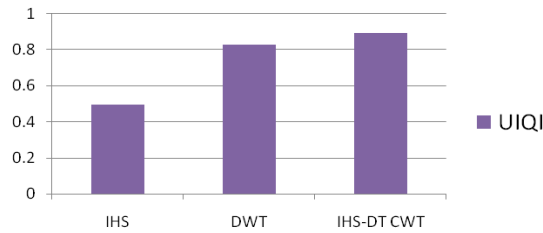


Fig. 8(d). UIQI of fusion methods

From the visual inspection results and the quantitative results, it can be seen that the experimental results are in compliance with the theoretical analysis.

When comparing the results, the CC of the proposed hybrid method is significantly higher which shows that it preserves the color of the MS image better than the other two fusion methods. The RMSE between the hybrid fusion image and the original MS image shows a lower value than other methods which means the color distortion of the fused image is less. When the ERGAS value is compared, proposed method has the lowest value and is less than three which represents satisfactory quality.

The UIQI for the proposed method is higher than other two methods and is nearest to one which shows a large amount of salient information is transferred to the fused image. The results reveal that the proposed hybrid IHS-DTCWT method performs better than the other two methods and produces the fused image closest to those the corresponding multi-sensors would observe at the high-resolution level.

VII. Conclusion

This paper presents a new hybrid approach to fuse high-resolution panchromatic image and low resolution multispectral image using integration of IHS-DTCWT technique. The performance of the proposed fusion method is analyzed using various quantitative indicators and compared with the commonly used stand-alone IHS and DWT method. The fusion techniques IHS and DWT provide superior visual high resolution MS images but ignore the requirement of high-quality synthesis of spectral information producing more spectral distortion.

The proposed hybrid method shows better performance in terms of the high-quality synthesis of spectral information while retaining the spatial content of the Pan image and it provides better result qualitatively and quantitatively.

This work infers that the new hybrid fusion technique using IHS-DTCWT offers computationally efficient image fusion techniques and intends to focus more on complex wavelet based fusion techniques to improve the existing.

References

[1] Edwards and P.A.Davis, "The use of Intensity-Hue-Saturation transformation for producing color shaded-relief images,"

Photogramm. Eng. Remote Sens., vol. 60, no. 11, pp. 1369–1374, 1994.

[2] Ivan W. Selesnick, Richard G. Baraniuk, and Nick G. Kingsbury "The Dual –Tree complex wavelet transform," *IEEE signal processing magazine*, pp.123-151, 2005.

[3] P. D. Shukla, "Complex wavelet transforms and their applications", PhD Thesis, The University of Strathclyde, 2003.

[4] Peter de Rivaz and Nick Kingsbury, "Bayesian image deconvolution and denoising using complex wavelets", *Proc. IEEE Conf. on Image Processing*, Greece, paper 2639, 2001.

[5] N.G. Kingsbury, "Image processing with complex wavelets," *Philos. Trans. R.Soc. London A, Math. Phys. Sci.*, vol. 357, no. 1760, pp. 2543–2560, 1999.

[6] Fernandes, "Directional, shift-insensitive, complex wavelet transforms with controllable redundancy", PhD Thesis, Rice University, 2002.

[7] Nick Kingsbury, "The dual_tree complex wavelet transform: A new efficient tool for image restoration and enhancement", *Proc. European Signal Processing Conference, EUSIPCO 98*, Rhodes, pp. 319-322, 1998.

[8] Wenbo W, Y.Jing, and K. Tingjun, "Study Of Remote Sensing Image Fusion And Its Application In Image Classification" *The Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Beijing*, Vol. XXXVI, Part B7, pp. 1141-1146, 2008.

[9] S G Mallat, "A theory for multiresolution signal decomposition: The wavelet representation", *IEEE Trans. PAMI*, 11(7), pp. 674-693, 1989.

[10] Chibani, Y., and A. Houacine, "The joint use of the HIS Transform and the redundant wavelet decomposition for fusing multispectral and panchromatic images", *International Journal of Remote Sensing*, 23(18), pp. 3821–3833, 2002.

[11] J.G.Liu, "Smoothing filter-based intensity modulation:A spectral preserve image fusion technique for improving spatial details," *Int. J. Remote Sens.*, vol. 21, no. 18, pp. 3461–3472, 2000.

[12] T.M.Tu, S.C.Su, H.C.Shyu, and P.S.Huang, "A new look at IHS-like image fusion methods," *Inf. Fusion*, vol.2, no. 3, pp. 177–186, 2001.

[13] J. N´unez, X. Otazu, O. Fors, A. Prades, V. c Pal´a, and R.Arbiol, "Multiresolution-Based Image Fusion with Additive Wavelet Decomposition," *IEEE Transactions on Geoscience and Remote Sensing*, vol.37, pp.1204 – 1211, 1999.

[14] T. Ranchin and L. Wald, "Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation," *Photogramm. Eng. Remote Sens.*, vol. 66, no. 1, pp. 49–61, 2000.

[15] B.Aiazzi, L.Alparone, S.Baronti, and A.Garzelli, "Context-driven fusion of high spatial and spectral resolution images based on over sampled multi-resolution analysis," *IEEE Trans. Geo sci. Remote Sens.*, vol. 40, no. 10, pp. 2300–2312, 2002.

[16] Y. Zhang, "A new merging method and its spectral and spatial effects," *Int. J. Remote Sens.*, vol. 20, no. 10, pp. 2003–2014, 1999.

[17] Ehlers M., S. Klonusa, P. Johan and P. Rosso, "Multi-sensor image fusion for pansharpening in remote sensing", *International Journal of Image and Data Fusion*, Vol. 1, No. 1, pp. 25–45, 2010.

[18] R. A. Schowengerdt, *Remote Sensing: Models and Methods for Image Processing* (2nd ed. Orlando, FL: Academic, 1997).

[19] Wang Z. and A.C. Bovik, "A universal image quality index," *IEEE Signal Process Lett.*, 9(3), pp. 81-84, 2002.

[20] K.Shivsubramani, P soman, Krishnamoorthy, "Implementation and Comparative Study of Image Fusion Algorithms", *International Journal of Computer Applications* (0975 – 8887) Volume 9, No.2, pp.3-6, 2010.

[21] Anjali Malviya and S.G.Bhirud,"Image Fusion of Digital Images", *Int. J. Recent Trends in Engineering*, Vol.2, No.3, pp. 2-4, 2009.

[22] V.P.S Naidu and J.R.Raol, "Pixel level Image fusion using wavelets and Principal component analysis", *Defence science Journal*, Vol.58, No.3, pp. 338-352, 2008.

[23] A. Goshtasby and S. G. Nikolov, "Image fusion: Advances in the state of the art", *Editorial- Science Direct, Special Issue on Image fusion*, 8(2), pp. 114-118, 2007.

[24] H.Wang, J. Peng, and W.Wu," Fusion algorithm for multisensor image based on discrete multi wavelet transform,"

- IEEE Proc. Visual Image Signal Process.*, 149(5), 2002.
- [25] K. Amolins, Y. Zhang, and P. Dare, "Wavelet based image fusion techniques - An introduction, review and comparison," *ISPRS Journal of Photogrammetry & Remote Sensing*, vol.62, pp. 249–263, 2007.
- [26] K. K. Gupta, R. Gupta, Wavelet Based Speckle Filtering of the SAR Images, (2006) *International Review on Computers and Software (IRECOS)*, 1 (3), pp. 224-232.
- [27] Nougrara, Z., Benyettou, A., Abdellaoui, A., Bachari, N.I., Lahmar, K., Comparative study between two proposed methods of an extracted road network and its nodes from satellite images of Algeria sites for contribution to the elaboration of a geographical information system GIS, (2012) *International Journal on Communications Antenna and Propagation (IRECAP)*, 2 (2), pp. 123-126.

Authors' information



G. Dheepa received her B.Sc computer science in 2001 , M.C.A in 2006 from Bharathidasan University, Tiruchirapalli and M.Phil in 2009 from Bharathiar University, Coimbatore, Tamilnadu, India. She is pursuing Ph.D degree in computer science at Bharathiar University. She has 2 years of experience as assistant professor in computer science at Excel Business

School(India). She has presented various research papers in national and international conferences and published a few in International journals. Her research interests include Image Processing, Computer Graphics, Data Mining and RDBMS.



Dr. S. Sukumaran graduated in 1985 with a degree in Science. He obtained his Master Degree in Science and M.Phil in Computer Science from the Bharathiar University. He received the Ph.D degree in Computer Science from the Bharathiar University. He has 25 years of teaching experience starting from Lecturer to Associate Professor. At present he is working as

Associate Professor of Computer Science in Erode Arts and Science College, Erode, Tamilnadu, India. He has guided for more than 40 M.Phil research Scholars in various fields and guided one Ph.D Scholar. Currently he is Guiding 5 M.Phil Scholars and 8 Ph.D Scholars. He published around 15 research papers in national and international journals and conferences. His current research interests include Image processing, Network Security and Data Mining. Dr.Sukumaran is a member of Board studies of various Autonomous Colleges and Universities.

Effect of Sensing Time Variation on Detection, Misdetction and False Alarm Probabilities in Cognitive Radio-Based Wireless Sensor Networks

J. A. Abolarinwa, N. M. Abdul Latiff, S. K. Syed-Yusof, N. Fisal, N. Salawu

Abstract – In this paper, we present an analytical derivation of the probability of detection, probability of misdetction and probability of false alarm in cognitive radio-based wireless sensor networks under a varying effect of sensing time. Sensing time is one of the most important parameters in cognitive radio-based networks. Particularly, in cognitive radio-based wireless sensor networks, the duration of channel sensing greatly impacts on probabilities of false alarm, misdetction and detection of primary user signal in a channel under investigation. In our analytical method, we have considered event that a given channel will be ON or OFF with different probabilities of ON and OFF states of the channel. From our simulation results, we found that, for different probability of ON channel state, there is significant change in probability of detection and misdetction as the sensing time increases. Simulation results also showed that for different probability of OFF channel state, probability of false alarm varies with sensing time. Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.

Keywords: Detection, Channel, Misdetction, Sensing-Time, False-Alarm

Nomenclature

W	Channel Sampling Frequency
T	Slot duration
λ_d	Detection Threshold
λ_f	False-Alarm Threshold
R	Data rate
t_1	Sensing slot
t_2	Transmission slot
C	Channel
H_1	Hypothesis PU available
H_2	Hypothesis PU absent
$i(t)$	Input signal
$i_f(t)$	Filtered signal
$[\]^2$	Squaring signal
$[\int]$	Integrator
$o(t)$	Output signal
$n(t)$	Noise signal
p	Channel ON duration
q	Channel OFF duration
P_d	Detection probability
P_f	False-alarm probability
P_m	Misdetection probability
Q	Q-function
SS_{eff}	Sensing efficiency
P_{ON}	Probability of Channel busy
P_{OFF}	Probability of channel free
L	Packet Size
$P_{success}$	Success probability
$P_{failure}$	Failure probability
σ_n^2	Noise Variance

σ_s^2	PU signal variance
P_1	Channel detection probability
P_2	Misdetection probability
P_3	Collision probability
P_4	False-alarm probability
P_5	Channel available probability

I. Introduction

There has been increase in research work in the field of wireless sensor networks (WSN) in recent time. This is due to many applications of sensor networks. Among others, disaster relief applications, battle-field surveillance, environment monitoring and biodiversity mapping, intelligent health care and medicine, facility management, intelligent building, telematics, and remote underwater surveillance are examples of WSN applications. WSN operate in the industrial scientific and medical (ISM) frequency band. 2.4GHz ISM band is unlicensed band that is open to other applications such as, IEEE 802.11 systems, IEEE 802.15.4 WPAN, wireless microphones, Bluetooth and microwave oven.

As a result of numerous applications and devices operating in this band, there is spectrum overcrowding problem. This leads to interference among dissimilar wireless applications using this band. In some locations, the allocation of the 2.4GHz frequency band has reached all-time height of 90% according to [1].

Due to spectrum scarcity and interference in the licensed and unlicensed band, a new efficient spectrum utilization paradigm has been proposed.

This paradigm is called cognitive radio (CR). Cognitive radio idea makes it possible for communication devices to adaptively and dynamically utilize the limited spectrum channels in an opportunistic manner.

Cognitive radio-based sensor networks (CRWSN) have the capability to sense its radio environment, intelligently adapt its communication parameters and reconfigure them accordingly. This unique feature of CR provides an avenue for solving the spectrum scarcity problem confronting wireless communication within the licensed and unlicensed frequency bands. As a result of the numerous potential advantages derivable from CR deployment, few authors have proposed a new sensor network called CRWSN.

This type of sensor network combines the CR functionalities with the traditional WSN to give a more robust sensor network. It is possible for this sensor networks to operate in both the licensed and unlicensed band in an opportunistic manner by sensing the spectrum for available channel. This is called opportunistic spectrum access (OSA). This is the core of cognitive radio technology. The new CRWSN networks prove to be more challenging than the traditional wireless sensor networks because CRWSN combines the features of WSN and CR in one system. Among other challenges are energy consumption and processing constraints.

The CRWSNs are also deployed in remote locations and they are low-power-battery-driven systems. In many cases, battery recharge is not possible. As a result of miniaturization of sensor nodes, computation complexity has to be kept simple.

In addition to these problems, simple antennas and radios are to be used in order to mitigate the problem of cost and affordability. Our focus in this paper is on the effect of spectrum sensing time on the detection, misdetection and false alarm probabilities during spectrum sensing in CRWSN.

II. Related Works

In [1], the authors presented recent developments and open research issues in spectrum management based on CR networks. Specifically, the work focused on the development of CR networks that does not require modification of existing networks. This work failed to address the issue of coexistence and interference between the primary user signal and that of secondary user CRWSN. CRWSN is a resource-constrained network. It is therefore very important to develop a suitable spectrum sensing technique that will optimize the limited channel resources of the network. In view of this need, the authors in [2] proposed energy detection-based spectrum sensing for cognitive radio network. The authors majorly focused on theoretical analysis based on the work done previously in [3].

The authors in [4] did a survey work on spectrum management in cognitive radio network, and they came up with four major open research questions about

spectrum sensing, spectrum decision, and spectrum sharing and spectrum mobility.

However, in furtherance to this work, authors in [5] proposed optimal spectrum sensing for cognitive radio network. On their part, authors in [6] proposed a scheme for channel access in cognitive radio sensor network with specific focus on energy efficiency.

In [7], the authors tried to answer the question of how frequently should spectrum sensing be carried out. The authors came up with trade-off between throughput and collision based on the effect of sensing time and PU activity pattern within the channel.

In a similar manner to the work done in [6], authors in [8] focused on energy efficiency as they considered spectrum sensing and access in cognitive radio networks.

In their work, they proposed optimal spectrum sensing and access mechanism using energy cost as their objective function. However, their work does not apply to CRWSN. The authors in [9] were concerned majorly with the order of sensing in a multi-channel cognitive radio network. They applied the principle of optimal stopping rule to decide when to stop sensing operation.

It is well known that sensing time is a major factor in spectrum sensing operation. Authors in [10] attempted using soft decision scheme to optimize the sensing time based on other parameters in a cooperative cognitive radio network.

The outcome of their work shows that optimal sensing time decreases with increase in signal to noise ratio. Authors in [11] looked at the challenges and solutions to spectrum sensing in cognitive radio network. This is a review work that described various spectrum sensing approaches available in literatures. Work in [12] directly focused on cognitive radio wireless sensor with specific thrust in the mode of channel selection for the cognitive radio sensor network secondary user.

Weighted combining cooperative energy detection spectrum sensing was used for dynamic spectrum access in [13]. Comparing their approach to equal gain and hard decision combining, they showed that weighted combining approach provided a better improvement. In addition, authors in [16] proposed optimal fuzzy fusion scheme to improve performance and reduce complexity of cognitive radio systems. A new reinforcement learning approach called DAC-PS is developed for solving the problem of an autonomous mobile robot navigating in an unknown and dynamic environment in [17].

From various works available in literature, we observed to the best of our knowledge that non have considered the effect of the probabilities of ON and OFF activities of primary user to determine how sensing time affects probabilities of detection, misdetection and false alarm in a CRWSN.

This is our focus in this work. The justification for this however is, for different probability of ON state of the channel, we will have different sensing time. This also determines what the detection, misdetection and false alarm probabilities will be. The rest of this paper is organized as follows; section III described the network

model. In section IV, we did the analysis of the various probabilities, while in section V we analyse the behaviour of the various channel states.

Section VI shows the simulation and result discussion. Finally, section VII gives the conclusion.

III. Network Model

Fig. 1 illustrates a simple cognitive radio network scenario. This network comprises of PU network and SU network operating within the same spectrum band in an overlay manner. The PU is the licensed user of the communication channel. The secondary user uses the licensed channel in an opportunistic manner when the primary user activity is not detected within the channel at a given period of time.

Cognitive radio is one critical enabling technology for future communications and networking that will bring about a more efficient and flexible way of utilizing the limited network resources. In cognitive radio based communication systems, the radio devices can adapt their operating parameters such as, transmission power, frequency, and modulation dynamically to the surrounding radio environment. This differentiates it from the traditional communication systems in which spectrum utilization is based on fixed spectrum allocation.

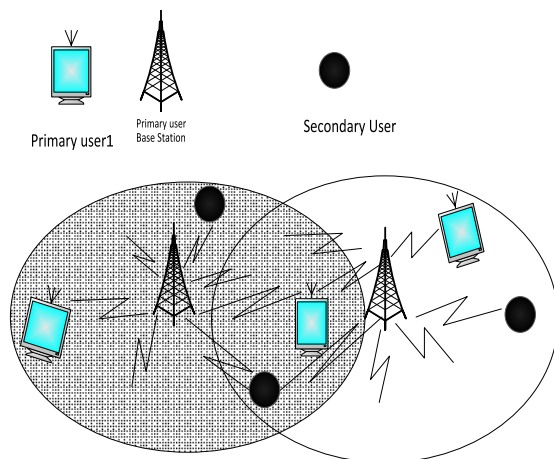


Fig. 1. Cognitive radio network scenario

III.1. Cognitive Radio-based Wireless Sensor Network Model

We considered a cluster-based cognitive radio-based wireless sensor network made up of cluster heads (CH), and member nodes (MN). The nodes are assumed to be static.

This is illustrated in Fig. 2. Cluster heads are full functioning device (FFD) with cognitive radio capabilities, while the member nodes are reduced functioning devices (RFD) which only send their captured data to the CH through sensed available channel. The CH coordinates communication within the network.

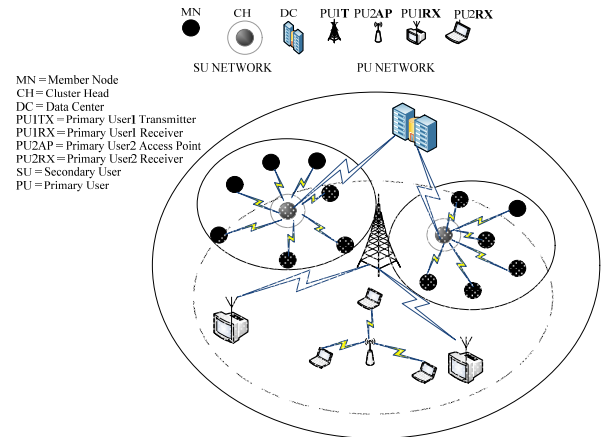


Fig. 2. Network model of cluster-based cognitive radio-based wireless sensor networks

III.2. Cognitive Radio-based Wireless Sensor Networks Channel Access Model

Timing pattern of the sensor network is shown in Fig. 3. This is a time-slotted scheme that is divided into sensing and transmission slots respectively. During the sensing slot, the CH scans the primary user (PU) spectrum for any available channel for communication by the sensor network.

Depending on the sensing out, the CH decides whether to transmit or to sense another channel. When a channel is found available and possesses good communication condition depending on the quality of service requirements of the application-specific CRWSN, the CRWSN switch into transmission time slot. During the transmission time slot, packet transmission between the member nodes and the CH takes place for a transmission time that is less than or equal to the entire transmission slot duration.

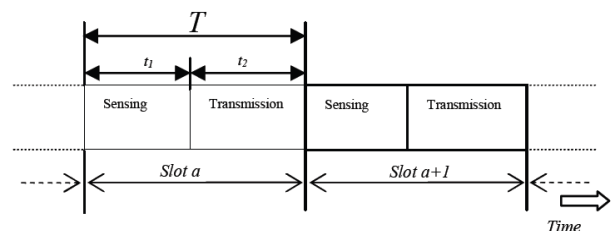


Fig. 3. Channel access timing framework

Time t_1 and t_2 represents the slot durations for sensing and transmission respectively. The higher the sensing duration, the more accurate the sensing outcome is expected to be. However, channel sensing is energy consuming. Also, in order to ensure accurate packet transmission, duration for transmission should be kept sufficiently large enough to accommodate the entire data packet size. Excessively high transmission time can be unbeneficial to the CRWSN in the event that PU begins transmission suddenly during the transmission slot.

Therefore, there is a trade-off between increasing sensing time and transmission time.

This also results in trade-off between sensing time and throughput. Accurate probability of detection and high spectrum sensing efficiency in terms of near zero misdetection probability are two important aims of spectrum sensing described in our work.

In Fig. 4, we show our channel access scheme designed for CRWSN. This is a time-slotted scheme based on the PU behavior in any particular channel of interest. Each frame is divided into two time slots which are, sensing and transmission slots as mentioned above.

Channel sensing and transmission is done based on time framework shown in Fig. 3. The local common control channel (LCCC) is introduced in the access scheme for the purpose of information control. For each cluster, there are C channels and a LCCC. A channel is considered available when there is no PU activity in the channel during sensing operation, and channel condition is suitable for data transmission by the secondary user (SU) CRWSN. Whatever be the outcome of the sensing operation by the CH of the CRWSN, the CRWSN proceeds to one of the following, transmit-receive state or switching state. Fig. 4 clearly depicts how this operation takes place.

As a result of radio frequency (RF) front-end hardware limitation and energy constrain of sensor networks, energy detection spectrum sensing is considered. With energy detection spectrum sensing technique, prior knowledge of the PU activity is not known by the SU CRWSN. Hence, the CH does not have to keep statistical record of the PU activity which is random in nature.

III.3. Energy Detection Spectrum Sensing Process

This is a spectrum sensing technique that makes use of the received signal at the CR receiver to determine the presence of the primary user within the channel. This process can be carried out both in the time and frequency domain.

With reference to [3] and [14], the energy detector consists of band-pass filter, which pre-filters the noise bandwidth, the squaring device and an integrator. Fig. 5 shows the block diagram of energy detection process in time domain.

Energy detection spectrum sensing technique is a test of two hypothesis based on the primary user activity in the channel. Hypothesis H_0 signifies the event that primary user signal is not available in the channel. Under this condition, the CRWSN only detects noise signal.

Hypothesis H_1 signifies the event that the primary user signal is available in the channel. Hence, the CRWSN detects both the primary user signal in addition to noise signal:

$$\begin{cases} H_0 : & o(t) = n(t) \\ H_1 : & o(t) = i(t) + n(t) \end{cases} \quad (1)$$

From (1), $o(t)$ is the received output of the integrator which decides if the channel is available or not available.

For H_0 , the received output is only noise sample $n(t)$.

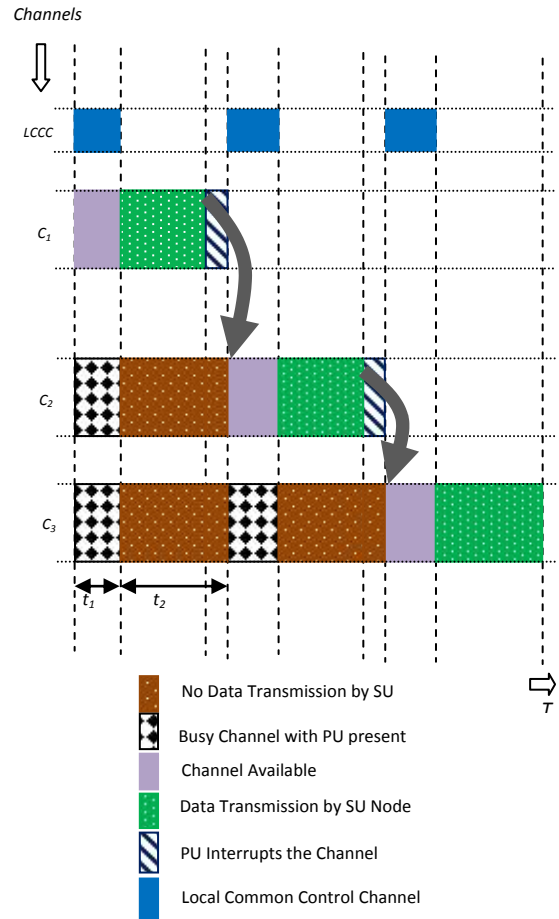


Fig. 4. Sensing, channel access and switching operations of cognitive radio-based wireless sensor networks

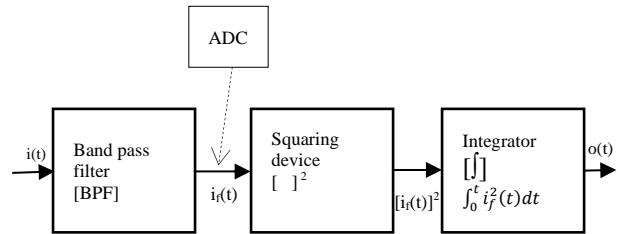


Fig. 5. Block diagram of energy detection process

We assume additive white Gaussian noise (AWGN) with zero average and σ_n^2 variance. For H_1 , received output is addition of sampled primary user signal $i(t)$ with σ_s^2 variance and AWGN noise. The test decision criterion of energy detection is given as:

$$o(t) = \int_0^{t_1} i_f^2(t) dt \quad (2)$$

III.4. Primary User Behavior Model

PU behavior is modeled as a two-state, time-homogenous discrete Markov process Wang et al (2012).

This is a preferred model because of the inter-dependence between the present state and the previous state.

The two-state Markov process is shown in Fig. 6. We considered a spectrum band consisting of C channels, each having different bandwidth BW . The PU can either be occupying the channel, which means, the channel is in ON state with PU signal present (that is, channel busy), or PU signal is absent, which means PU is in OFF state and the channel is available (that is, channel available) at any given time. When the PU occupies a channel, the channel is not available for the CR user to transmit.

Otherwise, the CR user transmit within the available channel assuming that other channel conditions such as fading and noise effects are favourable for packet transmission.

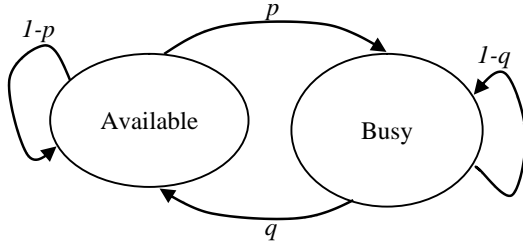


Fig. 6. Two state Markov model for primary user channel behavior

The PU occupancy of the channel follows independently identically distributed (i.i.d) exponential random variables p and q representing duration of ON and OFF respectively. Similarly, based on our previous work in [15], probabilities of ON and OFF are derived as follows:

$$P_{ON} = \frac{p}{p+q} \quad (3)$$

$$P_{OFF} = \frac{q}{p+q} \quad (4)$$

IV. Analysis of Detection, Misdetction and False-alarm Probabilities

From energy detection spectrum sensing described above, there are three important detection performance metrics which are, probability of detection P_d , probability of false alarm P_f , and probability of misdetection P_m . Detection probability is the probability that PU signal is correctly detected to be present in the channel within a sensing time. Probability of false alarm is the probability that reports the channel being occupied by the PU while the PU is actually absent at the instant of sensing the channel. Misdetection probability is also a probability that depicts the channel sensing outcome as free of PU signal while in actual sense; the PU is active in the channel during the sensing period. Generally, in terms of the hypothesis in (1), we define these probabilities mathematically define as:

$$\begin{cases} P_d = P_r\{\text{decision} = H_1|H_1\} \\ P_f = P_r\{\text{decision} = H_1|H_0\} \\ P_m = P_r\{\text{decision} = H_0|H_1\} \end{cases} \quad (5)$$

where:

$$P_m = 1 - P_d \quad (6)$$

For a given detection and false alarm decision thresholds λ_d and λ_f respectively, probabilities of detection and false alarm from (5) can be written as:

$$P_d = P_r\{o(t) > \lambda_d|H_1\} = Q\left(\frac{\lambda_d - 2t_l W(\sigma_s^2 + \sigma_n^2)}{\sqrt{4t_l W(\sigma_s^4 + \sigma_n^4)}}\right) \quad (7)$$

$$P_f = P_r\{o(t) > \lambda_f|H_0\} = Q\left(\frac{\lambda_f - 2t_l W\sigma_n^2}{\sqrt{4t_l W\sigma_n^4}}\right) \quad (8)$$

$o(t) > \lambda_d$ decides PU signal is ON

$o(t) < \lambda_f$ decides PU signal is OFF

$$Q = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left(-\frac{u^2}{2}\right) du$$

where λ_d and λ_f are the detection and false alarm decision threshold values respectively, σ_n^2 and σ_s^2 are noise and primary signal variances, W is the sampling frequency, and t_l is the sensing time of the SU. Q is the tail probability of the standard normal distribution or the probability that a normal (Gaussian) random variable will obtain a value larger than x standard deviations above the mean. It is formally called Q-function. From (3), (4), (7) and (8), we can formulate the P_d and P_f as a function of sensing time, ON and OFF event probabilities as:

$$P_d(t_l) = \left(\frac{p}{p+q}\right) Q\left(\frac{\lambda_d - 2t_l W(\sigma_s^2 + \sigma_n^2)}{\sqrt{4t_l W(\sigma_s^4 + \sigma_n^4)}}\right) \quad (9)$$

$$P_f(t_l) = \left(\frac{q}{p+q}\right) Q\left(\frac{\lambda_f - 2t_l W\sigma_n^2}{\sqrt{4t_l W\sigma_n^4}}\right) \quad (10)$$

Accurate probability of detection and high spectrum sensing efficiency in terms of near zero misdetection probability are two important aims of spectrum sensing.

Therefore, sensing efficiency is defined as the ratio of the transmission time to the total CR operation time.

From Fig. 3 we determine the spectrum sensing efficiency SS_{eff} as:

$$SS_{eff} = \frac{t_2}{t_1 + t_2} \quad (11)$$

where, t_1 and t_2 corresponds to sensing and transmission time slots respectively.

V. Channel State Probabilities Analysis

Five channel states were considered. These states are detection, misdetection, false alarm, collision and success states. We describe each state as shown in Fig. 7.

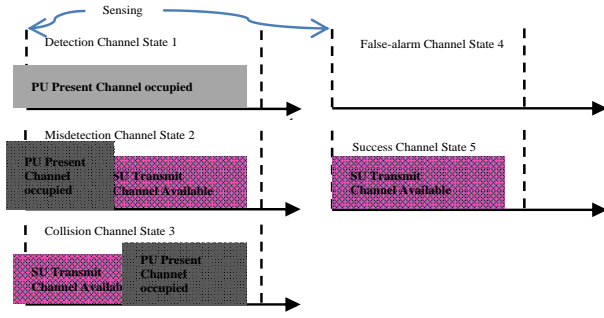


Fig. 7. Channel States for CRWSN under Dynamic Spectrum Access

There are two possible outcomes of sensing by the SU if the PU signal is present in the channel. It is either the PU is detected or is mis-detected. It therefore implies that, the probability that the PU is ON, which means, the probability that the channel is busy with PU activity is the summation of the probability of detection and of misdetection.

The PU must actually be occupying the channel in the event of these two states. As shown in Fig. 7, let P_1, P_2, P_3, P_4, P_5 represent the probability of each channel state accordingly.

Therefore, the probability of ON state can be written as:

$$P_{ON} = P_1 + P_2 \quad (12)$$

Substituting for P_{ON} in (9), then we solve for P_2 , where P_2 is the same as the probability of misdetection state, we have:

$$P_m(t_l) = P_{ON}Q\left(\frac{2t_lW(\sigma_s^2 + \sigma_n^2) - \lambda}{\sqrt{4t_lW(\sigma_s^4 + \sigma_n^4)}}\right) \quad (13)$$

A channel will only be in OFF state when the PU signal is not available in the channel as at the time of sensing by the SU. From Fig. 7, this only occurs in states 3 to 5. Therefore, probability of OFF channel state is the summation of the probabilities of states 3, 4 and 5, which are denoted as P_3, P_4 and P_5 :

$$P_{OFF} = P_3 + P_4 + P_5 \quad (14)$$

P_4 is the same as the probability of false alarm. This has been determined in (10) as $P_f(t)$. P_4 and P_5 are dependent on the packet size, L being transmitted by the SU. At every decision instant, the number of available channels is calculated as CP_{OFF} by the SU, where C is the number of channels. Based on [7], transmit-receive time t_2 as shown in Figure 3 is given as $t_2 = L/R$. Where L is the packet size and R is data rate for the SU transmission. If the probability of channel failure is given as $P_{failure} = 1 - e^{-\frac{L}{Rq}}$ and the probability of success is given as $P_{success} = e^{-\frac{L}{Rq}}$, the probability of P_4 and P_5 can be determined as follows:

$$P_4(L) = 1 - e^{-\frac{L}{Rq}}(P_{OFF} - P_3) \quad (15)$$

$$P_5(L) = e^{-\frac{L}{Rq}}(P_{OFF} - P_3) \quad (16)$$

VI. Simulations and Results

Channel access in cognitive radio-based wireless sensor networks depends on the following parameters to determine their efficiency. These performance metrics are, probability of detection, probability of misdetection, and false alarm probability as a function of sensing duration. We used MATLAB to carry out simulations and we obtained unique results of the relationships that exist between each of the probability and sensing time under different probability of channel busy and available.

Our simulation parameter set is shown in Table I.

TABLE I
SET OF SIMULATION PARAMETERS

Parameter	Description	Value
BW	Channel Bandwidth	1MHz
ISM-Band	Operating frequency Band	2.4GHz
λ	Energy Detection Threshold	5
d_{SU-PU}	Distance between PU and SU	50m
R	Data rate	40kbps
δ	Path loss component	2.5
P_{ON}	Probability of Channel busy	0.8, 0.5, 0.25, 0.1
P_{OFF}	Probability of channel free	0.2, 0.5, 0.75, 0.9
N_o	Noise Power	1.38×10^{-22}
P_{U_p}	PU Transmission Power	10dB
I_m	Maximum Interference Ratio	0.1
σ_n^2	Noise Variance	1
σ_s^2	PU signal variance	1

In Fig. 8, the variation of sensing time and the probability of detection is shown under a varying ON state probabilities. For different Pr_{ON} , there is a similar trend in the relationship between the sensing time and probability of detection. The result shows that, as the sensing time increases, there is an initial sharp rise in the probability of detection. This sharp rise is particularly noticeable at $Pr_{ON}=0.8, 0.5$ and 0.25 . However, beyond $0.4\mu s$ sensing time, there is a descent in the sharpness of increase in probability of detection. This is so because of finite sensing slot duration.

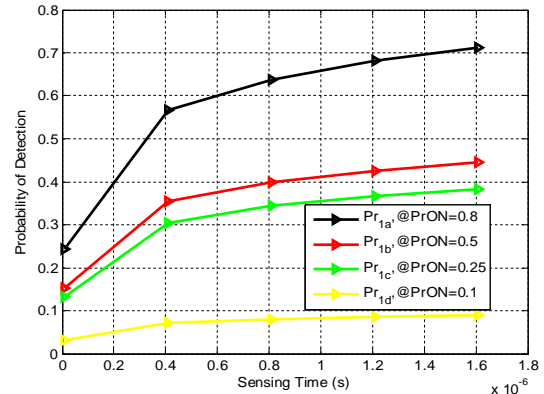


Fig. 8. Variation of probability of detection with sensing time under varying Pr_{ON}

Fig. 9 shows the variation of sensing time with the probability of misdetection state. Misdetection occurs when the SU cognitive radio sensor network could not detect the presence of a primary user in the channel even though the PU signal is still present within the channel.

The sensing time has significant effect on the probability of misdetection within the sensing slot of the SU frame. There is sharp decrease in the probability of misdetection with increase in sensing time.

This is logical because when the sensing time increases, it creates better chances for the detection of PU signal within the channel and definitely reduces the chances of misdetection. As expected, this sharp decrease is witness at higher values of Pr_{ON} .

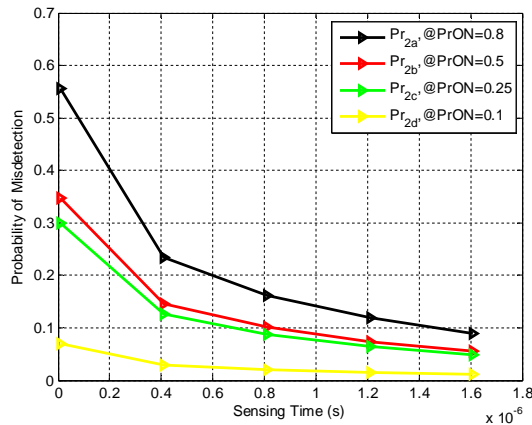


Fig. 9. Variation of probability of misdetection with sensing time under varying Pr_{ON}

From Fig. 10, it could be seen that the probability of false alarm, which is a situation in which the CR CH report to other MNs the outcome of its sensing that the PU signal is present in the channel, whereas the PU is not occupying the channel. This is understandable from the learning of the SUs. Based on the learning outcome of the past, the probability of false alarm tends to vary as the sensing time changes. This will ultimately reduce the reward rate of channel access by the SU. It could be seen from Fig. 10 also that as the Pr_{OFF} state increases, the increase in probability of false alarm witnessed is sharp.

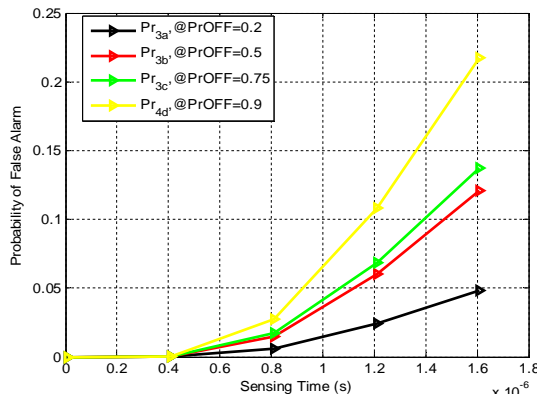


Fig. 10. Variation of probability of false alarm with sensing time under different Pr_{OFF}

VII. Conclusion

In this paper, we have presented an analytical derivation of the probability of detection, probability of misdetection and probability of false-alarm in CRWSN under varying probability of ON and OFF states of channel. We have shown how sensing time affects these parameters.

From simulation results, we found that; for different probability of ON channel state, there is significant variation in the probability of detection and misdetection respectively as the sensing time also changes. Simulation results also showed that for different probability of OFF channel state, probability of false alarm varies with sensing time.

In this work, we have limited our consideration to sensing duration. However, for future work, we will consider transmission time slot. As a form of limitation, this analytical approach is applied specifically to cognitive radio-based wireless sensor networks. This may not be applicable to other type of networks. Hence, our analytical approach is not generic.

Acknowledgements

This work was supported by Ministry of Higher Education (MOHE) Malaysia, Research Management Centre (RMC) of Universiti Teknologi Malaysia (UTM) and Federal University of Technology Minna, Nigeria.

References

- [1] O. Akan, O. Karli, and O. Ergul, "Cognitive radio sensor networks," *IEEE Network*, vol. 23, pp. 34-40, 2009.
- [2] Z. Xuping and P. Jianguo, "Energy-detection based spectrum sensing for cognitive radio," *IET Conference on Wireless, Mobile and Sensor Networks*, pp. 944-947, 2007.
- [3] H. Urkowitz, "Energy detection of unknown deterministic signals," *Proceedings of the IEEE*, vol. 55, pp. 523-531, 1967.
- [4] I. F. Akyildiz, L. Won-Yeol, M. C. Vuran, and S. Mohanty, "A survey on spectrum management in cognitive radio networks," *IEEE Communications Magazine*, vol. 46, pp. 40-48, 2008.
- [5] L. Won-Yeol and I. F. Akyildiz, "Optimal spectrum sensing framework for cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7, pp. 3845-3857, 2008.
- [6] Z. Huazi, Z. Zhaoyang, C. Xiaoming, and Y. Rui, "Energy Efficient Joint Source and Channel Sensing in Cognitive Radio Sensor Networks," *IEEE International Conference on Communications*, pp. 1-6, 2011.
- [7] P. Yiyang, H. Anh Tuan, and L. Ying-Chang, "Sensing-Throughput Tradeoff in Cognitive Radio Networks: How Frequently Should Spectrum Sensing be Carried Out?," *IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1-5, 2007.
- [8] S. Wang, W. Yue, J. P. Coon, and A. Doufexi, "Energy-Efficient Spectrum Sensing and Access for Cognitive Radio Networks," *IEEE Transactions on Vehicular Technology*, vol. 61, pp. 906-912, 2012.
- [9] C. Ho Ting and Z. Weihua, "Simple Channel Sensing Order in Cognitive Radio Networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, pp. 676-688, 2011.
- [10] S. Dafei, S. Tiecheng, W. Ming, H. Jing, G. Jie, and G. Bin, "Optimal sensing time of soft decision cooperative spectrum sensing in cognitive radio networks," *IEEE Wireless Communications & Networking Conference*, pp. 4124-4128, 2013.

- [11] Y. Zeng, Y.-C. Liang, A. Hoang, and R. Zhang, "A Review on Spectrum Sensing for Cognitive Radio: Challenges and Solutions," *EURASIP Journal on Advances in Signal Processing*, p 381465, 2010.
- [12] L. H. A. Correia, E. E. Oliveira, D. F. Macedo, P. M. Moura, A. A. F. Loureiro, and J. S. Silva, "A framework for cognitive radio wireless sensor networks," *IEEE Symposium on Computers and Communications*, pp. 611-616, 2012.
- [13] F. E. Visser, G. J. M. Janssen, and P. Paweczak, "Multinode Spectrum Sensing Based on Energy Detection for Dynamic Spectrum Access," *IEEE Vehicular Technology Conference*, pp. 1394-1398, 2008.
- [14] S. V. Omkar, M. K. Vijaya, "Analysis of Energy Detection based spectrum sensing over wireless fading channels in cognitive radio networks," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3 (3), 2013.
- [15] J. A. Abolarinwa, N. M. Abdul Latiff, and S. K. Syed Yusof, "Energy Constrained Packet Size Optimization For Cluster-based Cognitive Radio-based Wireless Sensor Networks", *Australian Journal of Basic and Applied Sciences*, vol. 7, pp. 138-150, 2013.
- [16] Aldar, D.S., Distributed fuzzy optimal spectrum sensing in cognitive radio, (2012) *International Review on Computers and Software (IRECOS)*, 7 (6), pp. 2788-2793.
- [17] Zemzem, W., Tagina, M., A new approach for reinforcement learning in non stationary environment navigation tasks, (2012) *International Review on Computers and Software (IRECOS)*, 7 (5), pp. 2078-2087.

Authors' information

Faculty of Electrical Engineering, Universiti Teknologi Malaysia, UTM-MIMOS Center of Excellence, Johor, Malaysia.



J. A. Abolarinwa obtained Bachelor's degree in Electrical and Computer Engineering from Federal University of Technology Minna, Nigeria. He obtained Master's degree (M.Eng) in Electrical and Electronic Engineering from University of Port-Harcourt, Nigeria. He is currently a Ph.D research student in the Faculty of Electrical Engineering, UTM. He is a member of the Telematic Research Group (TRG). His research interests are in, Cognitive Radio Networks, Software Defined Radio, and Wireless Sensor Networks. He is a registered member of professional organizations such as IEEE, IET, and IEICE.



N. M. Abdul Latiff received her Bachelor of Engineering (B.Eng) degree in Electrical-Telecommunications in the 2001 from UTM, Malaysia. She obtained Master of Science (M.Sc) degree in Communications and Signal Processing, and Ph.D in Wireless Telecommunication Engineering from Newcastle University, UK in 2003 and 2008 respectively. Currently, she is a senior lecturer at the Faculty of Electrical Engineering, UTM and she is a member of the Telematic Research Group (TRG). Her research interest includes, Cognitive Radio, Wireless Sensor Networks, Mobile Ad Hoc Networks, Network Optimization, Bio-inspired Optimization Algorithm, Evolutionary Algorithm and Clustering Algorithm. She is a registered member of IEEE and IET.



S. K. Syed Yusof received BSc (cum laude) in Electrical Engineering from George Washington University USA in 1988 and obtained her MEE and PhD in 1994 and 2006 respectively from UTM. She is currently an Associate Professor with Faculty of Electrical Engineering, UTM, and collaborating with UTM-MIMOS Centre of Excellence. Her research interest includes wireless communication, Software define Radio, Network Coding and Cognitive radio. She is a member of Eta Kappa Nu (HKN), and Phi Beta Kappa society.



N. Fisal received her B.Sc. in Electronic Communication from the University of Salford, Manchester, U.K. in 1984. M.Sc. degree in Telecommunication Technology, and PhD degree in Data Communication from the University of Aston, Birmingham, U.K. in 1986 and 1993, respectively. Currently, she holds university professor position at the Faculty of Electrical Engineering, UTM and she is the director of UTM-MIMOS Centre of Excellence. Her research interests are in the areas of multimedia networking, wireless sensor network and cognitive radio.



N. Salawu is currently pursuing his Ph.D degree in the department of Communication Engineering, Faculty of Electrical Engineering, Universiti Teknologi Malaysia. He is a research student in the Telematic Research Group (TRG). He received B. Eng. degree in Electrical and Computer Engineering, and M. Eng. degree in Communication Engineering from Federal University of Technology Minna, Nigeria in 2002 and 2010 respectively. His research interests include, radio resource management in cellular networks including 4G networks.

An Efficient Hybrid Segmentation Algorithm for Computer Tomography Image Segmentation

V. V. Gomathi¹, S. Karthikeyan²

Abstract – Medical Image segmentation plays a major role in medical image processing. During last decades, developing robust and efficient algorithms for medical image segmentation has been a demanding area of growing research interest. Extensive research has been done in creating many different approaches and algorithms for image segmentation, but it is still difficult to assess whether one algorithm produces more accurate segmentations than the other. The proposed method utilizes clustering with distance based segmentation approach for Computer tomography image segmentation. This paper provides new hybrid segmentation method based on K-Means, Medoid shift and Signature Quadratic Form Distance algorithm for computer tomography images. We validate the Hybrid segmentation approach with the parameters in terms of sensitivity, specificity, accuracy and number of fragments. The Real time dataset is used to evaluate the performance of the proposed method. The results obtained from the experimentation show that the proposed approach attains reliable segmentation accuracy and also clear that it is more efficient, robust and more appropriate for organ classification. **Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords: Computer Tomography, Hybrid Segmentation, K-Means, Medical Image Segmentation, Medoid Shift, Signature Quadratic Form Distance

Nomenclature

x_i	i-th element of the input image
v_j	j-th element of the cluster array
c	Number of clusters
$\left (x_i - v_j)^2 \right $	Square of absolute difference between i-th element of the input image and j-th element of the cluster value to find the distance between input image vector and cluster vector
$(1/C_i) * \left(\sum_{j=1}^{C_i} (x_i) \right)$	Identifying the new cluster by taking mean using i-th element of the input image and total number of clusters
C_i	Total number of Clusters
A	Similarity Matrix
P	Intensity Vector Pixel Values
Q	Input Image Pixel
T	Transpose Matrix

I. Introduction

Medical Image Segmentation is the process of automatic or semi-automatic detection of boundaries within a 2D or 3D image.

A major difficulty of medical image segmentation is the high variability in medical images. Medical images and imaging techniques play vital role in assisting health

care providers to access patients for diagnosis and treatment. Studying medical images depends mainly on the visual interpretation of the radiologists. However, this consumes time and usually subjective, depending on the experience of the radiologist and also suffer from intra-observer and inter-observer variability. Consequently the use of computer-aided systems becomes very necessary to overcome these limitations [1].

Structures are identified and segmented using either a manual, semi-automatic or fully automated methods. In the past few decades, many effective algorithms have been proposed to perform the computer-aided segmentation. The medical image Segmentation algorithm is suitable depends upon the different modalities images.

Various medical imaging techniques such as computed tomography (CT), magnetic resonance imaging (MRI), Ultrasound(US), Positron Emission Tomography (PET), etc provide different perspectives on the human body. Computer tomography is very important imaging modalities to provide radiotherapy for tumor patient. Manual segmentation is time consuming task and be prone to errors, especially due to fatigue.

Manual segmentation also gives inter and intra expert variability results. In this scenario reliable algorithms are essential for the delineation of anatomical structures and other regions of interest (ROI) to assist and automate the radiological tasks. Techniques for performing segmentations vary widely depending on the specific application, imaging modality, and other factors.

There is no universal algorithm for segmentation of every medical image. Each imaging system has its own specific limitations [2].

Clustering with distance based algorithm is suitable for Computer tomography images. One main drawback of the HMSK (Hybrid medoid shift with K-Means) and Signature Quadratic form distance method (SQFD) is over fragments. These pitfalls can be suppressed by using proposed hybrid segmentation method which uses HMSK and SQFD algorithm. In this paper, we proposed Hybrid Segmentation algorithm compared with Hybrid medoid shift with K-Means algorithm (HMSK) and Signature Quadratic Form distance (SQFD) algorithm.

The Experimental result proves that the Hybrid Segmentation algorithm gives most promising results with less number of fragments.

The rest of this paper is organized as follows. Section II reviews other published segmentation solution. Section III describes Materials used in this research work and explains and illustrates the proposed Hybrid Segmentation algorithm for Computer tomography Images. Section IV presents its Experimental results and compares with other existing methods and also presents the discussion. Finally, some conclusions are drawn in Section V.

II. Related Works

Ladak HM et al., used model-based initialization and the efficient discrete dynamic contour for semiautomatic segmentation of the prostate from 2D ultrasound images [3]. Yiqiang Zhan et al., presents a novel deformable model for automatic segmentation of prostates from three-dimensional ultrasound images, by statistical matching of both shape and texture [4]. Djamal Boukerroui et al., proposed a robust adaptive region segmentation algorithm within a Bayesian framework [5]. Ashish Thakur et al., presents the region based segmentation method for ultrasound images using local statistics. In this segmentation approach the homogeneous regions depends on the image granularity features, where the interested structures with dimensions comparable to the speckle size are to be extracted [6].

Elnomery Zanaty et al., presented reliable algorithms for fuzzy k-means and C-means that could improve MRI segmentation. They have estimated accurate clusters automatically even without knowing prior knowledge of the true tissue types and the number of cluster of given images [7]. Mohamed N. Ahmed et al., proposed a novel algorithm for fuzzy segmentation of magnetic resonance imaging (MRI) data and estimation of intensity inhomogeneities using fuzzy logic [8]. Jianzhong Wangm et al., presented a modified fuzzy c-means (FCM) algorithm for MRI brain image segmentation.

The proposed method incorporates both the local spatial context and the non-local information into the standard FCM cluster algorithm using a novel dissimilarity index in place of the usual distance metric[9].

Paresh Chandra Barman et al., proposed a medical diagnosis system by using level set method for segmenting the MRI image which investigates a new variational level set algorithm without re- initialization to segment the MRI image. They have used the speed function and the signed distance function of the image in segmentation algorithm. Their system consists of thresholding technique, curve evolution technique and an eroding technique [10]. Shan Shen et al., described a robust segmentation technique based on an extension to the traditional fuzzy c-means (FCM) clustering algorithm which is dependent on the relative location and features of neighboring pixels, is shown to improve the segmentation performance dramatically [11]. Iraky khalifa et al., presented a novel manipulation or utilization of Fuzzy C- Means (FCM) Clustering by using wavelet Decomposition for feature extraction and feature vector treat as input to FCM [12].

Chung-Yi Huang et al., proposed a region growing algorithm for constructing the triangular models of anatomic structures from two-dimensional slices of CT images. A modified marching cubes algorithm, a 3D reconstruction algorithm, is then employed to establish the triangular model and a data reduction algorithm that combines a pair of voxels along each coordinate direction is developed, in which piecewise linear interpolation is implemented to maintain the accuracy of the reduced model. [13]. Moreno A. et al., proposed an automatic and robust method, based on anatomical knowledge about the heart, in particular its position with respect to the lungs. This knowledge is represented in a fuzzy formalism and it is used both to define a region of interest and to drive the evolution of a deformable model in order to segment the heart inside this region [14].

Pardo X.M et al., presented a deformable contour that combines region and edge/gradient information in order to solve problems in the classical formulation that affect the segmentation of images obtained from CT bone scans [15]. Zikuan Chen et al., presented and developed an automatic method for 3D reconstruction of vascular trees using computed-tomography angiographic (CTA) images. All the involved algorithms are presented with the emphasis given to the skeleton pruning and tree construction algorithms. The skeletons obtained using a 3D thinning algorithm may contain cycles, spurs, isolated sticks, and non-unit-width parts, which hinder tree construction. As a solution to this problem, a skeleton pruning and tree construction algorithm is proposed [16].

III. Materials and Methods

III.1 Data Set Description

Different type of Tumor patient dataset was collected by a SIEMENS SOMATOM EMOTION SPIRAL CT scanner located at Multi Speciality Hospital, Coimbatore.

Besides a normal scan performed at a routine clinical dosage (130 mA), an additional scan from the same patient was acquired at a much lower tube current, i.e. 20 mA.

The 3D image data consisted of DICOM (Digital Imaging and Communications in Medicine) consecutive slices, each slice being of size 512 by 512 and having 16-bit grey level resolution. Each of the organs of interest in this research was manually contoured by the expert for the comparison of auto segmented output with manual contoured image.

III.2 Methodology

This paper proposes a new Hybrid segmentation algorithm based on Medoid shift, K-Means and Signature quadratic form distance segmentation method for computer tomography images. Medoid shift algorithm is also a nonparametric clustering approach. It is a mode seeking method that computes shifts towards areas of greater data density using local weighted medoids. The use of medoids to discover structure in data is natural since, locally, the medoid can be considered a good representative of its neighborhood. Unlike means, medoids do not need an explicit feature space and require only a valid distance measure [17]. The medoid shift algorithms also automatically calculate the number of clusters during execution like mean shift [18].

The most popular method for image segmentation is k-means clustering [19] [20]. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters fixed a priori. The clustering results of the K-means algorithm greatly depend on its initialization. The number of clusters must be known in advance.

The distance measure plays an important role in acquiring exact clusters. It is used to discover the similarity and dissimilarity between the pair of objects in the clustering techniques. Clustering techniques are based on measuring similarity and dissimilarity between data objects by calculating the distance between each pair. The choice of distance measure between clusters has a large effect on the shape of the resulting clusters [21].

Signature Quadratic form distance is a generalization of the Quadratic form distance. It (SQFD) [22] is an adaptive distance-based similarity measure. Signature Quadratic Form Distance measure which allows efficient similarity computations based on flexible feature representations. This approach bridges the gap between the well-known concept of Quadratic Form Distances and feature signatures. The Signature Quadratic Form Distance (SQFD) is a recently introduced distance measure for content-based similarity. It makes use of feature signatures, a flexible way to summarize the features of a multimedia object. The SQFD is a way to measure the similarity between two objects [21].

An efficient Hybrid Segmentation algorithm is proposed to avoid the pitfalls present in the HMSK algorithm and SQFD segmentation algorithm. One of common drawback was over fragmentation produced by the above said two algorithms. Over fragmentation produces many connected components.

Classifying many connected components is a tedious process. This will also cause wrong segmentation result and also leads to wrong decision making by the radiologist.

The above said algorithm is also not well suited to separate the joined organs. Some organs are joined together. For example heart and liver is joined together and also heart and spleen is joined together. In these circumstances, the proposed Hybrid segmentation algorithm is well suited for segmenting the joined organs efficiently. The Proposed Hybrid Segmentation is as follows.

Hybrid Segmentation Algorithm

Step 1: Consider the Single Dicom image or slices of Dicom images

Step 2: Apply the ECFT (Enhanced Curvelet Filter Technique) algorithm to get a noiseless image

Step 3: Obtain the Histogram of the input image

Step 4: Initialize the control parameter

Step 5: Find the gray level cluster values based on an initialized control parameter

Step 6: Find no of pixel values present between each range of all gray level cluster values.

Step 7: Cluster the pixels which lies between the ranges to the respective gray level cluster value

Step 8: Each cluster are considered as data points

Step 9: Find the distance between each cluster to all the data points

Step 10: Make the data point allocation by using
$$\left(\sum_{i=1}^c * \sum_{j=1}^c \left((x_i - v_j) \right)^2 \right)$$

Step 11: Find the new cluster center by using
$$(1/C_i) * \left(\sum_{j=1}^{C_i} (x_i) \right)$$

Step 12: Obtain the Clustered Image

Step 13: Initialize the cluster step value

Step 14: Generate the cluster centers based on the cluster step value.

Step 15: Calculate the similarity matrix A using cluster centers P and input image pixel values Q.

Step 16: Compute the Signature Quadratic Form Distance (SQFD) value using the following formula

$$SQFDA(Q, P) = \sqrt{((Q/P) * A * (Q/P))^T}$$

where

A - Similarity Matrix

P - Intensity Vector Pixel Values

Q - Input Image Pixel

T - Transpose Matrix

Step 17: Find the minimum distance value

Step 18: Find the cluster center value based on minimum distance value and assign that cluster center value to the respective pixel position in the image

Step 19: Repeat the step 15, 16, 17 and 18 until convergence is attained (i.e. no pixels change clusters).

In the proposed hybrid segmentation algorithm consists of medoidshift with K-means and signature quadratic form distance method. Initially an ECFT (Enhanced Curvelet Filtering Technique) has been applied for removal of noise in the CT images.

These noiseless images are the input images. Histogram is found for the input images. Based upon the histogram of the input image, the control parameter is initialized. The random cluster values have taken based on the control parameter. The Closest cluster value is found based on the occurrences of cluster values.

Here each cluster is considered as data points. The distance between data points and cluster points (Closest cluster) has been calculated and found the new cluster.

Finally the similar cluster image has been obtained. The initial cluster step value has chosen either by manually or randomly. Then find the cluster centers based on the initialized cluster step value. The number of clusters in the image is equal to the number of cluster centers. Then the distance measure has been calculated between every pixel and the cluster centers. Signature Quadratic form distance is used to find the distance. For this process the two vectors P and Q is formed. By using P and Q the similarity matrix is generated.

Then the SQFD similarity between the P and Q is identified. The position with minimum SQFD value is identified. The cluster center consists of the cluster value.

The minimum value in the cluster center position is replaced with the original pixel value in the same position. Repeat the same process till the convergence attained (ie there is no change in the pixel value of an image). Finally we obtained the segmented image.

IV. Experimental Results and Discussion

Experimentation was carried out on 100 numbers of different tumor patients contains 100 to 1000 slices of Computer Tomography images using Segmentation algorithms. The image format is DICOM (Digital Imaging Communications in Medicine). The algorithm has been implemented in Matlab environment. Manual Segmentation done by the medical expert. Experimental results of the images are illustrated here. Fig. 1(a) depicts input CT image. Fig. 1(b) describes segmentation result generated with HMSK. Fig. 1(c) depicts SQFD algorithm on the CT images. Fig. 1(d) shows Hybrid Segmentation result. Fig. 2(a) shows segmentation result generated with HMSK for Liver region. Fig. 2(b) depicts results generated with SQFD algorithm for Liver Region. Fig. 2(c) shows segmentation result generated with proposed Hybrid segmentation. Fig. 2(d) depicts the manual segmentation results contoured by the medical experts for liver organ.

IV.1. Performance Analysis of HMSK, SQFD, Hybrid Segmentation Method

Selecting the suitable segmentation evaluation measure is a complex task. A variety of performance

measures to assess the medical image segmentation methods are available in present scenario. Generally sensitivity, specificity and accuracy are used to evaluate the segmentation methods in a good manner.

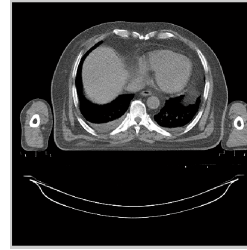


Fig. 1(a). Input Image

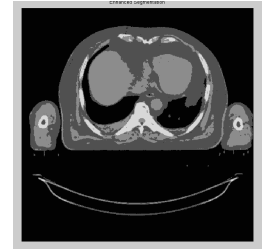


Fig. 1(b). HMSK Segmentation



Fig. 1(c). SQFD Segmentation

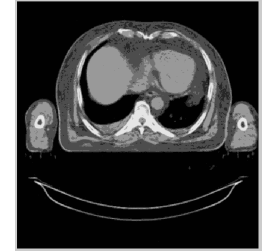


Fig. 1(d). Hybrid Segmentation

Liver Segmentation Output



Fig. 2(a). HMSK Segmentation

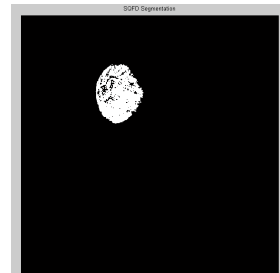


Fig. 2(b). SQFD Segmentation output

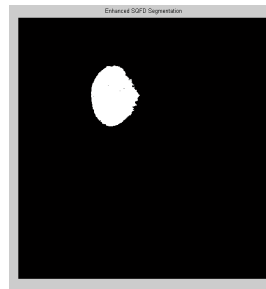


Fig. 2(c). Hybrid Segmentation output



Fig. 2(d). Manual Segmentation done by the experts

They are defined as:

$$\text{Sensitivity} = \frac{TP}{TP + TN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP (True Positive) is the number of pixels of the foreground that are correctly classified, TN (True Negative) is the number of pixels of the background that are correctly classified, FP (False Positive) is the number of pixels of the background that are classified as foreground and FN (False Negative) is the number of pixels of the foreground that are Classified as background.

Accuracy refers to the degree to which the segmentation results agree with the true segmentation i.e. Correct segmented pixels in the object. Fragments indicate that the number of connected components in the required region to identify as organ.

In this paper we also consider the fragments parameter. If more number of fragments exists in the image, the segmentation task is also complicated.

TABLE I
PERFORMANCE ANALYSIS OF HMSK, SQFD,
HYBRID SEGMENTATION ALGORITHM

Quantitative Parameters	CT Image Segmentation Algorithm		
	HMSK	SQFD	Hybrid Segmentation (HMSK with SQFD)
Sensitivity	93.19	98.12	98.12
Specificity	99.26	99.99	99.95
Accuracy	99.05	99.22	99.65
Number of fragments	350	173	133

We have proposed and successfully implemented a new integrated method for segmenting real time Computer tomography images. This paper mainly concentrates to propose a new algorithm with comparison of HMSK and SQFD algorithm.

In our previous research, we have compared many segmentation algorithms. The segmentation results of HMSK and SQFD are considered the best algorithm that gives better segmentation results for real CT images. The HMSK algorithm is based on clustering based segmentation algorithm and SQFD algorithm is a distance based algorithm.

In cluster based medical image segmentation algorithms, more number of unwanted fragments present and also fragments are not consistent when executed for a certain number of times i.e. when the same image executed for different number of times, the result were not holding the same number of fragments, position of fragment and size of fragment and also were dynamic.

For diminishing these drawbacks, the distance based segmentation algorithm has been proposed. In our previous distance based research, we have compared five distance measures namely Euclidean distance, Manhattan Distance, Minkowski distance, Chebyshev distance and Signature Quadratic form Distance(SQFD) measures[21].

The SQFD algorithm finds the similarity between all cluster elements and every pixel values such that the feature space have the highest possible similarity values of cluster vector. One of the drawback exist in this SQFD is Still number of fragments exist is more. The HMSK and SQFD algorithm is also not good fit for exact computer tomography image segmentation. Hence we proposed a Hybrid Segmentation method.

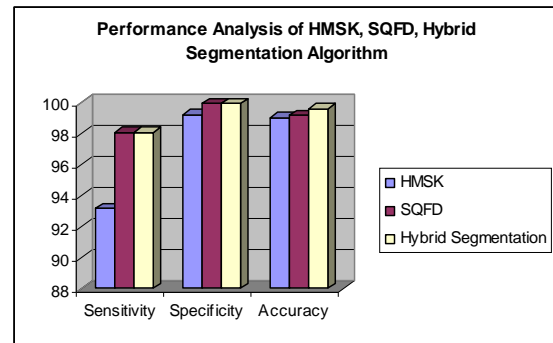


Fig. 3(a). Performance analysis of HMSK, SQFD, Hybrid segmentation algorithms in terms of sensitivity, specificity and accuracy

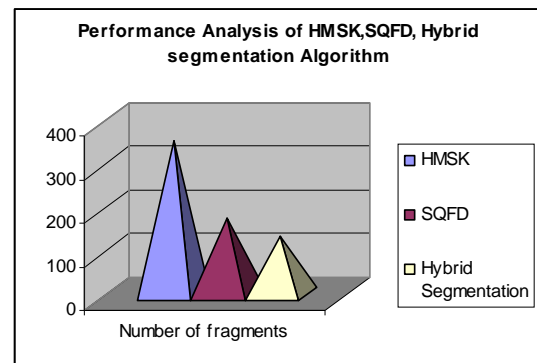


Fig. 3(b). Performance analysis of HMSK, SQFD, Hybrid segmentation algorithms in terms of number of fragments

There are more number of quantitative metrics are available to evaluate the segmentation accuracy. Here four important parameters are used to determine the precision of the HMSK, SQFD and Hybrid segmentation algorithms such as Sensitivity, specificity, accuracy and number of fragments. The important parameters mentioned above i.e. Sensitivity, specificity, accuracy and number of fragments are specified in Table I, for the HMSK, SQFD, Hybrid segmentation algorithms. In this paper, the liver organ has taken for demonstrating the segmentation performance. Segmentation of liver is a tedious process.

Liver is joined with some other organs such as heart, lung, stomach, spleen etc. In Figs. 2(a), (b) and (c) describes the liver segmentation done by HMSK, SQFD and Hybrid Segmentation algorithm.

Hybrid Segmentation algorithm having less number of fragments compared to HMSK and SQFD algorithm based on the visible part and analysis part. The segmentation result is compared with manual contoured image already done by the medical expert. Manual segmentation result is represented in Fig. 2(d).

The proposed approach is precise, robust and provides good quality results. From the experimental results obtained it can be concluded that our proposed Hybrid segmentation algorithm performs well in segmenting the real CT images with less number of fragments (connected components). The liver and heart organs are well separated and heart and spleen are also well

separated. The proposed hybrid segmentation method gives the highest accuracy than other methods.

V. Conclusion

In this paper, a hybrid segmentation algorithm is proposed to get the precise fragments from Computer tomography images. We found that the clustering with distance based algorithm is well suitable for exact computer tomography image segmentation.

The method integrates medoidshift with K-means algorithm and Signature quadratic form distance algorithm. Such integration reduces the drawbacks of both the methods. The benefit of the methodology is that it produces high quality segmentations of Computer tomography images and also separates the joined organ efficiently.

The computational results show that the proposed methodology achieves high quality precise segmentation results with less number of fragments.

Hence the complexity is reduced for exact organ classification.

References

- [1] Dinesh D. Patil, Sonal G. Deore, Medical Image Segmentation: A Review, *International Journal of Computer Science and Mobile Computing*, Vol. 2, n.1, pp.22 – 27, 2013.
- [2] Neeraj Sharma and Lalit M. Aggarwal, Automated medical image segmentation technique. *Journal of Medical Physics*, Vol.35, n.1, pp.3–14, 2010.
- [3] Ladak HM, Mao F, Wang Y, Downey DB, Steinman DA, Fenster A., Prostate boundary segmentation from 2D ultrasound images, *Journal of Medical Physics*, Vol.27,n.8,pp.1777-1788,2000.
- [4] Yiqiang Zhan and Dinggang Shen, Deformable Segmentation of 3-D Ultrasound Prostate Images Using Statistical Texture Matching Method, *IEEE Transactions on Medical Imaging*, Vol. 25, n. 3, pp.256-272, 2006.
- [5] Djamal Boukerroui , Atilla Baskurt c, J. Alison Noble , Olivier Basset , Segmentation of ultrasound images—multiresolution 2D and 3D algorithm based on global and local statistics, *Pattern Recognition Letters*,Vol.24 , pp.779–790, 2003.
- [6] Ashish Thakur Radhey Shyam Anand, A Local Statistics Based Region Growing Segmentation Method for Ultrasound Medical Images, *International Journal of Medical, Health, Pharmaceutical and Biomedical Engineering Vol.1*, n.10, pp.570-575, 2007.
- [7] Elnomery Zanaty and Sultan Aljahdali, Improving Fuzzy Algorithms for Automatic Magnetic Resonance Image Segmentation, *The International Arab Journal of Information Technology*, Vol. 7, n.3, pp.271-279, 2010.
- [8] Mohamed N. Ahmed, Sameh M. Yamany, Nevin Mohamed, Aly A. Farag, and Thomas Moriarty, A Modified Fuzzy C-Means Algorithm for Bias Field Estimation and Segmentation of MRI Data, *IEEE Transactions on Medical Imaging*, Vol. 21, n. 3, pp.193-199, 2002.
- [9] Jianzhong Wangm, Jun Kong, Yinghua Lu,,Miao Qi, Baoxue Zhanga, A modified FCM algorithm for MRI brain image segmentation using both local and non-local spatial constraints, *Computerized Medical Imaging and Graphics*, Vol. 32, n.8, pp.685–698, 2008.
- [10] Pares Chandra Barman, Sipon Miah, Bikash Chandra Singh and Mst. Titasa Khatun, MRI mage segmentation using level set method and implement an medical Diagnosis system, *Computer Science & Engineering: An International Journal (CSEIJ)*, Vol.1, n.5, pp.1-10, 2011.
- [11] Shan Shen, William Sandham, Member, IEEE, Malcolm Granat, and Annette Sterr , MRI Fuzzy Segmentation of Brain Tissue Using Neighborhood Attraction With Neural-Network Optimization, *IEEE Transactions On Information Technology In Biomedicine*, Vol. 9, n.3, pp.459-467, 2005.
- [12] Iraky khalifa , Aliaa Youssif , Howida Youssry, MRI Brain Image Segmentation based on Wavelet and FCM Algorithm, *International Journal of Computer Applications*, Vol.47, n.16, pp.32-39, 2012.
- [13] Chung-Yi Huang, Lai-Jun Luo, Pei-Yuan Lee, Jiing-Yih Lai,,Wen-Teng Wang, Shang-Chih Lin, Efficient Segmentation Algorithm for 3D Bone Models Construction on Medical Images, *Journal of Medical and Biological Engineering*, Vol.31, n.6, pp.375-386, 2010.
- [14] A .Morenoa, C.M. Takemuraa, O .Colliotc ,O .Camarad, I .Blocha, Using anatomical knowledge expressed as fuzzy constraints to segment the heart in CT images, *Pattern Recognition*,Vol.41,n.8, pp. 2525 – 2540, 2008.
- [15] X.M. Pardo. , M.J. Carreira , A. Mosquera, D. Cabello, A snake for CT image segmentation integrating region and edge information, *Image and Vision Computing*,Vol.19, n.7, pp.461-475, 2001.
- [16] Zikuan Chen, Sabee Molloi, Automatic 3D vascular tree construction in CT angiography, *Computerized Medical Imaging and Graphics*,Vol.27, pp.469–479,2003.
- [17] Yaser Ajmal Sheikh, Erum Arif Khan, Takeo Kanade, Mode-seeking by Medoidshifts. *Computer Vision (ICCV)*, IEEE International conference on, pp.1-8,2007.
- [18] V.V.Gomathi, Dr.S.Karthikeyan, An Efficient Clustering based Segmentation Algorithm for Computer Tomography Image Segmentation, *Journal of biomedical engineering and medical imaging*, vol.1, n.3, pp. 1-11, 2014.
- [19] J.L Marroquin, F. Giosi, Some Extensions of the K-Means Algorithm For Image Segmentation and Pattern Classification, Technical Report, MIT Artificial Intelligence Laboratory,1993.
- [20] M.Luo, Y.F.Ma ,H.J. Zhang, A *Special Constrained K-Means approach to Image Segmentation* ,proceedings of the Fourth International Conference on Information Communications and Signal Processing and the Fourth Pacific Rim Conference on Multimedia,Vol.2,pp.738-742,2003.
- [21] V.V. Gomathi, S. Karthikeyan, Performance Analysis of Distance Measures for Computer tomography Image Segmentation, *International Journal of Computer Technology and Applications*, Vol. 5, n.2, pp. 400-405,2014.
- [22] Beecks.C, Uysal M.S, Seidl.T, Signature Quadratic Form Distances for Content-based Similarity, *ACM CVIR* 2010.
- [23] V.V. Gomathi , S. Karthikeyan, A Proposed Hybrid Medoid Shift with K-Means (HMSK) Segmentation Algorithm to Detect Tumor and Organs for Effective Radiotherapy, *Lecture Notes in Computer Science(Springer)*, Vol. 8284, pp.139-147, 2013.
- [24] Ebrahim, M.J., Pourghassem, H., A novel automatic synthetic segmentation algorithm based on mean shift clustering and canny edge detector for aerial and satellite images, (2012) *International Review on Computers and Software (IRECOS)*, 7 (3), pp. 1122-1129.
- [25] Ali Hassan Al-Fayadh, Hind Rostom Mohamed ,Raghad Saaheb Al-Shimsah, CT Angiography Image Segmentation by Mean Shift Algorithm and Contour with Connected Components Image, *International Journal of Scientific & Engineering Research*, Vol.3, n. 8, pp.1-5, 2012.
- [26] Keh-Shih Chuang , Hong-Long Tzeng , Sharon Chen , Jay Wu , Tzong-Jer Chen, Fuzzy c-means clustering with spatial information for image segmentation, *Computerized Medical Imaging and Graphics*,Vol.30,n.1, pp. 9–15, 2006.
- [27] Wenbing Tao, Hai Jin, Yimin Zhang, —Color Image Segmentation Based on Mean Shift and Normalized Cuts, *IEEE Transactions on systems, man, and cybernetics—part b: Cybernetics*, Vol. 37, n. 5, pp.1382-1389, 2007.
- [28] Zhou Wang and Alan C. Bovik, Ligang Lu, Why is image Quality Assessment So Difficult.
- [29] Zhou Wang, Member, Alan C. Bovik, Image Quality Assessment: From Error Visibility to Structural Similarity, *IEEE Transactions On Image Processing*, Vol. 13, n. 4, pp.1-14, 2004.

Authors' information

¹Ph.D Research Scholar, Research and Development Centre, Bharathiar University, Coimbatore, Tamilnadu, India.

²Assistant Professor, Department of Information Technology, College of Applied Sciences, Sohar, Oman.



V. V. Gomathi Completed MCA in Bharathidasan University, India in 2003 and MPhil in Bharathiar University, Coimbatore, India in 2005. She worked as a Senior Lecturer in Karpagam University, Coimbatore, India till 2013. At present she is a PhD Research Scholar in Research and Development Center, Bharathiar University, Coimbatore, India. Her main scientific interests are Medical Image Processing, Medical Image Mining, and Database Technologies. She has Published 7 papers in International Journal, 2 papers in National Magazine, presented papers in 3 International Conferences and 10 National Conferences. Her Current interests lie in the Development of Algorithms for Medical Image Mining, Medical Image Processing Problems, and real time medical domain problems.



S. Karthikeyan got a Ph.D. in Computer Science and Engineering from Alagappa University, Karaikudi, India in 2008. He worked as a Head, Department of Computer Science, Karpagam University, Coimbatore, India from 2001 to till 2008. At present he is working as a Assistant Professor in Information Technology, College of Applied Sciences, sohar, Sulatanate

of Oman till this date. His main scientific interests are Cryptography, Network Security and Data Mining. He published 50 papers in International Journal, 8 papers in National Journal and presented papers in 12 International Conferences 8 National Conferences. His Current interests lie in the Development of Network security Algorithms for Confidentiality, Integrity, Authentication and Key management. At present his research focused on network security for sensors.

Dr.S.Karthikeyan is a senior member in Association of Computer Electronics and Electrical Engineers (ACEEE), India, International Association of Computer Science and Information Technology, Singapore and a member in Computer Science Teachers Association (CSTA), ACM, New York, USA, International Association of Engineers - Computer Science/Wireless Networks, Canada/Spain, and Indian Science Congress Association, Calcutta, India.

Practical Analysis of Impact of Transmitter Hardware Impairments for MIMO Channel Measurement

P. Vijayakumar¹, S. Malarvizhi²

Abstract – Multiple Input Multiple Output MIMO is one of the promising technologies for the high data rate wireless Communication and adapted in many wireless standards. MIMO Channel measurement is an essential process based on which the decoding of symbols, equalization and adaptive transmission can be implemented. Hence, real time measurement of MIMO channel response and characterization is important one. Many research works has been carried out in the area of MIMO channel estimation and measurements under the assumption of no hardware impairment. But the practical hardware suffer from hardware impairments like IQ gain imbalance, skew and phase noise. MIMO systems are very sensitive for the above impairments which degrades the achievable capacity of the MIMO system. This paper presents an experimental real time study of indoor MIMO capacity measurement incorporating the hardware impairments. Algorithms are written in LABVIEW and implemented in SDR platform of National Instrument 6.6 GHz vector signal generator PXIe 5673 and vector signal analyzer PXIe 5663. **Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords: Channel Measurement, Experimental Study, MIMO Transmitter Hardware Impairment, Least Square Estimate

I. Introduction

Multiple Input Multiple Output MIMO technology gives a reliable high data rate transmission and hence adapted in many standard. There are many research work going on in this area ([1]-[17]).

But the MIMO system performance degrades to a great extend in presence of hardware impairments like IQ gain imbalance, phase noise, quadrature skew, DC offset etc., Many research work are towards the channel estimation and capacity evolution and a few experimental work has been reported considering the hardware impairments.

Simulation study on the impact of hardware impairments have been reported in literature. Phase noise effect on multi antenna based OFDM system was studied [2], [10]-[13] and IQ imbalance effect on Zero Force-MIMO system was reported in [3], [18]. Impact of Tx-RF impairments on the performance of MIMO detection algorithms has been studied in [4], where the hardware impairment has been modelled as noise. This work claims that the performance of linear MIMO detection schemes may improve with the addition of residual Tx-RF impairments as noise source. In recent work [5] impact of the hardware impairments on a large scale MIMO was studied, where a new system model incorporates the transceiver hardware impairments both at the Base Station BS with large antenna arrays and at the single antenna in User Equipment UE was studied.

Results proves that the hardware impairments create a finite ceilings on the channel estimation accuracy and the

downlink/uplink capacity of each UE, furthermore the results proved that a huge degrees of freedom offered by massive MIMO can be used to reduce the transmit power to tolerate larger hardware impairments.

The impact of transceiver impairments in a two way amplify and forward relay network configuration was investigated in [6], the effective signal-to-noise and distortion ratio at both transmitter nodes are obtained and used to deduce exact and asymptotic closed form expressions for the outage probabilities, as well as tractable formulations for the Symbol Error Rate (SER) was studied under simulation. All the above works reported have analyzed the impact of hardware impairments in simulation study only. This paper analyze the impact of the hardware impairments practical by conducting an experiment for a 2X2 MIMO setup on real time using National Instrument PXIe SDR platform.

The remaining sections of this paper is organized as follows: section II discusses the channel estimation and capacity evolution in the presence of hardware impairments, section III deals with the experimental setup, section IV discuss the experimental results and conclusion are drawn in section V conclude the reach work with summary of current and future work

II. Hardware Impairments on Channel Estimation and MIMO Capacity

Phase noise is a random disturbance in the phase of the carrier signal.

This causes a rotation and noise like blurring of the signal constellation. When phase noise present in the local oscillator, rotation and blur due to noise can be observed. Phase noise creates instantaneous frequency error of the baseband signal. Modulation schemes like QAM, the phase noise prevents carrier recovery and causes the spinning of the constellation plot.

Quadrature skew is a source of error which occurs in the LO splitter of arbitrary wave form generator, IQ modulator modules, divides the LO signal into I and a Q signals with exact 90° phase different. While an ideal system would result in each of these being exactly 90° out of phase but in a practical system there will be some amount of skew. Lower order modulation schemes like QPSK quadrature skew has a relatively little effect on system throughput. But if sources of error is high then there will be significant impairment.

IQ mismatch or imbalance is another noise source that introduced in hardware of IQ modulator and demodulator during IQ modulation at the transmitter and IQ demodulation at the receiver. I and Q being two separate signals, each one is created and amplified independently. Inequality of this gain between I and Q paths is called as IQ gain imbalance that results in incorrect positioning of each symbol in the constellation causing errors in recovering the data. In heterodyne receiver, when IF-down conversion takes place IQ imbalance causes the respective image channel to mix partially with the desired signal.

This noise not only created in I and Q signal generator but It may be created by slightly different conversion losses in I and Q up convertor or in down convertor mixer hardware's or by different filter losses in I and Q signal paths [1]. Hardware impairments like skew, phase noise and IQ imbalance can be modeled as additive noise in the RF transmitter [6] as shown the Fig. 1. At the receiver side in the presence of transmitter impairment, the received signal $y(k)$ of MIMO system can be modeled as:

$$y(k) = (s(k) + N_h(k))H(k) + N \quad (1)$$

where $y(k)$ is the k^{th} instant received signal, $s(k)$ is k^{th} instant transmitted signal; $N_h(k)$ is the k^{th} instant total hardware impairment noise, $H(k)$ is MIMO channel matrix at k^{th} instant and N is the additive noise in the channel.

In our model all the hardware impairments are modeled as a total noise $N_h(k)$ and the level of the total impairments is considered as α , with this the impact of the hardware impairment on the MIMO channel estimation using Least Square (LS) method and the impact of the impairment noise on the MIMO capacity is studied. The MIMO LS channel estimator for the ideal hardware is:

$$\hat{H} = Y \cdot S_t^H (S_t S_t^H)^{-1} = Y \cdot M_t^+ \quad (2)$$

where:

$$M_t^+ = S_t^H (S_t S_t^H)^{-1}$$

In presence of the hardware impairment of level α , the estimated channel is:

$$\hat{H}_{im} = Y \cdot S_t^H (S_t S_t^H + (\alpha^2 SNR + 1)I_{MR})^{-1} \quad (3)$$

where, I_{MR} is identity matrix with size of $MR \times MR$ and MR is no of receiver antennas. From \hat{H}_{im} equation, one can observe that the estimation of H introduces estimation error which is the function of square of the level of the hardware impairment α^2 .

Transmitter hardware impairment noise appears as spatially colored noise at the receiver. By combining the hardware noise and channel noise the total noise can be written as:

$$N_{total}(k) = N_h(k)H + N \quad (4)$$

In a compact form, the received signal is:

$$y = HS + \sqrt{K}W \quad (5)$$

where $W \sim CN(0, IMR)$ white noise at receiver.

K is the covariance of the aggregate noise that is given by:

$$K = \sigma_t^2 H H^H + \sigma_r^2 I_{MR} \quad (6)$$

Channel capacity in the presence of spatially colored noise. Based on the covariance matrix K is given as below:

$$C_{im} = \log_2 \det \left(I_{MR} + \frac{H H^H}{K} \right) \quad (7)$$

From the above equation it is very clear that the channel capacity of the MIMO system get reduced because of the noise covariance of the hardware impairments. This reduction of the capacity and channel estimation error because of the hardware impairments is going to be analyzed experimentally on 2X2 MIMO s.

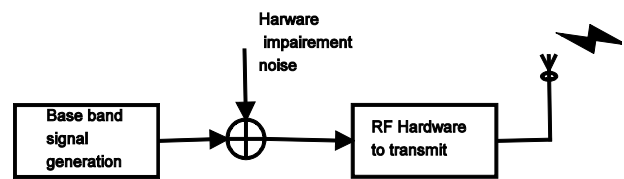


Fig. 1. System model

III. Experimental Setup

The 2X2 MIMO transmitter with hardware impairment is constructed using two NI PXIe 5673 Vector Signal Generator (VSG). PXIe 5673 VSG consist of 5652 RF signal generator which act as a RF local oscillator, 5450 arbitrary wave form generator and a 5611 module called IQ modulator acts as a RF up convertor by mixing the user signal by RF LO signal.

To study the impact of the hardware impairment different level of impairment of IQ gain imbalance, quadrature skew and LO phase noise is introduced that are marked as N_h additive noise and shown in the Fig. 2.

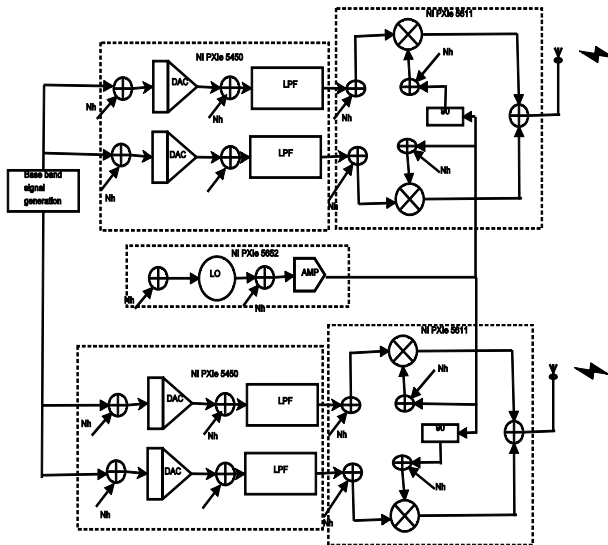


Fig. 2. 2X2 Transmitter setup with hardware impairments

The 2X2 receiver is constructed by using two NI PXIe5663 Vector Signal Analyzers (VSA). Each vector signal analyzer consist of a RF down convertor NI PXIe 5601, a digitizer NI PXIe 5622 modules and a NI PXIe 5652 RF signal generator that acts as LO source for mixing to both VSA. The RF down convertor is used to convert the RF signal into IF signal, then the digitizer is used to convert the analog signal into digital signal, finally the digitized samples are given to the MIMO LABVIEW code to decode and analyze the received signal.

The channel estimation algorithm and the capacity evolution code are written by using LABVIEW and the performance is analyzed. Fig. 3 shows the receiver setup that is used to receive and analyze the signal.

The transmitter and receiver set up parameter used for generation and reception of RF signal is given in the Table I with possible range of impairments that can be added with the transmit signal. Fig. 4 shows the pictorials representation of the 2X2 MIMO hardware that is used for this experimental study.

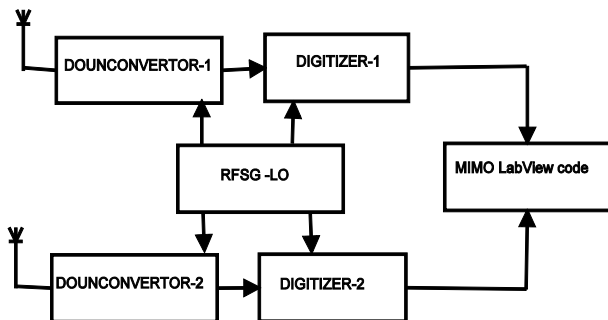


Fig. 3. 2X2 Receiver setup (RFSG-Radio Frequency Signal Generator)

TABLE I
TRANSMITTER AND RECEIVER SYSTEM SET UP

Parameter	Transmitter / receiver side
Transmitter Power	0dBm;-10dBm;-30dBm / NA (Not Applicable)
Carrier frequency	1GHz /1GHz
Trigger power level	NA /-40dBm
Symbol rate	100k samples per sec/100k samples per sec
Modulation	4QAM /4QAM
Pulse shaping filter	Root raised cosine with alpha=0.5 and order=4 / Root raised cosine with alpha=0.5 and order=4
I/Q gain imbalance	0 to 6dBm/ NA
Quadrature Skew	0 to 30 °/NA
Phase noise	-50 to 20dBm at 4KHz offset / NA



Fig. 4. Hardware set up for measurement

IV. Result and Discussions

The impact of the hardware impairments IQ gain imbalance, quadrature skew and phase noise on channel estimation and 2X2 MIMO system capacity are evaluated for different transmitter power 0dBm,-10dBm,-30dBm.

Fig. 5 shows that the capacity of the MIMO system linearly decrease with respect to the increased IQ gain imbalance. since our hardware will only allow to apply IQ gain imbalance up to 6dBm. We have varied the gain from 0 to 6 dBm and the impact on the capacity is analysed.

At zero IQ gain imbalance the capacity is for ideal case, which is higher for 0dBm transmitted power.

Fig. 6 shows that percentage capacity decrement will happen from without impairment level with the increased IQ gain imbalance, at a gain imbalance of 6dBm around 90% of the capacity is reduced from the ideal case. To study the impact of the IQ gain imbalance at various SNR level condition, we have analyzed the capacity for transmitter power from -30dBm to 0dBm, from Fig. 7 we observe that at high SNR level the capacity decreasing linearly with respect to the increased impairment but at low SNR i.e., for the -30dBm transmitter power the change in the capacity is very minimum.

Similarly to study the impact of the channel estimation on IQ gain balance impairment at various transmitter power is studied. The average channel gain from the H matrix is calculated as channel gain that channel gain vs IQ gain imbalance for various transmitter power is plotted as in Fig. 8.

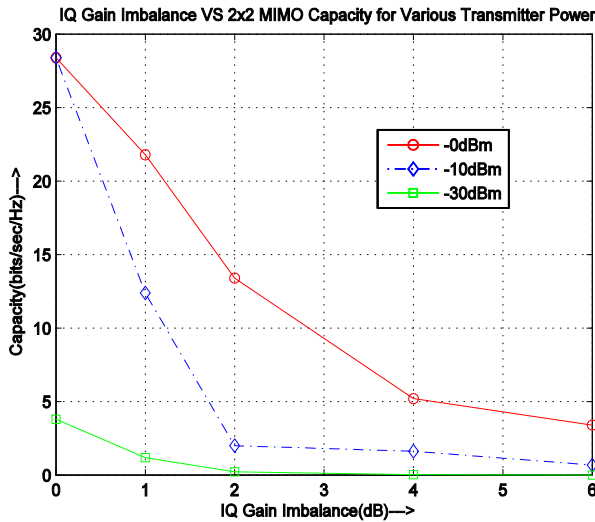


Fig. 5. Impact of IQ Gain Imbalance for different transmitter power

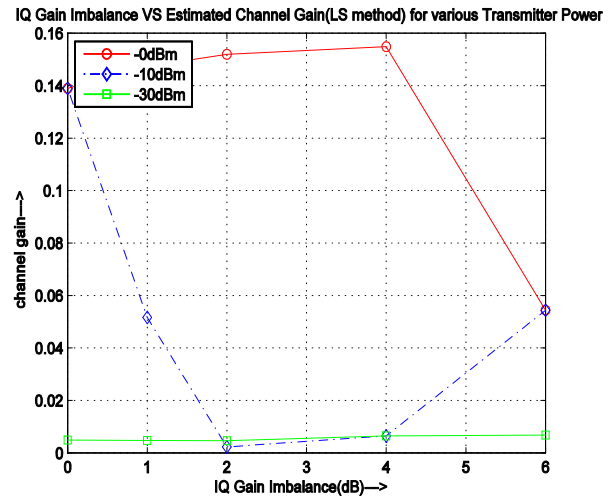


Fig. 8. IQ Gain imbalance VS Estimated Channel Gain for different Transmitter Power

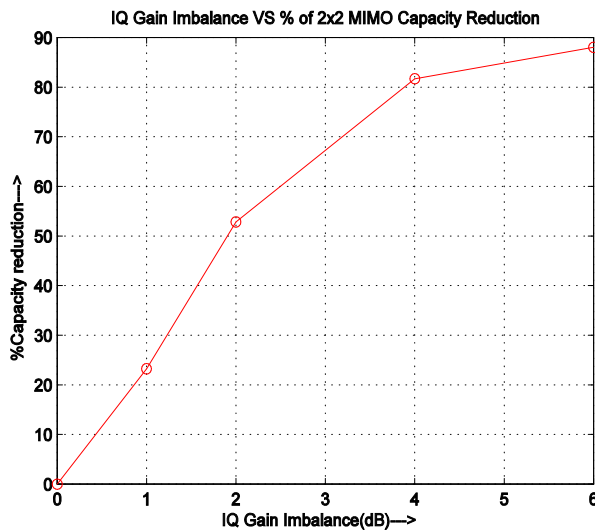


Fig. 6. Percentage capacity reduction with respect IQ gain imbalance

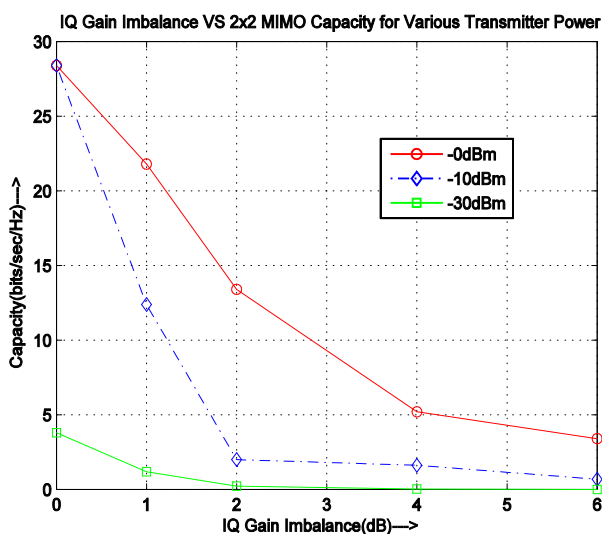


Fig. 7. Impact of IQ Gain Imbalance for different transmitter power

It is observed that at high SNR region (-0dBm, -10dBm) the average estimated channel gain exhibits an oscillatory behavior but at low SNR case at -30dBm transmitter power, the channel gain undergo a very minimum change on estimation, Same kind of the analysis is done for quadrature skew on capacity and channel gain that is plotted in Fig. 9, Fig. 10 and Fig. 11 respectively.

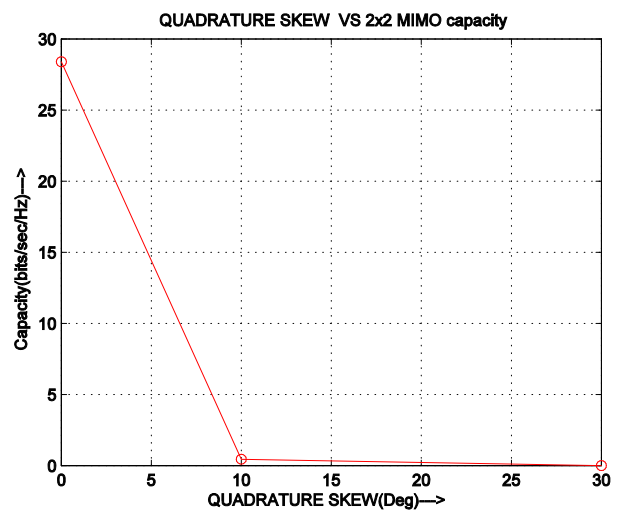


Fig. 9. 2X2 MIMO capacity at two different quadrature skew level

From that graphs we observe that for skew of 10° there is a high degradation in performance after that increasing skew has very little impact on capacity and channel gain. We can conclude that the presence of the skew irrespective of value has the same impact more or less. The impact of the phase noise is studied by keeping the transmitter power at 0dBm and the plotted on Fig. 12 and Fig. 13 by varying the phase noise from -60dB to -20dB at 4kHz frequency offset. From Fig. 12 we can observe that there is a linear decrement of capacity with respect to increased phase noise.

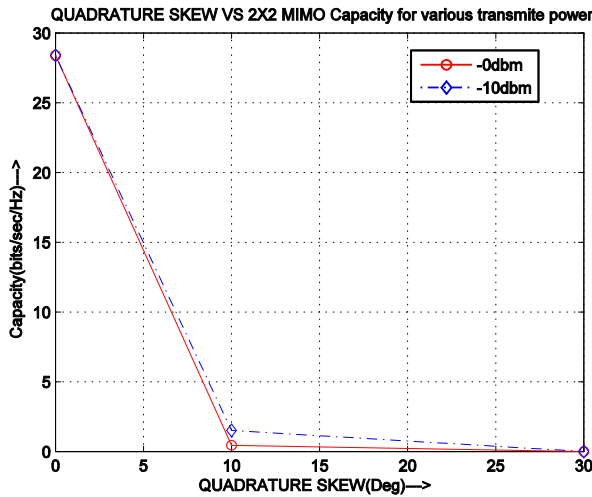


Fig. 10. Impact of quadrature skew on 2X2 MIMO capacity at two different Transmitter Power

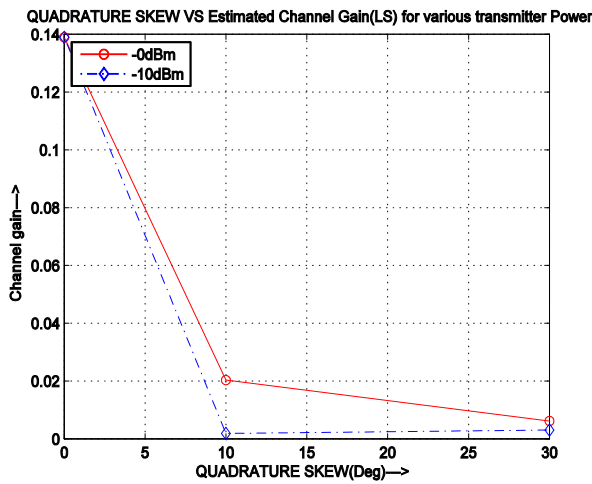


Fig. 11. Estimated channel gain for different Skew and Transmitter Power

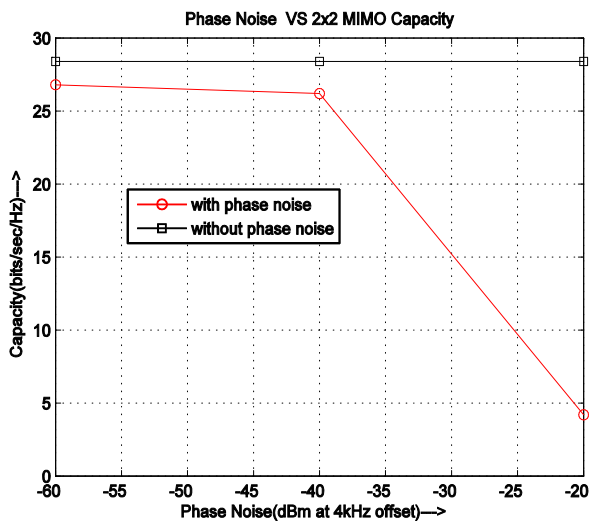


Fig. 12. Impact of Phase Noise on 2X2 MIMO capacity

From Fig. 13 we can observe that the estimated average channel gain exhibits an oscillator behavior.

The observations from the graph are tabulated as Table II. From the table we observe that, the quadrature skew has dominant impact in comparison with other impairments. Finally giving all the impairments at the same time the system performance is observed. We have applied the I/Q gain imbalance of 3 dB, quadrature skew of 30 degree and phase noise density of -40 dBc/Hz at 4kHz offset.

Fig. 14 shows the 4QAM constellation diagram with all impairments of the above quantity added with the system.

TABLE II
IMPACT OF TRANSMITTER HARDWARE IMPAIRMENTS

Impairment	Value	Impact on Channel gain	Impact on 2X2 Capacity
I/Q gain imbalance	6dB	71% reduction	90% reduction
Quadrature Skew	30Degree	93% reduction	97% reduction
Phase Noise	-20dBm at 4KHz offset	42% reduction	86% reduction

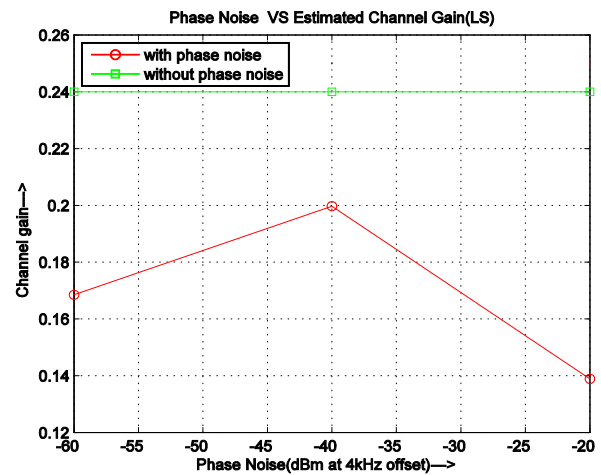


Fig. 13. Estimated channel gain with respect phase noise

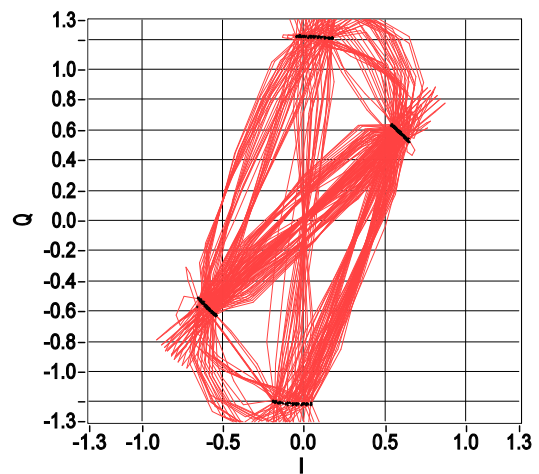


Fig. 14. 4-QAM signal constellation graph with all impairments

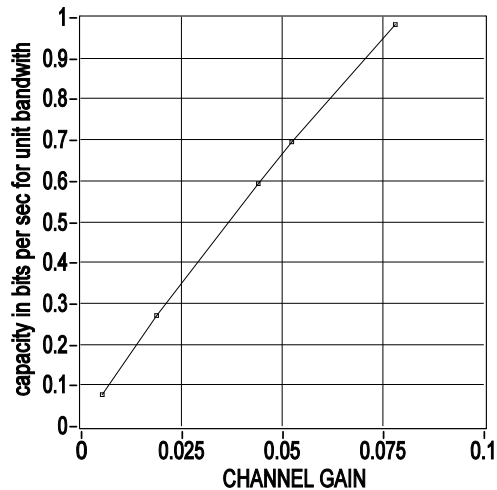


Fig. 15. Observed 2x2 MIMO capacity in presence of all impairments

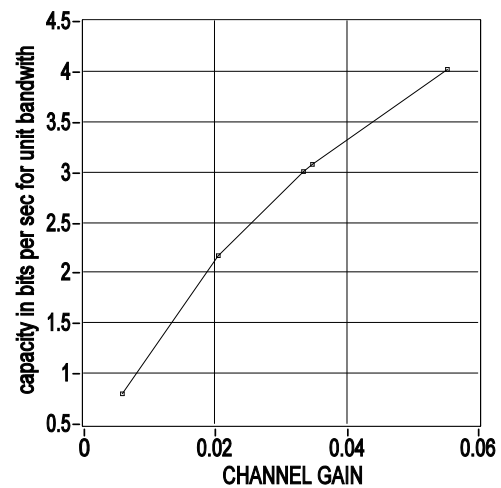


Fig. 16. Instantaneous 2x2 MIMO capacity for I/Q gain Imbalance of 6 Db

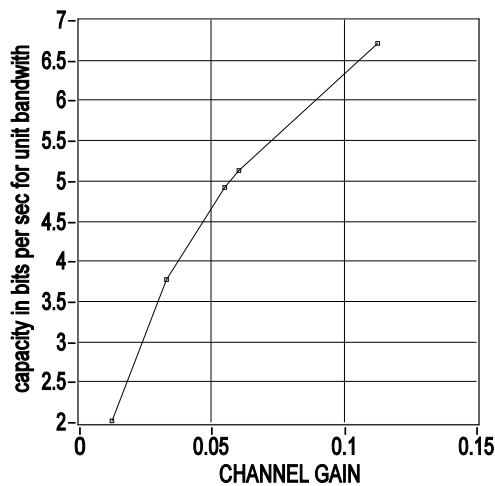


Fig. 17. Instantaneous 2X2 MIMO capacity for skew 30 degree

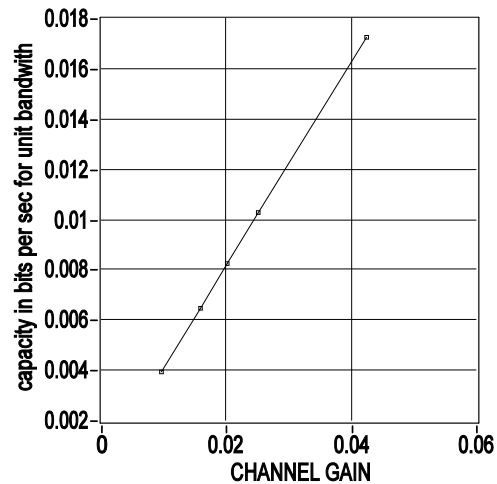


Fig. 18. Instantaneous 2x2 MIMO capacity for phase noise 20 dBm/Hz at 4kHz offset

The capacity of 2x2 MIMO system observed for the above case is given the Fig. 15 we can observe that in presence of the all impairments the capacity of the system is more or less same as that of skew maximum impairment case. The above analysis graphs shows time averaged measured performance for illustrate the real scenario we have captured some instantaneous performance same graph for peak impairments values for each case.

From Figs. 16 to 18 the instantaneous capacity that is captured on real time is shown. The phase noise impairment is dominating among all other impairments.

V. Conclusion

An experimental real time study of indoor 2X2 MIMO channel estimation and capacity measurement has been carried with various transmitter hardware impairments for various transmitter power. The capacity analyses prove that capacity of the MIMO system decreases

linearly with respect the increased impairment level, moreover it is the evident that the noise model based MIMO capacity in presence of hardware impairments is a valid. From the channel estimation gain results, it is proved that the channel estimation error is more of oscillatory behavior at high SNR level (at high transmitter power), at low SNR level the estimation error is very minimum.

Acknowledgements

This work is supported by the infrastructure created under Fund for Improvement of Science & Technology FIST - Department of Science and Technology DST of ECE department, SRM University.

References

- [1] R. W. Lowdermilk and f. j. harris, "Vector signal analyzer implemented as a synthetic instrument," IEEE Trans. Instrument. & Measurement, vol. 58, pp. 281-290, Nov. 2009.

- [2] T. Schenk, X.-J. Tao, P. Smulders, and E. Fledderus, "Influence and suppression of phase noise in multi-antenna OFDM," in Proc. IEEE VTC 2004-Fall, vol. 2, Sep. 2004, pp. 1443–1447.
- [3] T. C. W. Schenk, E. R. Fledderus, and P. F. M. Smulders, "Performance analysis of zero-IF MIMO OFDM transceivers with IQ imbalance," *Journal of Comm.*, vol. 2, no. 7, Dec. 2007.
- [4] H. Suzuki, T. Tran, I. Collings, G. Daniels, and M. Hedley, "Transmitter noise effect on the performance of a MIMO-OFDM hardware implementation achieving improved coverage," *IEEE Journal on Sel. Areas in Comm.*, vol. 26, no. 6, pp. 867–876, Aug. 2008.
- [5] Emil Björnson, Jakob Hoydis, Marios Kountouris, and M'rouane Debbah, "Massive MIMO Systems with Non-Ideal Hardware: Energy Efficiency, Estimation, and Capacity Limits", arXiv:1307.2584v2 [cs.IT] 7 Jan 2014.
- [6] Michail Matthaiou, Agisilaos Papadogiannis, Emil Björnson, and M'rouane Debbah, "Two-Way Relaying under the Presence of Relay Transceiver Hardware Impairments", *IEEE COMMUNICATIONS LETTERS*, VOL 17, NO. 6, JUNE 2013.
- [7] Ranjitham, G., Shankar Kumar, K.R., A mathematical modeling and simulation analysis of lattice reduction algorithms for large MIMO detections, (2014) *International Review on Computers and Software (IRECOS)*, 9 (6), pp. 1065-1073.
- [8] Mattered, D., Paura, L., Sterle, F., Widely linear decision-feedback equalizer for time-dispersive linear MIMO channels, (2005) *IEEE Transactions on Signal Processing*, 53 (7), pp. 2525-2536.
- [9] Mattered, D., Paura, L., Sterle, F., Widely linear MMSE equaliser for MIMO linear time-dispersive channel, (2003) *Electronics Letters*, 39 (20), pp. 1481-1482.
- [10] Mattered, D., Tanda, M., Blind symbol timing and CFO estimation for OFDM/OQAM systems, (2013) *IEEE Transactions on Wireless Communications*, 12 (1), art. no. 6397549, pp. 268-277.
- [11] Mattered, D., Tanda, M., Bellanger, M., Frequency-spreading implementation of OFDM/OQAM systems, (2012) *Proceedings of the International Symposium on Wireless Communication Systems*, art. no. 6328353, pp. 176-180.
- [12] Mattered, D., Tanda, M., Preamble-based synchronization for OFDM/OQAM systems, (2011) *European Signal Processing Conference*, pp. 1598-1602.
- [13] Mattered, D., Tanda, M., Data-aided synchronization for OFDM/OQAM systems, (2012) *Signal Processing*, 92 (9), pp. 2284-2292.
- [14] Kalkan, Y., On the advantages of frequency-only MIMO radar, (2013) *International Journal on Communications Antenna and Propagation (IRECAP)*, 3 (3), pp. 163-168.
- [15] Tami, A., Keche, M., Ouamri, A., New joint blind channel estimation and data detection through a time varying MIMO channel, (2013) *International Journal on Communications Antenna and Propagation (IRECAP)*, 3 (5), pp. 255-260.
- [16] Manasra, G., Najajri, O., Rabah, S., Arram, H.A., DWT based on OFDM multicarrier modulation using multiple input output antennas system, (2012) *International Journal on Communications Antenna and Propagation (IRECAP)*, 2 (5), pp. 312-320.
- [17] Nawaz, T., Baig, S., Khan, A., The performance comparison of coded WP-OFDM and DFT-OFDM in frequency selective rayleigh fading channel, (2011) *International Journal on Communications Antenna and Propagation (IRECAP)*, 1 (6), pp. 500-505.
- [18] Mattered, D., Sterle, F., ML estimation of receiver IQ imbalance parameters, (2007) *2007 International Waveform Diversity and Design Conference*, WDD, art. no. 4339401, pp. 160-164.

Authors' information

¹Assistant professor (senior grade), SRM University, Chennai.

²Professor, SRM University, Chennai.



Vijayakumar P. is from Thirumanoor, Salem, currently living in Chennai. He has completed Master in Engineering (M.E) applied electronics in college of engineering, Guindy, Anna University, Chennai, Tamilnadu, India at 2005. He has completed B.E in electronics and communication engineering from Madras University at 2000. His current research interests are in the area of MIMO wireless communication, cognitive radio networks, Software Defined Radio, intelligent systems. He is member of professional societies: IEEE, ISTE, International association of computer science and information technology, International Association of Engineers (IAENG), Universal Association of Computer and Electronics Engineers



Dr. S. Malarvizhi is working as professor in SRM University, Chennai. She finished her Ph.D. Wireless communication College of Engineering, Guindy, Chennai-25, 2006. M.E. Applied Electronics, GCT Coimbatore, Anna University, 1990. B. E. Electronics Communication Engineering Arulmigu Meenakshi amman College of Engineering, Kanchipuram, Madras University, 1989. Her research interests are in the area of Wireless Communication. Sensor Communication, Communication algorithms implementation in FPGA.

Meta-Classifer Based on Boosted Approach for Object Class Recognition

Noridayu Manshor, Amir Rizaan Abdul Rahiman, Raja Azlina Raja Mahmood

Abstract – Object class recognition deals with the classification of individual objects to a certain class. In images of natural scenes, objects appear in a variety of poses and scales, with or without occlusion. Object class recognition typically involves the extraction, processing and analysis of visual features such as color, shape, or texture from an object, and then associating a class label to it. In this study, global shape and local features are considered as discriminative features for object class recognition. Both local and shape features are combined in order to obtain better classification performance for each object class. A meta-classifier framework is proposed as a model for object class recognition. Meta-classifier is used to learn a decision classifier that optimally predicts the correctness of classification of base classifier for each object. In this framework, base classifiers based on boosting approach are trained using the local and global shape features, respectively. Then, these classifiers results are combined as input to the meta-classifier. The results from classification experiments showed that meta-classifier based on boosted approach performs better compared to some state-of-the-art approaches in object class recognition. **Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords: Meta-Classifier, Boosting, Classifier Fusion, Object Class Recognition

I. Introduction

Due to the recent developments in technology, huge amounts of images are easily generated using relatively affordable devices such as digital cameras, video camcorders and mobile phones. The Internet has also allowed easy and ubiquitous access, indirectly contributing to the massive consumption of images data.

Due to these, and also due to the wide availability of mass storage devices, the amount of images data is growing to colossal proportions.

In order for effective and intuitive retrieval, these images should be annotated. Annotation is the process of assigning meaningful labels to data, mostly via a set of keywords. To perform object-level annotation, regions of interest (ROI) or the object itself has to be extracted firstly from the image and its spatial relationships identified.

The task of object classifications is known as Object Class Recognition. The main task in Object Class Recognition is to discriminate between objects of one class and those of other classes. The challenges of Object Class Recognition are to find class models that are invariant to changes in appearance within a class, while being discriminative enough to distinguish between objects from different classes.

Specifically, Object Class Recognition involves extracting features from an identified object, and then associating a label to it representing the object's class. An object class furthermore can contain various objects of the same genre.

For example, the object class “flower” may consist of a variety of flowers and an object class “cars” may consist of cars of different brands and models with a variety of shapes and sizes. For instance, Fig. 1 shows some examples of images where the car class appears. It is straightforward to perceive that these three cars are very different in terms of visual appearance, but all must be classified within the same class.



Fig. 1. The ‘car’ class

Objects can have a variety of poses, scales, with or without occlusion, depending on the viewing direction, angle, and distance. The object class can be understood by a computer based on its visual features such as shape, color and texture. The challenge is to map or relate these visual features to a higher level conceptual representation that is closer to human understanding. The discrepancy in understanding between machines and humans is known as the “Semantic Gap”. At this juncture, most related research efforts work on mapping an object within an image to a suitable concept [1]-[3]. The challenges of Object Class Recognition are not only related to the features point of view but also depending on the classifier design. In the past, fusing different classifiers has managed to improve classification accuracy [1], [4], [5].

The main idea is that, by combining different classifier outputs, higher accuracy can be achieved as opposed to using just one classifier. To improve classification accuracy, a suitable fusion method and selection of appropriate classifiers have to be taken into account.

The fusion of classifiers can be done at the feature-level and decision-level. In the past, Content-based Image Retrieval (CBIR) researches fused several features into a single feature vector [6], [7]. However, this has its limitations such as increased computational time due to the curse of dimensionality [8], [9], [40]. To overcome this, fusing at the decision-level is more promising by constructing a multiple classifier for each image feature [8], [10]. The final decision is identified based on combination outputs from each classifier.

In the case of this work, the classifier fusion is adopted due to the diversity of information from the local and global features. The final predicted object class result is produced through the integration of outputs obtained from the discriminant function of different classifiers.

The computational burden of the base classifier also motivates us to adopt classifier combination. Since the different classifier may produce different results, thus, classifier fusion can be used to balance the performance of a set of classifiers in order to increase the classification accuracy. In order to do so, this study needs to exploit the different learning algorithms for improving the performance of object class recognition. As stated in [1], [4], [5], [11], boosting approach has improved the accuracy of object class recognition. Thus, in this study, boosting technique is applied in constructing and developing the meta-classifier approach by combining global shape and local features.

II. Related Work

This study deals with the problem of object class recognition. This study concentrates on the use of different features and aggregation techniques for combining global shape and local features. Previous research works focus on combining several features to increase the object class recognition performance. These various existing approaches differ mainly in three ways, 1) the types of features being considered and their combination approaches, ii) the learning algorithms being adopted for classification, and iii) the application domain.

Not all objects of an image can be recognized by a single feature. The usage of one feature of an object causes the problem of ambiguity in recognition. For example, if color features are extracted from 'green banana' and 'green apple' images, it will be very difficult for the computer to perform classification based on this feature alone, since both images may have similar green color. Misclassifications can potentially occur leading to poor recognition accuracy.

To resolve the above mentioned scenario, current research has adopted a combination of features in order to improve object class recognition.

Mostly, an approach combining various local features is used, or alternatively, both global and local features are combined [1][5][12]. Previous research in CBIR used a combination of various global features of color and shape [6], [7]. The methods to combine features may be categorized into two types.

The first type is known as feature fusion, whereas the second type is decision fusion. Feature fusion is a method of combining multiple features into a single feature vector before it is fed it into a classification engine. Decision fusion on the other hand is a method where each feature type is firstly learned by their individual classifier.

The output of these individual classifiers is then combined through another classifier that may be based on algebraic rules, fuzzy integral or meta-classifier by using classifier fusion approaches.

As stated in [13]-[15], classifier fusion is a powerful tool for finding the final decision of object classes, specifically by combining many features. Many approaches for classifier fusion can be classified into different ways [8], [16]. Firstly, different learning algorithms can be used for different features, or same data, and second, similar base learning algorithms can be used for different subsets of the training examples or different features. The outputs of the classifiers are then combined by using the various combination approaches mentioned in Fig. 2. This diagram presents the hierarchy of feature fusion and decision fusion approach.

Boosting is one of the popular approaches for solving object class recognition problems [1], [3], [5], [11]. It is based on the decision trees algorithm, which is fast to train and has well-established default parameter settings. Boosting manipulates training data to generate multiple hypotheses. The class label with the highest confidence factor is assigned as the final label.

However, the features are combined into one feature pool (i.e. global and local features), where it will cause misclassification occurred by combining these different types of features.

In this study, the object classes are categorized by constructing a meta-classifier framework.

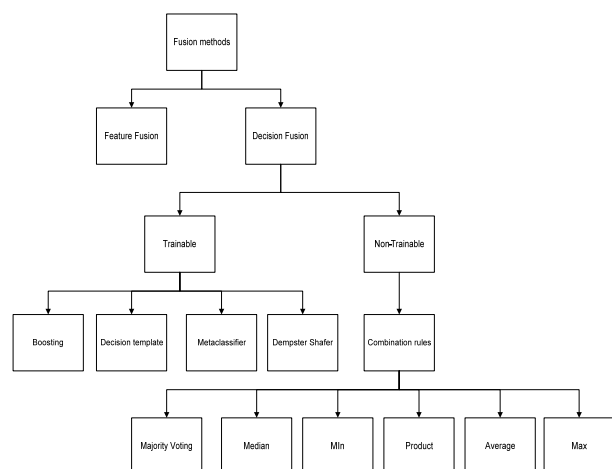


Fig. 2. A hierarchy of fusion methods

The both features are trained separately by individual base classifier and each output of base classifier is combined to be trained by meta-classifier based on boosted approach for final decision. The intuitiveness of this method is to improve the performance on the base learner's predictions by producing the new set of hypotheses from the base learners' outputs.

III. Proposed Meta-Classifier Methodology

In this work, the object features come from different sources, such as global and local features. It consists of heterogenous features. The manner in which features are extracted is totally different. Therefore, a single classifier is not suitable to learn the information contained in both features.

To handle the two types of features considered in this study, different feature combination models are needed. Performance, time and dimensionality of data are important factors that influence the design of the feature combination technique. Thus, decision fusion may overcome the problem of feature fusion where each type of feature is trained by individual classifiers, where the output of these classifiers will determine the final class of the objects.

III.1. Features Extraction

Feature extraction is divided into two parts: global shape features extraction and local features extraction.

For shape description, this study not only focuses on the contour but also on the region. Hence, we introduce both contour based descriptors: Fourier Descriptors (FD) and Elliptical Fourier Descriptors (EFD) as well as region based shape descriptors, namely Moment Invariants (MI). For objects that have prominent regions and less local pattern details, MI is better because focusing on region based to capture not only for contour descriptors but also on the region itself.

For local features, Scale Invariant Feature Transform (SIFT) is used to form the feature vector along with the shape features. To recognize the object class, discriminative features have to be identified to ensure the objects can be grouped into their respective classes. This section briefly explains how these features are extracted from the pre-segmented objects.

A. Global shape features

Shape is an important part of the semantic content of images and is the main feature for recognizing object classes [17] [18]. A global image annotation does not require segmentation process. The shape feature extraction process is done directly from the entire image or image partitions without considering the problem of segmentation. However, for object annotation, it can produce a more precise concept than global image annotation [19]. This is probably because, the classification can be done based on extracted feature in each object.

But, it has the disadvantage of requiring object segmentation and automatic segmentation process may affect the performance of object annotation since automatic image segmentation may not be completely reliable. For example, edge-based segmentation approaches detect the changes of image intensity between two dissimilar regions in order to partition an image. The problem with such approaches occurs when the image has noise such as broken edges and/or overlapping regions [20].

Previously, segmenting an object using region based segmentation will create multiple segments in one object.

This is because an object will consist of two dissimilar groups of pixels which may create two partitions in an object. Consequently, the current automatic segmentation algorithms cannot produce accurate enough shape representations as expected by the user as shown in Fig. 3 [21], [22]. For instance, it is difficult to separate an object from its background, if its boundary's color is similar to the background color.



Fig. 3. Segmentation of objects with complex background [22]

As stated, object annotation best reflects the contents of the image by describing its objects. However, as mentioned earlier, this is highly dependent on the ability to segment objects in the image. This area has been receiving much attention lately and several satisfactory approaches are being proposed [23], [24]. In our study, the focus is on selecting representative features and classifier design assuming image is segmented into objects. Thus, proposed work uses either publicly available presegmented datasets or manually performs segmentation where presegmentation is not available.

i. FD

FD is the boundary descriptors, which are based on the silhouettes of segmented objects [25]. The primary factors, such as invariance under translation, rotation, and scaling, are reasons of why FD is chosen [26].

This invariance factor is very important for object class recognition applications to capture the variability of shapes in an object due to the presence of intra-class variation, pose changes and size changes.

ii. EFD

Similar to FD, EFD is extracted on the closed contour of an object based on its boundary information.

However, EFD generates only minimal dimensions compared to FD for representing the general shape of the object [27]. This feature is widely used in medical image processing [28][29]. It is because EFD gives a good result in shape recognition and low computational cost. The closed contour was defined with differential chain codes, represented as a point coordinate of closed contours. EFD able to represent the shape objects with complex shape or high variability of curvature such as bikes [30]. Due to this ability, EFD are chosen of course to strengthen the global shape features as well as to be as complementary role to the FD.

iii. MI

MI is a shape feature frequently used in image processing, shape recognition and classification [1], [31], [32]. This feature can be extracted from the boundary and interior region of an object.

The moment invariant from Hu [33] proposed seven normalized central moments that are invariant to object scales, translations and rotations. This feature is used in this research because its ability to represent a variety of geometrical features in an object.

It can also deal with disjoint shapes where the boundary information is not available and hence cannot be supported by FD [34].

B. Local features

Local features refer to the features that are extracted based on the interest points detected on the object. The features are extracted around the interest points in an object patch. Local features are computed at multiple points in the object and are consequently more robust to occlusion and clutter.

The local features are most widely used to overcome object class recognition accuracy. SIFT is a very good local feature for objects with different view, scale, image blur, light change and translation [35],[36].

The difference-of-Gaussian is applied to identify the interest points of an object.

128 features are extracted around multiple interest points of object patches. This produces multi-dimensional features for a single object. In order to produce a single feature vector, the Bag of Keypoints (BoK) approach is being adopted in this study [37].

III.2. Meta-classifier Based on Boosted Approach

The decision fusion contributes to the improvement of object class recognition. Decision fusion is a method where each feature type is firstly learned by their individual classifier. In this study, the global features and local features are mapped onto the individual feature space. This may reduce the computation time due to the reduction of the number of feature dimension.

Another reason is the result of individual classifiers will support each other to produce the improvement of classification results. This is because the individual classifier will produce different generalization error.

By using decision fusion approach, the discrepancy of errors is reduced and accuracy is improved.

Boosting in particular, has gained wide interest as a well suited learning approach, because Boosting selects a collection of very diverse features (weak classifiers) that are combined to form a diverse final (strong) classifier [1]. The intuitive idea behind boosting by Freund and Schapire [38] is to train a series of weak classifiers and to iteratively improve the performance of these classifiers.

The algorithm relies on continuously changing the weights of the training set so that those that are frequently misclassified get higher weights. Once all the weak classifiers have been trained, their predictions are then combined through a weighted majority voting scheme.

This way, new classifiers that are added to the set are more likely to classify those hard examples correctly

IV. Results and Discussion

Our approach emphasizes the use of different global shape features combined with SIFT features.

The classification performance of our proposed methods is compared to that of works in [1] and [5] using the benchmark dataset, namely Graz02 [1]. [1] and [5] used different combination of local features only, whereas we used the combination of local and global features, known as Boost_GSLF. The Boosting parameters from [1],[5] are used in this experiment.

The weight of the training dataset are initialized to 1 and the number of iterations for training a weak learner is $T=150$. Table I shows that Boost_GSLF improves the classification accuracy by 10-15% for 'bikes' and 'persons' classes but decrease in performance for 'cars' class in comparison to work proposed by [1] and [5]. It is difficult to identify which features degrade the performance on the 'cars' dataset because different types of features are placed into one feature pool.

TABLE I
CLASSIFICATION ACCURACY USING BOOSTING APPROACH

	Boost_GSLF	[5]	[1]
Bikes	0.891	0.747	0.778
Cars	0.737	0.813	0.705
Persons	0.920	0.813	0.812

To address this problem, the earlier mentioned meta-classifier is proposed, which trains the different features independently. In this study, the use of similar base classifier with different meta-classifier gives different performance. It is important in meta-classifier approach to identify a better base classifier to improve the classification performance. Therefore, empirical studies need to be conducted to design a good classifier fusion that can improve the classification performance for object class recognition.

First, a Support Vector Machine (SVM) is used as a base classifier and as a meta-classifier namely as SVM².

The SVM algorithm may find the hyperplane with the maximum margin that optimally separates the training sets for the base classifiers.

Second experiment, the SVM is used as a base classifier as similar to first experiment, but the Naïve Bayes (NB) is used as meta-classifier to produce the final decision of object class (SVM_NB). NB is a simple density algorithm which is based on probabilistic models that has been proven to provide good prediction accuracy in meta-learning [39].

The output of base classifiers is used as an input for training the meta-classifier. The meta-classifier captures the class posteriori probabilities output from a set of base classifiers. Then, the meta-classifier will identify the final decision of each instance. Table II presents the recognition results of SVM² and SVM_NB using Graz02 dataset. These experiments are performed to show the comparison of the two different meta-classifiers' performances, SVM and NB, using the similar base classifier results of SVM. It shows that simple learning algorithm (NB) for meta-classifier provides better classification accuracy result.

TABLE II
CLASSIFICATION ACCURACY USING SVM AS BASE CLASSIFIERS
AND DIFFERENT META-CLASSIFIER
(SVM² AND SVM_NB)

Object Class	SVM ²	SVM_NB
Bikes	0.621	0.735
Cars	0.670	0.820
Persons	0.800	0.884

Motivated by the performance of Boosting as shown in Table I, this paper adopts the boosting learning algorithm as a base classifier for the proposed meta-classifier approach. However, the feature combination method is different where global shape and local features are not combined together in one feature pool as proposed in Boost_GSLF. Both features' classifiers are trained independently by the boosting algorithm due to the different representation of features.

Then, the output of the base classifiers are again trained to obtain the final decision by the NB (Boost_NB) and SVM (Boost_SVM) algorithms. It is expected that such approach would improve the classification accuracy results.

For boosting, the learning parameters used are discussed in [1] and [5]. For the meta-classifier, the parameter C of SVM and γ of the RBF kernel is chosen through 10-fold cross-validation. Each boosting base classifier gives a confidence measure which is then combined to be used as input to the meta-classifier. In the problem of object class recognition, this is a novel meta-classifier approach using boosting algorithms as a base classifier to train global and local features independently.

Table III shows the classification accuracy obtained by using AdaBoost as a base classifier produces highly accurate prediction result compared to using the well known algorithms such as SVM (Table II). This table compares the results using two different meta-classifiers, namely NB and SVM to gain the final decision of each object class. Generally, the result of Boost_NB shows better classification results than Boost_SVM. From this result, the simple meta-classifier (NB) improved the

accuracy of classification rather than the complicated SVM meta-classifier.

Even though different meta-classifiers are used, there is not much difference in terms of recognition performance except for the 'cars' class. This is not necessarily that complex learning algorithms will give the best result for the meta-classifier approach in producing the final decision of object class recognition.

This is because the selection of a good base classifier by separating global shape and local features has resulted in improvement of object class recognition using meta-classifier approach.

TABLE III
CLASSIFICATION RESULTS BETWEEN BOOST_NB AND BOOST_SVM

	Boost_NB	Boost_SVM
Bikes	0.861	0.861
Cars	0.861	0.719
Persons	0.925	0.926

In addition, increasing the number of features by considering global shape and local features is an important factor in order to improve the result in decision fusion. However, the suitability of how those features are combined and classifier design need to be taken into account as well.

Fig. 4 presents the error rates of the proposed meta-classifier approach with different learning algorithms, compared with traditional boosting (Boost_GSLF) using the proposed global shape and local features. Similar sample sizes with similar features and 150 boosting iterations were used in this comparison. The global shape and local features are trained independently by a classifier except for the Boost_GSLF algorithm, where both features are put into one feature pool and trained by a classifier.

The Boost_NB provides the least error in comparison to others. Although the meta-classifier used is a simple algorithm, Naïve Bayes performs better than SVM. This is probably because the generative models (e.g., Naive Bayes) often outperform discriminative models (e.g., SVM) on smaller datasets, since the latter may overfit.

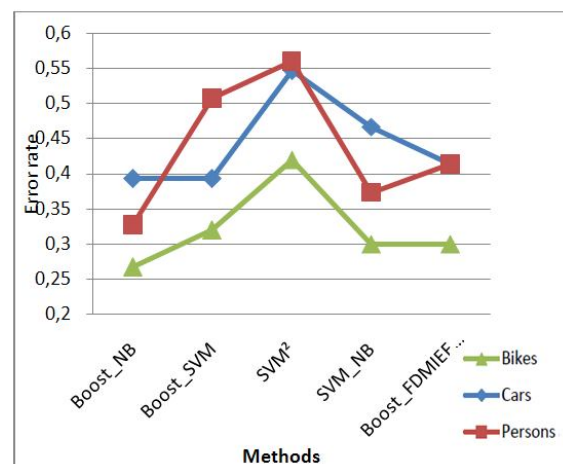


Fig. 4. Error rates for different classifier fusion methods for 'bikes', 'cars' and 'persons' class

V. Conclusion

Object class recognition poses many challenges. For example, various approaches have shown very good classification results but none has dealt with object recognition problem in the real world. All the researches deal with a very limited number of object classes because of the nature of the available datasets. A real world application would require a huge amount of training data which is very time consuming to collect. Furthermore, the stored models have to be in a form that is easy to manipulate in order to make an efficient inference.

In this paper, the combination technique of both global shape and local features is emphasized due to different types of information available in both features. The combination was performed based on a meta-classifier approach. Specifically, both the output from global and local features' classifiers, which are represented by posterior probability, are used as an input for the final classifier to obtain the final label of object class. The proposed meta-classifier is evaluated using simple algorithms such as Naïve Bayes and Adaboost as the base classifiers. The experimental results have shown that the proposed method outperforms other discussed approaches. This can potentially be used within other applications, such as image annotation and retrieval. We have observed that Adaboost as the base classifier, and Naïve Bayes as the meta- classifier (Boost_NB), gave high performance with a classification accuracy of up to 86.1%.

References

- [1] Opelt, A. Pinz, A. Fussenegger, M. and Auer, P.. Generic object recognition with boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 28(3): 416-431, 2006a.
- [2] Opelt, A. Pinz, A. and Zisserman, A. A boundary-fragment-model for object detection. *Proceedings of the 9th European conference on Computer Vision (ECCV'06)*, May 7-13 Graz, Austria. 575-588, 2006b.
- [3] Shotton, J. Winn, J. Rother, C. and Criminisi, A. Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*. 81(1): 2-23, 2009.
- [4] Manshor, N., Abdul Rahiman, A.R., Raja Mahmood, R.A., Fusion of global shape and local features using meta-classifier framework, (2013) *International Review on Computers and Software (IRECOS)*, 8 (9), pp. 2113-2117.
- [5] Hegazy, D. and Denzler, J. Generic object recognition using boosted combined features. *Proceedings of the 2nd international conference on Robot vision (RobVis'08)*, Auckland, New Zealand. 355-366, 2008.
- [6] Oliveira, L. and Nunes, U. On integration of features and classifiers for robust vehicle detection. *Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems*, Oct. 12-15, Beijing, China. 414-419, 2008.
- [7] Veltkamp, R. C. and Tanase, M. A survey of content-based image retrieval systems. *Content-Based Image and Video Retrieval: Multimedia Systems and Applications Series*. 47-101, 2002.
- [8] Mangai, U.G. Samanta, S. Das, S. and Chowdhury, P.R. A survey of decision fusion and feature fusion strategies for pattern classification. *IETE Technical Review*. 27(4): 293-307, 2010.
- [9] Faundez-Zanuy, M. Data fusion at different levels. *Multimodal Signals: Cognitive and Algorithmic Issues, Lecture Notes in Computer Science*. 5398. 94-103, 2009.
- [10] Antenreiter, M. Ortner, R. and Auer, P. Combining classifiers for improved multilabel image classification. *Proceedings of the ECML/PKDD 2009 Workshop on Learning from Multi-Label Data (MLD'09)*, Sept. 7-11, Bled, Slovenia, 2009.
- [11] Hatami, N. and Ebrahimpour, R. Combining multiple classifiers: Diversify with boosting and combining by stacking. *International Journal of Computer Science and Network Security (IJCSNS)*. 7(1): 127-131, 2007.
- [12] Jeong, J. Hwang, C. and Jeon, B. An efficient method of image identification by combining image features. *Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication (ICUIMC'09)*, Jan. 15-16, Suwon, Korea. 607-611, 2009.
- [13] Kittler, J. Hatef, M. Duin, R.P.W. and Matas, J. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 20(3): 226-239, 1998.
- [14] Kuncheva, L. Combining pattern classifiers: Methods and algorithms. *Wiley-Interscience Publication*, 2004.
- [15] Polikar, R. Ensemble based system in decision making, *IEEE Circuit and Systems Magazine*, Third Quarter, 2006.
- [16] Oza, N.C. and Tumer, K. Classifier ensembles: Select real-world applications. *Information Fusion*. 9(1): 4-20, 2008.
- [17] Zhang, D. and Lu, G. A comparative study on shape retrieval using Fourier descriptors with different shape signatures. *Journal of Visual Communication and Image Representation*. 14 (1): 41-60, 2003.
- [18] Eakins, J. P. Towards intelligent image retrieval, *Journal Pattern Recognition*. 35(1): 3-14, 2002.
- [19] Kuettel, D., Guillaumin, M., and Ferrari, V. Combining Image-Level and Segment-Level Models for Automatic Annotation. *Advances in Multimedia Modeling*, 16-28, 2012.
- [20] Sonka, M. Hlavac, V. and Boyle, R. Image processing, analysis, and machine vision. *PWS Publishing*, 1999.
- [21] Campilho, A., Kamel, M., Han, D., Li, W., Lu, X., Wang, T., and Wang, Y. Automatic Segmentation Based on AdaBoost Learning and Graph-Cuts. *Image Analysis and Recognition*, 4141: 215-225, 2006.
- [22] Chen, J. J., Su, C. R., Grimson, W. L., Liu, J. L., and Shiu, D. H. Object Segmentation of Database Images by Dual Multiscale Morphological Reconstructions and Retrieval Applications. *IEEE Transactions on Image Processing*. 21(2): 828-843, 2012.
- [23] Brox, T., Bourdev, L., Maji, S., and Malik, J. Object segmentation by alignment of poselet activations to image contours. *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2225-2232, 2011.
- [24] Carreira, J. and Sminchisescu C. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE Transaction on Pattern Analysis and Machine Intelligence*. 34(7), 1312-1328, 2012.
- [25] González, E. Adán, A. Feliú, V. and Sánchez, L. Active object recognition based on Fourier descriptors clustering. *Pattern Recognition Letters*. 29(8): 1060-1071, 2008.
- [26] Sarfraz, M. Object recognition using Fourier descriptors: some experiments and observations. *Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation (CGIV)*, July 26-28, Sydney, Australia. 281-286, 2006.
- [27] Soldea, O. Unel, M. and Ercil, A. Recursive computation of moments of 2D objects represented by elliptic Fourier descriptors. *Pattern Recognition Letters*. 31(11): 1428-1436, 2010.
- [28] Jeong, Y. and Radke, J.R. Reslicing axially sampled 3d shapes using elliptic Fourier descriptors. *Medical Image Analysis*. 11(2): 197-206, 2007.
- [29] Reig-Bolaño, R. Martí-Puig, P. Rodriguez, S. Bajo, J. Parisi-Baradad, V. and Lombarte, A. Otoliths identifiers using image contours EFD. *Distributed Computing and Artificial Intelligence*. 79: 9-16, 2010.
- [30] Carlo, J. M., Barbeitos, M. S., & Lasker, H. R. Quantifying Complex Shapes: Elliptical Fourier Analysis of Octocoral Sclerites. *The Biological Bulletin*, 220(3), 224-237, 2011.
- [31] Yuan, R. and Hui, W. Object identification and recognition using multiple contours based moment invariants. *Proceedings of the International Symposium on Information Science and Engineering (ISISE '08)*, Dec. 20-22, Shanghai, China. 140-144, 2008.
- [32] Nabatchian, A. Abdel-Raheem, E. and Ahmadi, M. Human face recognition using different moment invariants: A comparative

study. *Proceedings of the 2008 Congress on Image and Signal Processing (CISP '08)*, May 27-30, Sanya, Hainan, China. 661-666, 2008.

- [33] Hu, M.K. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*. 8: 179-187, 1962.
- [34] Chen, Q. Evaluation of OCR algorithms for images with different spatial resolutions and noises. *Masters Abstracts International*, University of Ottawa, 2004.
- [35] Mikolajczyk, K. and Schmid, C. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 27(10): 1615-1630, 2005.
- [36] Vijayarajan, V., Dinakaran, M., Feature based image retrieval using fused sift and surf features, (2013) *International Review on Computers and Software (IRECOS)*, 8 (10), pp. 2500-2506.
- [37] Csurka, G. Dance, C. Fan, L. Willamowski, J. and Bray, C. Visual categorization with bags of keypoints. *Proceeding of the Pattern Recognition and Machine Learning in Computer Vision Workshop*. Grenoble, France. 1-22, 2004
- [38] Freund, Y. and R.E. Schapire. Experiments with a new boosting algorithm. *Morgan Kaufmann Publishers, Inc.* 1996.
- [39] Bennett, P. N. Building reliable metaclassifiers for text learning. *PhD thesis*, Carnegie Mellon University, 2006
- [40] Venkatesh Kumar, S., Karthikeyan, T., Hyperspectral Image Classification Using Multiple Kernel Learning SVM with FA-KLD-LFDA for Multi Feature Selection, (2014) *International Review on Computers and Software (IRECOS)*, 9(8), pp. 1338-1347.
doi:<http://dx.doi.org/10.15866/irecos.v9i8.2598>

Authors' information



Noridayu Manshor received the B.S degree in Computer Science from Universiti Putra Malaysia (UPM) in 2000, the M.Sc in 2004 from Universiti Teknologi Malaysia (UTM), and Ph.D degree in Computer Science at Universiti Sains Malaysia (USM), Malaysia in 2013.

Currently, she is a senior lecturer at Faculty of Computer Science and Information Technology, UPM. Her current research interests include image processing, computer vision and pattern recognition.



Amir Rizaan Abdul Rahiman received the B.S degree in Computer Science from University Putra Malaysia (UPM) in 2000 and the M.Sc and Ph. D degrees in Computer Science from University Teknologi Malaysia (UTM), and University Sains Malaysia (USM), Malaysia in 2004 and 2011, respectively. Currently, he is senior lecturer at Faculty of Computer Science

and Information Technology, UPM. His current research interests include multimedia applications, e-learning solution, flash-based storage systems, and multimedia storage systems.



Raja Azlina Raja Mahmood received the B.S degree in Computer Science from University of Michigan (Ann-Arbor), Michigan USA in 1996, the M.Sc 1999 from Universiti Teknologi Malaysia (UTM), and currently pursuing Ph.D degree in Computer Science at Monash University, Australia. Currently, she is lecturer at Faculty of Computer Science and Information

Technology, UPM. Her current research interests include computer networking, wireless and mobile networks.

Exploration of Heterogeneous Resources in Embedded Systems

Aissam Berrahou, Nassim Sefrioui, Ouafaa Diouri, Mohsine Eleuldj

Abstract – In this article we will address the multi-objective exploration problem of heterogeneous resources in embedded systems. The solution searched is to minimize the cost of execution and communication based on the following constraints: tasks with precedence constraints, deterministic scheduling (the execution time of each task is known), communication model, distributed platform with uniform cores, load balancing, size of memories associated... For this, we used linear programming with boolean variables communally know by BIP (Binary Integer Programming) as a method of multi-objective optimization to find the optimal solutions. The choice of such a method is motivated by the fact that it allows the realization of solutions called high-performance Pareto. Although this problem is NP-complete, this method provides optimal solutions in reasonable computation time. It determines in a single execution an optimal solution and this even if the problems are not convex. **Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords: Multi-Objective Optimization, Embedded System, Mapping, Scheduling, Accurate Method, Binary Integer Programming

Nomenclature

GP	Platform model
R	Set of resources
L	Set of links
GA	Application model
T	Set of tasks
E	Set of data (represent the precedence constraints between the tasks)
$P(r_i, r_j)$	Path of r_i to r_j
TC_{IN}	Communication cost
BP	Bandwidth
LT	Network latency
A	Adjacency matrix that denotes the allocation of tasks to resources
B	Adjacency matrix that expresses the dependence of the tasks towards the data
M	Adjacency matrix that expresses the mapping of data to paths
D	Adjacency matrix that expresses the relationship between paths and resources
F	Adjacency matrix that denotes the relationship between all paths resulting from a given routing algorithm and the links which make up these trajectories
Cap	Adjacency matrix that denotes if the cores of the platform is available to execute the tasks of the application
$Type$	Adjacency matrix that describes if the resources of the platform are core or non-core
$T1$	Adjacency matrix that expresses the completion time of all tasks on all available cores

$T2$	Column vector, which denotes the time dedicated for the execution of a task by his landlord
$S1$	that expresses the storage capacity of the memories associated with all resources of core type
$S2$	Column vector that expresses the storage size required for each task
Map	Adjacency matrix that designates the initial mapping of tasks to resources of core type
Par	Adjacency matrix that denotes the tasks should be executed in parallel. The algorithm used to obtain this matrix is the breadth First Search (BFS) algorithm

I. Introduction

Today, the evolution of the computer science is such that the majority of systems are parallel and embedded.

Since electronic accessories (mobile phones, GPS, etc) to the great centers of scientific computing, the systems we use are composed of several processing units. Parallel architectures is so pervasive that the SISD (Single Instruction Single Data) flow execution are almost an exception.

The problem of effective exploitation of these resources then arose. The main problem of such a system is to judiciously assign calculations to be made on the different execution resources to optimize several objectives. The scheduling theory is the branch of operational research that considers this type of optimization problem.

A typical scheduling problem is to choose for each task to be processed, the execution resource where the task will execute and on what date.

The scheduling problems are numerous, according to a set of parameters and constraints that affect the tasks (such as different time calculation, precedence relationships ...), according to the execution platform (homogeneous, heterogeneous, dedicated, ...) and according to the performance index that is to be optimized (eg, the total completion time: makespan [3], the communication time between tasks IPC: Inter-Process-Communication [2], the power consumption, the implementation surface, etc).

Unfortunately, most of these scheduling problems are NP-complete. It is therefore generally not easy to obtain in reasonable time an optimal scheduling. And that, for the reason that these objectives depend on a multitude of constraints that related with:

- The number of tasks and resources;
- The capacity of the associated memories;
- The parallelism;
- The topology of interconnection network: type (regular or irregular), pines (unidirectional or bidirectional)
- The routing algorithm;
- The capacity of the communication channels;
- The type of energy source used in the diet of the platform (battery, renewable energy ...) / consumption (low, medium, high), static or dynamic voltage.
- The load balancing;
- The fault tolerance.

In addition, these objectives are often contradictory, for example, to minimize the IPC, we can think affect the entire tasks graph into a single node of the resources graph, but this will result in a load imbalance, because the task graphs have different sizes in terms of number of nodes. Another example is the minimization of the execution cost, of course we can think of assigning tasks to the most powerful cores. However, this will result in a higher energy consumption, or even a higher execution cost if the size of memory associated with these cores is very low compared to that of cores which are not powerful, and therefore if we bear in mind this constraint (memory size) we can find an optimal combination best in comparison with the one we originally thought.

Indeed, the literature is rich with various resource allocation algorithms, we cite as an example the algorithms based on heuristics algorithms [3] [14] [15] [17], based on graph theory [6] [18], the load balancing algorithms [10] [16], genetic algorithms [1][8], etc.

However, these solutions do not take into account some very important constraints to improve the total execution time, such as:

- The size of the memory associated with each core should be sufficient;
- The parallel tasks should be mapped to different cores to improve the total runtime of the application.

In addition, they take account of the objectives at the expense of others. To properly study the system and be

able to find an optimal solution to the mapping of multi-objective problem, we must, first of all model the problem formally. For this, we used linear programming with Boolean variables communally know by BIP (Binary Integer Programming) [7] as a method of multi-objective optimization to find the optimal solutions. The choice of such a method is motivated by the fact that it allows the realization of solutions called high-performance Pareto. They allow to determine in a single execution an optimal solution and this even if the problems are not convex.

The rest of this article will be divided into three parts:

- The first part will be dedicated to introducing some preliminary concepts that we will use to model the optimization problem of mapping multi-objective;
- The second part presents the modeling of the optimization problem of mapping based on two objectives: minimization of the communication cost, and minimization of total execution cost with the taking into account of the load balancing constraint, the size of memories associated constraint, the parallel tasks constraint, etc;
- The third part consists of the presentation of simulation results obtained by the F4MS simulator (Framework for Mixed Systems) [4] which are based on a multi-objective solver named Gurobi [12].

II. Model of Embedded System

We consider a mixed system consisting of hosts and an interconnection network. Each host is fitted with a core (programmable or not) with their own resources, including a memory of limited size. The cores of different hosts are heterogeneous and can communicate through an interconnection network. A distributed application is composed of heterogeneous tasks, logical or physical, that can run on host synchronously or asynchronously. Each task must be assigned to a single core but it can execute several tasks within the limits of the memory it has. Some tasks need to communicate information through the network if they have not been assigned to the same host.

Each assignment of all tasks to host corresponds a runtime cost (total cost of tasks execution on heterogeneous hosts to which they are assigned), a communication cost (total cost of communication between pairs of unassigned tasks to the same processor and should be exchange information) and a cost of energy consumption.

II.1. Platform Model

The platform of execution considered in our study is heterogeneous: the different nodes put different times to accomplish the same task. In our model, we assume that are know the following information:

- Available tasks on each node (Potentially executable)
- The execution time of the tasks available on each node. As the platform is heterogeneous there is no

relationship between the execution time of the various nodes for the same task.

A platform model (see Fig1) is modeled by the graph $GP=(R,L)$:

- $R = \{r_1, \dots, r_k\}$: The nodes of GP, they represent the resources of the platform: switches, routers, memories, processing units or even host. These latter are fitted with the cores (specific or programmable) having its own resources, including a memory of limited size. The cores of different hosts are heterogeneous and can communicate through a communication model.
- $L = \{l_1, \dots, l_m\}$: The edges of GP, they represent the communication links between hosts and physically are the pins of an interconnection network.

In addition, each node r_j of R is able to execute a subset of the tasks that the application needs. Let T_i be the subset of tasks that the node r_j is able of achieving and $T = \bigcup_i T_i$ a set of the tasks available on the platform,

with $1 \leq i \leq n$ and $r_j \in R$. In addition, each communication link $l_k(r_a, r_b)$ has a bandwidth $Bw(r_a, r_b)$ that is the size of data that can be transferred through a communication link per unit of time.

The Fig. 1, illustrates a graphical representation of an example of possible allocation of four tasks of an application A on four cores of a platform P based on a 2×2 Mesh topology.

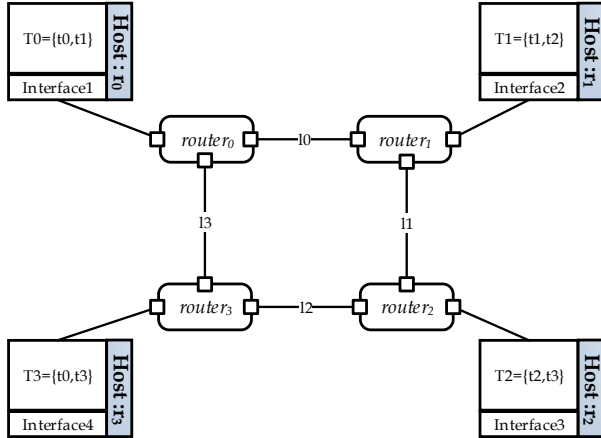


Fig. 1. Example of a platform with all possibilities for mapping four tasks

Thus, the Table I below represents the values correspond to the time required to execute the four tasks of the Application A on the four cores of the platform P .

When a cell of the table contains no value, it means that the core is not able to execute this task for the reason that we have for example physical tasks and we want to assign on programmable core, or for the reason the memory associated with a core is not sufficient to store these tasks at run-time, etc. For example, the core r_3 is able to execute the tasks t_0 and t_3 but is not able to execute the tasks t_1 and t_2 .

TABLE I
THE EXECUTION TIMES OF THE FOUR TASKS OF THE APPLICATION A
BY THE FOUR CORES OF THE PLATFORM P

	Core	r0	r1	r2	r3
Task	t0	35	-	-	15
	t1	10	10	-	-
	t2	-	10	10	-
	t3	-	-	10	10

II.2. Communication Model

As we mentioned above, hosts can either be connected by point-to-point communication links (unidirectional and / or bidirectional) or a complex interconnection network based on a specific topology (regular or irregular). In the literature, several models of communication exist, such as 1-port model or multi-port model. We chose to use the 1-port model due to its ease of modeling and implementation while remaining realistic. In this model, only one data can be sent simultaneously through a communications link and a node can realize at most one reception and one emission.

We define $P(r_i, r_j) = \bigcup l_{h,h'}$ with $l_{h,h'} \in L$ as the path of r_i to r_j .

II.3. Application Model

The model of platform defined previously can execute several times an application modeled under the shape of a graph with different games of entry. Each of these tasks must be executed according to constraints of order and/or priority. We suppose that all the input data is ready at the beginning and the same application is applied to this data set. Thus, we must schedule a workflow consists of instances of the same application on the platform with the model described above. Also there is no deadline (deadline) for tasks. In this study we focus only on the execution of a single workflow.

The application model (see Fig. 2) is modeled by the oriented graph $GA = (T, E)$, Where, $T = \{t_1, \dots, t_n\}$ is a set of n tasks and $E \subset T \times T$ represent the precedence constraints between the tasks. If a dependence constraint (t_i, t_j) at the level of the application exists between the task t_i and the task t_j it is translated by a sending of a data e_k of the task t_i to the task t_j for that the latter can run.

Each node of the graph is characterized by:

- The Complexity: defines the task structure which can express either a monolithic component, or a set of parallel components or a composite component.
- The size: defines the size of task expressed in cycle for each computing unit class.
- The data: The quantity of input and output data, and when they are produced.
- The communication: represents the interaction between the tasks of the same graph.
- The chronological parameters: that denotes delays and chronometric parameters wich indicate dates, such as:

- The date of awakening, represents the time of triggering of the first execution request;
- Maximum execution time when it has the processor itself, also called WCET (Worst Case Execution Time)[9];
- The critical period (deadline), represent the maximum acceptable time for its execution;
- The period when it comes to a periodic task;
- The deadline: when it comes to a task to strict constraints, otherwise is a date its overcoming causes a temporal fault.

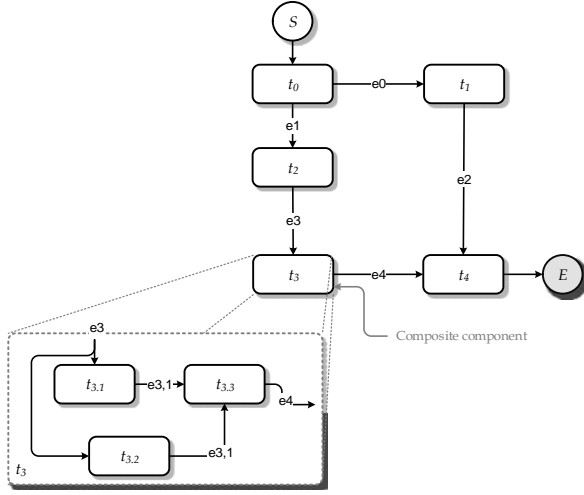


Fig. 2. Example of an application composed of four tasks

III. Problem Formulation

III.1. Minimization of the Communication Cost

The minimization of the communications cost between the cores with significant consequences that can improve the optimization of other objectives such as the minimization of energy consumption, total cost of execution and the implementation area. The cost of communication depends on four constraints:

- The communication model
- The topology of interconnection network (irregular or regular)
- The size of the messages exchanged
- Number of cores implicated in the communication.

For example, the communication cost in the case of communication topologies regularly found in the parallel applications:

- Unidimensional: frequent in the scientific computing problems based on grids. Cores exchange messages with their two neighbors (north and south)
- Ring: Each processor receives data from its left neighbor and sending to the right.
- Tree: often used in applications that have global operations.

The communication cost can be expressed as follows:

$$TC_{IN}(r_i, r_{i+1}, e_k) = LT + \frac{e_k}{B_w} \quad (1)$$

where r_i and r_{i+1} are two calculation resources on which the tasks t_i and t_k are mapped respectively to communicate, IN is the topology of the communication platform, e_k is the message size (in bytes) transmitted on one cycle of execution, LT is network latency and Bw is bandwidth.

In the case of the Ring topology, the cores are first of their left neighbor then send the right. The cost of communication is additive, it can be formulated as follows:

$$TC_{Ring}(r_i, r_j, e_k) = \begin{cases} \sum_{h=i}^{j-1} TC_{Ring}(r_h, r_{h+1}, e_k), \\ r_{h+1} \text{ is the successor of } r_h, \text{ if } r_i \neq r_j \\ 0, \text{ else} \end{cases} \quad (2)$$

For the algebraic formulation of this problem we will need:

Matrix A:

$$A_{ij} = \begin{cases} 1, \text{ if the task } t_j \in T \text{ is mapped to resource } r_i \in R \\ 0, \text{ else} \end{cases} \quad (3)$$

Adjacency matrix of size $|R| \times |T|$, which expresses the mapping of tasks on resources. It constitutes the first set of our decision variables:

$$A = \begin{pmatrix} x_{0,0} & \dots & x_{0,T-1} \\ \vdots & \ddots & \vdots \\ x_{R-1,0} & \dots & x_{R-1,T-1} \end{pmatrix} \quad (4)$$

For example the allocation matrix obtained from the example shown in the Fig. 1 is as follows:

$$A = \begin{pmatrix} x_{0,0} & x_{0,1} & 0 & 0 \\ 0 & x_{1,1} & x_{1,2} & 0 \\ 0 & 0 & x_{2,2} & x_{2,3} \\ x_{3,0} & 0 & 0 & x_{3,3} \end{pmatrix}$$

Matrix B:

$$B_{ij} = \begin{cases} 1, \text{ if } \exists t_k, e_j = (t_i, t_k) \in E \\ -1, \text{ if } \exists t_k, e_j = (t_k, t_i) \in E \\ 0, \text{ else} \end{cases} \quad (5)$$

Adjacency matrix of size $|T| \times |E|$, which expresses the dependence of the tasks towards the data.

For example the dependence Task/data matrix obtained from the example shown in the Figure 2 is as follows:

$$B = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & -1 & 0 & -1 \end{pmatrix}$$

Matrix M :

$$M_{ij} = \begin{cases} 1, \text{if the data } e_j \in E \text{ is assigned} \\ \text{to the path } p_i \in P \\ 0, \text{else} \end{cases} \quad (6)$$

Adjacency matrix of size $|P| \times |E|$, that expresses the data mapping paths to arrive at their destinations. It depends on the mapping of tasks to resources.

It constitutes the second set of our decision variables.

$$M = \begin{pmatrix} y_{0,0} & \cdots & y_{0,E-1} \\ \vdots & \ddots & \vdots \\ y_{P-1,0} & \cdots & y_{P-1,E-1} \end{pmatrix} \quad (7)$$

Matrix D :

$$D_{ij} = \begin{cases} 1, \text{if } source(p_j) = r_i \\ -1, \text{if } sink(p_j) = r_i \\ 0, \text{else} \end{cases} \quad (8)$$

Adjacency matrix of size $|R| \times |P|$, that expresses the relationship between paths and resources. For given a routing algorithm and a platform model

Matrix F :

$$F_{ij} = \begin{cases} 1, \text{if the link } l_j \in p_i \\ 0, \text{else} \end{cases} \quad (9)$$

Adjacency matrix of size $|P| \times |L|$, that denotes the relationship between all paths resulting from a given routing algorithm and the links which make up these trajectories.

Constraint 1: related to routing

We have derived the following linear equation which constrains the mapping of a task and the communication link with the other. This constraint comes from the routing algorithm implemented in interconnection networks:

$$D \times M = A \times B \quad (10)$$

Constraint 2: related to the mapping of tasks on the resources

The second constraint is to ensure that each task is assigned to exactly a single core.

However we can regroup several tasks on a single resource:

$$A \times 1_{|T|} = 1_{|R|} \quad (11)$$

Such as, $1_{|R|}$ is a unitary column vector of size R with all these elements are 1 and that for note $A^{RT} = (A^{TR})^T$.

Constraint 3: related to the data mapping path:

$$M \times 1_{|P|} \leq 1_{|E|} \quad (12)$$

A data may be assigned to at most one path.

Constraint 4: related to the capacity of the bandwidth

$$F^T \times M \times E \leq Bw \quad (13)$$

where DP is the bandwidth and E is a row vector of size $|E|$ that represents the applications data.

The total demand for bandwidth on a path must not exceed its capacity.

Objective function:

Total traffic on the links can be calculated as the sum of all requests e_i on the path links that arise in a given mapping:

$$\min : E^T \times M^T \times F \times 1_{|E|} \quad (14)$$

III.2. Minimization of the Runtime Cost

To realize this objective, we must seek a compromise between the following three constraints:

- The frequency of cores: we must try to place tasks those requiring high runtime on cores with high clock frequencies;
- The size of the memory associated with each core must be sufficient;
- Parallel tasks should be mapped on different cores to improve the total runtime of the application.

For the algebraic formulation of this problem we will need:

Matrix Cap :

$$\begin{aligned} Cap_{ij} &= \\ &= \begin{cases} 1, \text{if the host } c_j \in C \text{ is able} \\ \text{to execute the task } t_i \in T \\ 0, \text{else} \end{cases} \end{aligned} \quad (15)$$

Adjacency matrix of size $|T| \times |C|$, which denotes if a resource j of core type (reconfigurable, signal processing nonprogrammable, etc) is able to perform a task i .

Matrix Type :

$$Type_{ij} = \begin{cases} 1, & \text{if the resource } r_i \in R \text{ is of core type } c_j \in C \\ 0, & \text{else} \end{cases} \quad (16)$$

Adjacency matrix of size $|R| \times |C|$, that describes if resources of platform are core or non-core.

Matrix T1 :

$$T1_{ij} = \begin{cases} \text{Completion time of } t_i \text{ on host } c_j, & \text{if } Cap_{ij}=1 \\ 0, & \text{else} \end{cases} \quad (17)$$

Matrix of size $|T| \times |C|$, which expresses the completion time of all tasks on all available cores. The values of this matrix are obtained by dividing the theoretical complexity of the tasks by the clock frequency of cores.

Vector S1 :

$$S1_i = \text{storage capacity of the memory associated with the core } i \quad (18)$$

Column vector of size $|C|$, that expresses the storage capacity of the memories associated with all resources of core type.

Vector S2 :

$$S2_i = \text{storage size required for the task } i \quad (19)$$

Column vector of size $|T|$, that expresses the storage size required for each task.

Matrix Map :

$$Map = A^T \times Type \quad (20)$$

Adjacency matrix of size $|T| \times |C|$, which designates the initial mapping of tasks to resources of core type.

Matrix Par :

$$Par_{ij} = \begin{cases} 1, & \text{if the task } t_i \text{ should be executed in parallel with the task } t_j \\ 0, & \text{else} \end{cases} \quad (21)$$

Adjacency matrix of size $|T| \times |T|$, denoting the tasks should be executed in parallel.

The algorithm used to obtain this matrix is the breadth First Search (BFS) algorithm [11] (see Fig. 3).

Vector T2 :

$$T2 = (Map \cdot Cap) 1_{|C|} \quad (22)$$

Column vector of size $|T|$, which denotes the time dedicated for the execution of a task by his landlord.

Constraint 5: Execution capacity

All tasks must be mapped to cores that are able to execute them. The equation expressing this constraint is:

$$T2 \times H = Cap \quad (23)$$

Such that $H = A^T \times Type$

Constraint 6: storage size

The tasks should be mapped on the cores having a sufficient storage size:

$$Map \times S1 \leq S2 \quad (24)$$

Constraint 7: Parallel Tasks

Parallel tasks should be mapped (maximum) on different cores:

$$1_{|T|}^T \times Par \geq 1_{|C|} \quad (25)$$

Objective-function 2: minimize run-time

$$Min : (T2) \quad (26)$$

However, it should be noted that our minimization problem is difficult to solve the optimality. In this context we used the method of Lagrangian relaxation [13] to calculate a lower bound.

Bound thus provided gives a guarantee on the quality of the solution.

The idea is to penalize the hard constraints in the objective function and give a lower bound of the optimal solution.

In general this bound is close to the optimum.

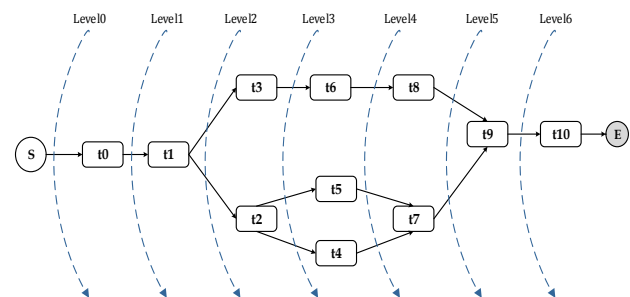


Fig. 3. The levels of a task graph

IV. Case Study

The case study that we propose in this paper is offline mapping of a task graph (see Fig. 5), a system composed of 8 logical tasks and 4 specific tasks, a heterogeneous runtime platform based on (see Fig. 4):

- Interconnection network : Mesh 2×3 ;
- X-Y routing;
- Two generic cores, two DSP, FPGA, and ASIC.

After the design phase of the application and platform model using the CoMMix profile [5], Optimizer F4MS [4] gives the following result in the form of mapping chart (see Fig. 6) and Gantt diagram (see Fig. 7) to represent scheduling operation, in a reasonable time (27 seconds) compared to the exhaustive method which requires 6^{12} combinations.

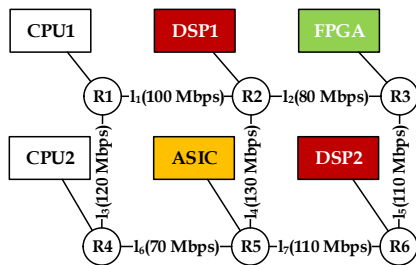


Fig. 4. Platform model

TABLE II
CHARACTERISTICS OF PLATFORM RESOURCES

Cores	Frequency (Ghz)	Memory size (MB)
CPU1	2.1	30
CPU2	2.3	40
DSP1	3	100
DSP2	3.3	100
ASIC	4	50
FPGA	6	300

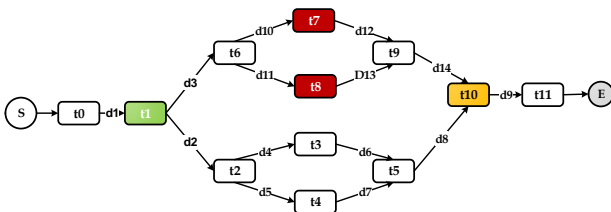


Fig. 5. Application Model

TABLE III
CHARACTERISTICS OF APPLICATION TASKS

Tasks	Type	Theoretical complexity (10^9 s)	Size (MB)
t0	Generic	504	2.5
t1	Reconfigurable	3808	4
t2	Generic	400	1
t3	Generic	280	1
t4	Generic	330	0.5
t5	Generic	440	1.2
t6	Generic	730	1.5
t7	Signal processing	2330	3.2
t8	Signal processing	1250	3
t9	Generic	630	1.3
t10	Physical	7330	4.5
t11	Physical	277	2.5

TABLE IV
RUNTIME ORDER AND THE TRANSFER OF APPLICATION TASKS

Data	Type	Size (MB)
d1	t0 → t1	1
d2	t1 → t2	34
d3	t1 → t6	28
d4	t2 → t3	58
d5	t2 → t4	28
d6	t3 → t5	18
d7	t4 → t6	65
d8	t5 → t10	34
d9	t10 → t11	28
d10	t6 → t7	38
d11	t6 → t8	98
d12	t7 → t9	8
d13	t8 → t9	65
d14	t9 → t10	15

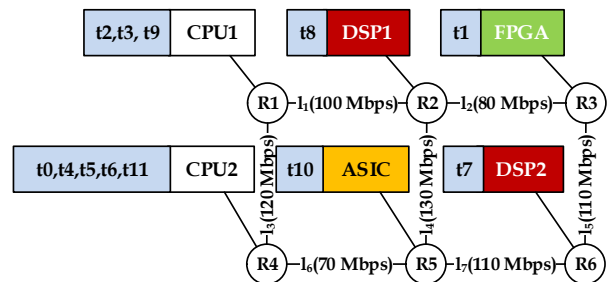


Fig. 6. Mapping chart

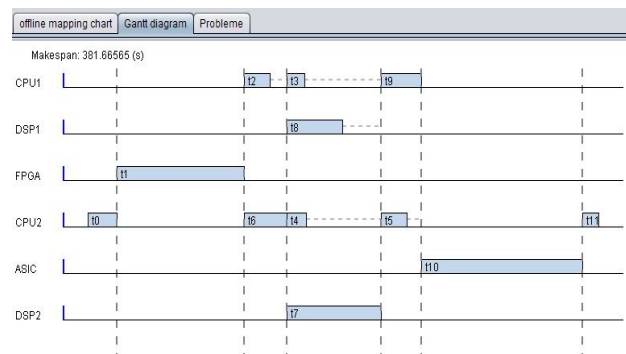


Fig. 7. Gantt diagram

TABLE V
TOTAL RUNTIME

Communication cost	Runtime cost	Total
14.643 s	381.665 s	396.308 s

V. Conclusion

In this paper we presented an accurate method of mapping multi-purpose an application to an execution platform for heterogeneous cores by minimizing the communication and execution time.

After having reviewed the various solutions proposed, we sought to take a new orientation for solving this problem. For that, and according to the research carried, we used linear programming with Boolean variables as a method of optimizing search for an optimal solution. As perspective, our method must take into account the mapping of composite tasks and non-deterministic tasks.

References

- [1] Rose, A.V.V., Ramachandran, R.S., Genetic algorithm based optimization of vertical links for efficient 3D NoC multicore crypto processor, (2013) *International Review on Computers and Software (IRECOS)*, 8 (5), pp. 1082-1090.
- [2] Aditya, H., Ravishankar, K., Kunal, T., Udi, W., *Minimum makespan scheduling with low rank processing times*, Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms (Page: 937 Year of Publication: 2013 ISBN: 978-1-611972-51-1).
- [3] Akturk, I., *ILP-Based Communication Reduction for Heterogeneous 3D Network-on-Chips*, 21st Euromicro International Conference on Parallel, Distributed and Network-Based Processing (Page: 514 Year of Publication: 2013 ISBN 978-1-4673-5321-2).
- [4] Berrahou, A. Raji, Y., Rafi.M, Eleuldj, M., *Framework For Mixed Systems*, 21th International Conference on Microelectronics (Page: 336 Year of Publication: 2009 ISBN 978-1-42445816-5).
- [5] Berrahou, A., Sefrioui, N., Diouri, O., Eleuldj, M., *CoMMix Profile for Modeling and Analysis Embedded Systems*, The 4th International Conference on Multimedia Computing and Systems (Page: 336 Year of Publication: 2014 ISBN 978-142445816-5).
- [6] Chien-chung, S., Wen-hsiang, T., *A Graph Matching Approach to Optimal Task Assignment in Distributed Computing Systems Using Minimax Criterion*, IEEE Transactions on Computers, (Page: 34 Year of Publication: 1985 ISSN 0018-9340).
- [7] Chinneck, J. W., *Practical optimization: A gentle introduction*, 2004.
<http://www.sce.carleton.ca/faculty/chinneck/po/Chapter13.pdf>.
- [8] E. Ilavarasan, P. Thambidurai, Genetic Algorithm for Task Scheduling on Distributed Heterogeneous Computing System, (2006) *International Review on Computers and Software (IRECOS)*, 1 (3), pp. 233-242.
- [9] F. Rammig, M. Ditzel, P. Janacik, T. Heimfarth, T. Kerstan, S. Oberthuer, K. Stahl, Basic concepts of real time operating Systems, W. Ecker, W. Müller and R. Dömer (Ed.), *Hardware-Dependent Software*, 2 (Netherlands: Springer; 2009, 15-45).
- [10] Gao, C., Liu, J. W. S., Railey, M., *Load Balancing Algorithms in Homogeneous Distributed Systems*, International Conference on Parallel Processing (Page: 302 Year of Publication: 1984).
- [11] J.Fournier, *Theorie des graphes et applications: Avec exercices et problèmes* (Hermes Science Publications).
- [12] Gurobi solver, <http://www.gurobi.com>
- [13] Marshall, L., Fisher, The Lagrangian Relaxation Method for Solving Integer Programming Problems, *Management Science*, Vol. 50, n. 12, pp. 1861-1871, 2004.
- [14] Perng-Yi, R., Edward Y. S. Lee, Masahiro, T., A Task Allocation Model for Distributed Computing Systems, IEEE Transactions on Computers (Page: 41 Year of Publication: 1984 ISSN 0018-9340).
- [15] Price C.C., Salama M. A. Scheduling of Precedence-Constrained Tasks on Multiprocessors, *The Computer Journal*, Vol. 33, n.3, pp. 219-229, 1990.
- [16] Rafael, A., Luis, L., *Sharing Jobs among Independently Owned Processors*, In Proceedings of the 8th International Conference Systems (Page: 282 Year of Publication: 1988).
- [17] Stuart, J., *Artificial Intelligence: A Modern Approach* (Prentice Hall/Pearson 2010).
- [18] Stone, H., Multiprocessor Scheduling with the aid of Network Flow Algorithms, *IEEE Transactions on Software Engineering*, Vol. 3, n.1, pp.85-93, 1977.

Authors' information



Aissam Berrahou received his DESA degree in Multimedia and computer network (2008). He is now a PhD student in the Mohammadia School of Engineers, Mohamed V University, Morocco. His current research area focuses upon the mixed systems, mapping optimization. His research interests in embedded systems, optimization, Component based-approach, MDA: Model Driven Architecture, other heuristics, and intelligence solutions on massively parallel and distributed architectures.



Sefrioui Nassim received the DESA degree in computer science, Telecom and Multimedia. He is now a PhD student in the Mohammadia School of Engineers, Mohamed V University, Morocco. His research focuses on interconnection networks on chip and parallel computer architectures, design methodologies for nanoscale systems on-chip, with a special interest on network-on-chip communication architectures.



Mohsine Eleuldj received the Ph.D. degree in computer science from the University of Montreal, Canada in 1989. He is currently Professor of Mohammadia School of Engineers, RABAT Morocco. His current research focuses on the development of design methodologies and software tools for the design of embedded systems.

Ouafaa Diouri received the D.E.S in computer science from the University of Paris XI, French in 1985. She is currently Professor of mohammadia engineering school, EMI RABAT Morocco. Her research interests include communication architectures, computer networking security and information systems security.

A Navigation-Aided Framework for 3D Map Views on Mobile Devices

Adamu Abubakar¹, Sadegh Ameri², Suhaimi Ibrahim³, Teddy Mantoro⁴

Abstract – The visualization of 3D objects on maps in mobile devices could enhance user perception of the objects on the map. However, in order to improve the interaction of the user with the mobile device that contains the 3D map for navigation, the path and location should be represented with a reasonable degree of accuracy and true to life. This paper utilizes a familiar concept of computational geometry that can be applied to the description of points and the path from one point to another: the Voronoi diagram conceptualizes the visualization of 3D objects on a 3D map in a mobile device used for navigation. The reason for considering Voronoi for this concept is that a system based on Voronoi was found to have a well-known geometric structure. This is similar to the fundamental construct of a navigation aid, which naturally evolves as a discrete set of points in moving from one location to another. The result of our concept could be implemented in the design of a device to aid navigation. **Copyright** © 2014 Praise Worthy Prize S.r.l. - All rights reserved.

Keywords: Voronoi, Mobile Device, 3D Objects, 3D Map, Navigation Aid

Nomenclature

rand	Random
\mathbb{R}^2	Square Of The Set Of Real Numbers
∞	Infinity
#	Any Number

I. Introduction

Google now provides indoor maps in association with its Google Maps application programming interface (API). This service is also available for mobile devices catering for indoor environments. The main idea behind the Google Indoor Map is to be able to point at a building and then zoom in and out of the indoor environment from floor to floor in order to search within the building. This is feasible with a degree of accuracy if the floor plan of the building is known to or used by Google Map. If not, Google Map requests the individual for a floor plan showing their building's layout, in order to add it to Google Maps, so that the indoor map of that building will be available through the Google Maps API for use in both mobile applications and websites.

The question which this research is trying to address is "can this be sufficiently accurate for use as a navigation aid?" especially on mobile devices. The major issue with navigation aids is to provide an easier means of finding an unfamiliar place. In the event where a location and the paths leading to it are miscalculated, this may even mean the difference between life and death.

3D objects on a 3D map will provide a realistic view which helps in recognizing a location of interest. Unfortunately, before recognizing the point of interest, individuals need to move in order to reach it.

As a result, pathways and location are the most essential attributes of applications for navigation, even when using a 3D map. Voronoi diagrams are proposed for determining paths and locations. The Voronoi diagram is a feature of computational geometry that uses points and paths, or node and edges, to uniquely define relationships between them [1].

There are many navigation-aided systems designed to suit mobile devices [2] for both pedestrian and in-car navigation systems, because mobile devices can also be used in a similar way to the dedicated devices uses for navigation. Most of the navigation systems on a mobile device platform with a 3D map do not normally use a 3D model [3]. Although for marketing reasons they mostly claimed to use 3D maps, the 3D representation of the environment is actually restricted or has no 3D components at all [4]. This kind of view is a 2D projection that gives an illusion of 3D [5]. Undoubtedly, creating a navigation aid with 3D maps for mobile devices is a complex task, but worth attempting. [6]-[7]. This is why a functional 3D map application capable of using a 3D dataset to display the most detailed description of a particular environment, with precise location and accurate paths, is important. This is the motivation of this research work. Mobile navigation systems available in the public domain rely on client/server remote rendering via wireless networks and assisted by a Global Positioning System (GPS). The client side is a 3D engine inside or installed on a mobile device, which is only functional when the device is equipped with an on-board GPS signal receiver.

Any expert in the field of 3D graphics will realize that navigation aids that claim to use 3D maps suggest that the intended interpretation of the 3D focuses on a 3D pictorial projection of the upcoming road layout that does not look

realistic or offer freedom to manipulate the viewpoint. This could strongly affect the location and path if the user relies on these systems as the only source of navigation. There are several techniques for establishing interactive positions and paths for navigation aids. Most research into the concept of navigation is for naval and air-space navigation, used by ships and planes respectively. However, navigation is within a general domain framework of spatial knowledge, whose development covers a three-stage discrete process consisting of landmark knowledge, route knowledge and survey knowledge [8]. This forms the theory popularly called navigation theory, which states “that humans first extract landmarks from an environment which are salient but static cues, orientation dependent, and also disconnected from one another” and consequently builds up what is known as a mental map [9].

Navigation theory allows researchers to consider the procedure for uncovering the state and behaviour of individuals “on-the-go”. One of the key aspects is landmark knowledge, which clearly indicates the levels of abstraction and understanding of the “knowledge about the presence of specific landmarks or scene instances in the environment regardless of their respective spatial characteristics” [8]-[9]. With this understanding it then leads to route knowledge, which is a transition between topological and landmark knowledge. Route knowledge is an evolving part of the individual’s ability to develop interconnected paths between landmarks [8], [10].

This will lead to sequential knowledge of “landmarks in an area coupled with navigation instructions which thus requires a collection of procedural as well as topological knowledge, that’s the knowledge of the relationships between locations and their links” [11]-[12]. Route knowledge evolves to a graphical organization of nodes and edges that is constantly growing as more nodes and edges are added [10], [13]-[15]. Based on the concept which this paper aims to establish, Voronoi diagrams will be used to organize nodes and edges which will also show the degrees of growth of each node and edge for use in the design of a navigation system that will use 3D maps for mobile devices. This paper consists of four further sections. Section II describes the Voronoi diagram for establishing paths and locations. Section III describes the conceptual establishment of path and location. Section IV provides the results, section V the conclusions of the work.

II. Establishing Path and Location

Voronoi diagrams are used to establish a navigation-aided framework to view 3D maps on mobile devices.

This can be implemented using various programming languages.

It requires building an algorithm and following each step for establishing points and regions within a space.

The implementation can be mostly in two-dimensional space, for which sets of points within a plane are divided into cells. Each cell contains exactly one region, as

shown in Fig. 1.

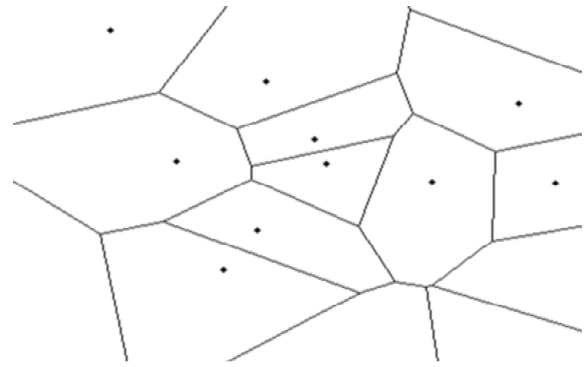


Fig. 1. Voronoi diagram within a certain space

The figure is generated with each cell containing a boundary which represents points and edges in a space. Points represent the location of objects and the edges are their boundaries. Although the number of nodes or points is arbitrary, it can be seen that they fall within an axis of x and y for which they are normalized to zero and one.

There are many spaces between the points, as depicted in the figure, because there are few nodes within each region. The relationship between the points in the space is such that each cell represents a Euclidean distance to the others and is bound to be smaller than the distance to others within the plane. This assertion, when it covers all points within the entire plane, leads to boundaries for cells and results in Voronoi edges.

This will result in the point’s equidistance from the nearest two sites. As a result, the point where multiple boundaries meet is the Voronoi vertex, which is theoretically equidistant from its three nearest sites. This theory is supported by a lemma which states that *the number of Voronoi vertices and edges are:*

$$2(n-1)-h \text{ and } 3(n-1)-h \quad (1)$$

This argument is true when n represents the number of spots within and h is considered as the number of spots on the convex hull of S . The proof of this is guided by considering the claim for a finite graph which satisfies Euler’s formula. Consider V , E and F to be the vertices, edges, and faces of the polyhedron (graph), then mapping members’ functions which can be projected on to a plane implies that a set of n points can be transformed into a finite graph; this is explained by presenting $2E = 3V$ for each side of the equality using Euler’s formula to give us Eq. (2):

$$V = 2(F-2) \text{ and } E = 3(F-2) \quad (2)$$

This function that produces $V = h + \#$ Voronoi vertices and $E = h + \#$ Voronoi edges. Therefore Voronoi vertices can be represented by $2(F-2)-h$ whereas its edges will be resolved in $3(F-2)-h$ based

on the fact that $F = n + 1$; such conditions will give rise to Eq. (3), which is the number of Voronoi vertices and edges:

$$2(n-1)-h \text{ and } 3(n-1)-h \quad (3)$$

This situation can be conceptualized in a design that will provide for a navigation aid as the platform for establishing the location and the number of interactions between each location in a well-defined pattern. This will lead to establishing that for a given space an application for a navigation aid should consider the points of interest as Voronoi vertices and the links to other points as Voronoi edges. Thus an increase in either an edge or a vertex leads to a corresponding increase in the other.

In order to support this, a second lemma is established. This lemma states that *the Voronoi diagram $V(S)$ has $O(n)$ many edges and vertices. The average number of edges in the boundary of a Voronoi region is less than 6.* In the case of a navigation aid, this implies that locations that can be used are such that the average number of edges in the boundary region should be less than 6. In order to prove this, the Euler formula for planar graphs is applied based on the established Voronoi diagram.

Hence, when V, E and F that is vertices, edges, faces respectively are connected, and putting these components into the Euler equation, we find Eq. (4):

$$V - e + f = 1 + c \quad (4)$$

Thus when this is applied to a finite Voronoi diagram, each vertex will contain at least three incident edges, and adding them together results in Eq. (5):

$$e \geq \frac{3v}{2} \quad (5)$$

Substituting this inequality with $c = 1$ and $f = n + 1$ yields Eq. (6):

$$v \leq 2n - 2 \text{ and } e \leq 3n - 3 \quad (6)$$

Furthermore, adding up the number of edges contained in the boundaries of all $n+1$ faces results in Eq. (7):

$$2e \leq 6n - 6 \quad (7)$$

This is because each edge is counted twice. Thus, the average number of edges in a region's boundary is bounded by a value less than six.

This condition supports a framework where a conceptualization can be well supported.

Since it has been proven that points and paths can be established and the number of these within a certain space is well defined, this generates a well established platform for the design of a navigation aid.

For establishing location and the number of interactions between each location a well-defined pattern

is required. This will lead to establishment of a given space in the navigation aid that should consider the points of interest as Voronoi vertices linked to the other points by Voronoi edges. The system starts by considering a given finite space with finite points or locations (see Fig. 2).



Fig. 2. Points within a region of Euclidean plane

Given a finite number (a set of two or more) of distinct points in a Euclidean plane, as shown in Fig. 2, there is an association of all locations in that space with the closest number(s) of the point set with respect to Euclidean distance.

The result is a tessellation of the plane into a set of the regions associated with members of the point set, as shown in Fig. 3. This is the planar ordinary Voronoi diagram generated by the point set, and the regions constituting the Voronoi diagram are ordinary Voronoi polygons. Since the points are the finite number n of points in the Euclidean plane, this will easily relate the pathfinding technique to the Voronoi metric space.

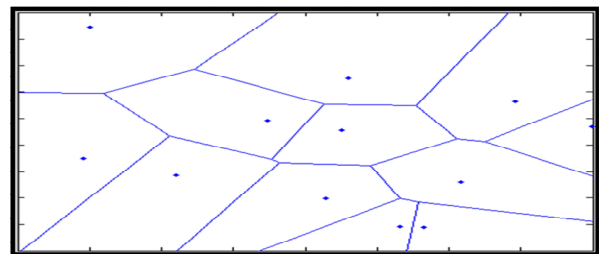


Fig. 3. Voronoi polygon (tessellation of points within a region)

Let $P = \{p_1, \dots, p_n\} \subset \mathbb{R}^2$ be the region given by:

$$2 < n < \infty \text{ and } x_i \neq x_j \text{ for } i \neq j, i, j \in I_n.$$

is:

$$V(p_i) = \{x \mid \|x - x_i\| \leq \|x - x_j\| \text{ for } j \neq i, i, j \in I_n\} \quad (8)$$

is called the planar ordinary Voronoi polygon associated with p_i or the Voronoi polygon of p_i and the set given by:

$$v = \{V(p_1), \dots, V(p_n)\} \quad (9)$$

is called the planar ordinary Voronoi diagram generated by P or the Voronoi diagram of P .

An application with a 3D map for use in mobile devices could implement this technique. In these applications, the approaches that will suit the implementation of this technique are when path and location are important. Obviously, in all navigation aids, path and location are the most important things to consider. The Voronoi diagram being generated is in a two-dimensional plane, which will be underneath a 3D-model file; its role will be to create a sub-division of the entire region of the 3D dataset into appropriate disjointed data points (nodes) to indicate the path differences between each node in the space. This means that the result of combining the two layers will be a separate layer which establishes the known points (nodes) and distance between the nodes (region). As a result these applications present a well-defined space.

III. Conceptualization of Establishment of Path and Location

Conceptualization is necessary (see Figure 4) in order to established path and position as the key aspects of a navigation aid for mobile devices that will use 3D map, and a Voronoi diagram that suits navigation theory has to be considered (see Fig. 5). Navigation theory will allow for an understanding of people's behaviour while they are moving, so that the Voronoi diagram must suit how it is applied. Navigation theory, as explained earlier, states that humans first extract landmarks from an environment.

These landmarks are salient but static cues, orientation-dependent, and also disconnected from one another. Accordingly, our study uses this concept to design the conceptual framework for the application.

IV. Results

The generated Voronoi diagram shown in Fig. 5 is the basis on which the landmark and route can be established in the application for navigation aid.

The figure is designed to provide an analogy of spatial data components and their relationship with navigation practices and computational geometric representation. It was generated through the Voronoi diagram technique in Matlab. Matlab has a function "voronoi.m" which is called by `voronoi(x, y)` in order to plot the Voronoi diagram for the given number of nodes. Each node is plotted randomly with two coordinates x and y within a metre square grid.

The code that generates Fig. 5 is as follows:

```
x = rand(1,90); y = rand(1,90); voronoi(x,y)
```

This code invokes the Matlab Voronoi generator to produce a Voronoi diagram in a 2D (x by y) metre square plane with ninety nodes. Due to the large number of nodes within the plane shown in Fig. 4, some nodes within neighbouring regions seems to be attached together; these nodes are as follows:

$y = 0.6 - 0.7$ } Two nodes are very close to each other
 $x = 0.3 - 0.4$ } within this grid
 $y = 0.6 - 0.7$ } Two nodes are very close to each other
 $x = 0.7 - 0.8$ } within this grid
 $y = 0.7 - 0.8$ } Two nodes are very close to each other
 $x = 0 - 0.1$ } within this grid

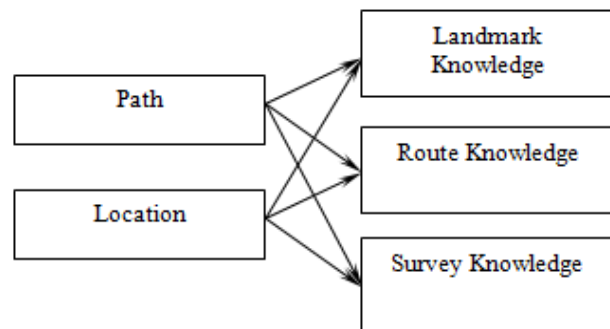


Fig. 4. Conceptual framework

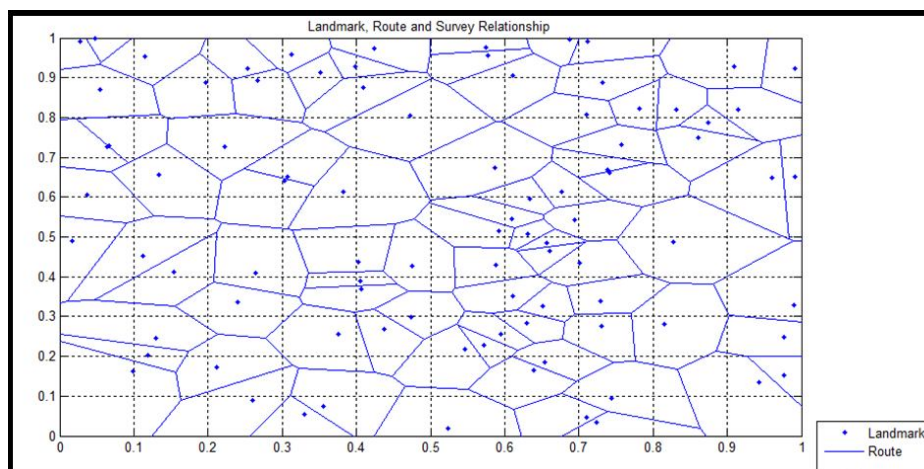


Fig. 5. Voronoi concept in landmarks (dots), routes (lines), and their connections

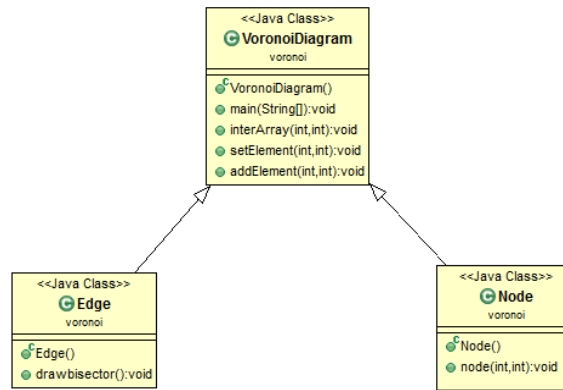


Fig. 6. Class diagram for the Voronoi diagram generation

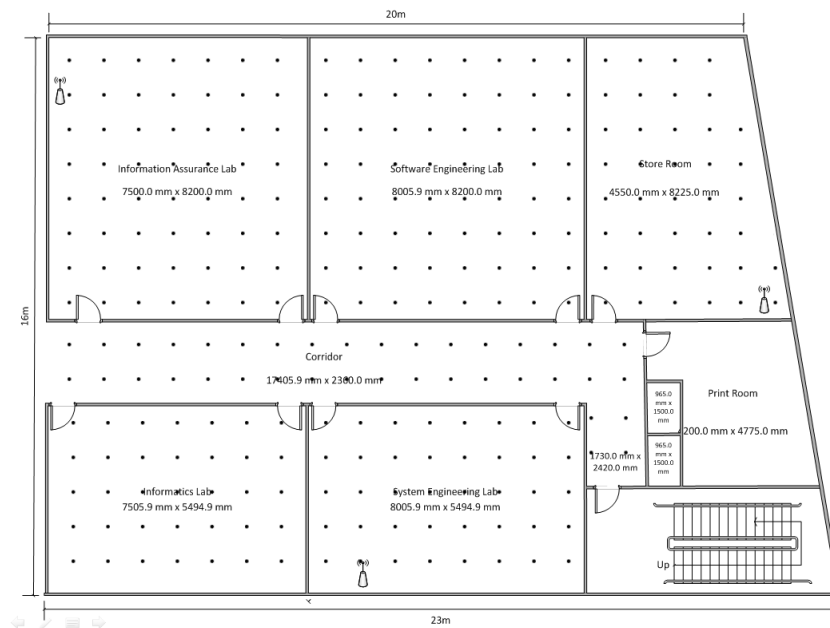


Fig. 7. Distributions of points within one floor of an indoor area

In terms of representation of landmarks, the above situation might not be reliable as nodes are so close to each other, and therefore the number of nodes has to be reduced. This clearly indicates a high-density populated area representation. A landmark in this research is considered as the representation of “city furnitures”, that is features that bring about recognizability such the position of ATM machines, trees, lamp-posts, bridges, buildings, etc.

At the layer in which the Voronoi technique is applied, a considerable number of different methods are used in different classes to generate the diagram [13]. The most important variables are the points or nodes and vertices or edges which generate the region. A node or point constitutes a place which stores the x and y coordinates of the location that it belongs to.

As a result, a node class will be responsible for storing all the nodes on the Voronoi diagram, as shown in Fig. 6.

For appropriate storage, an interArray method is provided to store all the points being generated for particular regions; when another node is added, this can

be updated by the addElement method. Such an element can be set with the setElement method. The lines joining and setting boundaries for nodes within the Voronoi diagram are the vertices or edges.

These are held in the vertex or Voronoi vertex class which stores the vertices or all boundaries for the entire region within the Voronoi diagram. In order to integrate the nodes and vertices, another class is required; this is the main class, also called the Voronoi diagram class.

The implementation of the algorithm considering the entire perimeter under consideration in the Voronoi diagram A is as follows:

Input = A, which is the set of all the n distinct nodes $A = \{a_1, a_2, a_3, \dots, a_n\}$ within the perimeter.

Output = A_{ij} as the decomposition of the entire perimeter into n number of regions connected by the their closest neighboring nodes.

Initialize the procedure with the random two nodes in set

A within the given perimeter.
while each node position established,
 do find the perpendicular bisector between
 the nodes,
 if more nodes are added from set $A \{a_n\}$
 then find the perpendicular bisector
 between the nodes
 else you are left with only two
 points and a Voronoi
 polygon could not be
 formed
 else update the triangulation
 edge indices and edge
 midpoints
 do find the vertex nearest the midpoints
 then update the triangulation edge indices and
 edge midpoints **until** the edges are exhausted
Initialize another random node to repeat the algorithm

The indoor area of the fourth floor of the Tun Abdul Razak building in UTM Jalan Samarak, Kuala Lumpur was used to show the feasibility of the proposed scheme (see Fig. 7). It is easy to draw the layout of a building with a distribution of points as in Figure 7. The most important part is that the selected points that will form a Voronoi diagram. Points very close to the edges of the building cannot generate a boundary as postulated in lemma 2. Therefore, using our concept for an application that will use Voronoi aided path and location of 3D objects in a mobile device for indoor navigation, only the central parts of the building would be suitable.

V. Conclusion

This paper provides the feasibility of establishing 3D objects on a 3D map in a mobile device's visualization with path and location precision. This is important because it could enhance user perception of the objects on the 3D Map. The reason for our assessment is that path and location on 3D maps inside a mobile device would improve user interaction with the device for navigation. The path and location were evaluated by using a Voronoi diagram to establish a node and edge in a space with a reasonable degree of accuracy; this will eventually ensure the precision of the ground-truth or real-life position of the 3D navigation application. This is because the Voronoi algorithm relies on the concept of computational geometry applied to the description of points and the paths between points

It can therefore be implemented on navigation systems on mobile devices to help people finding their way in unfamiliar places. The result of our concept could be implemented in the design of navigational aid devices in general.

References

- [1] A. Abubakar, T. Mantoro, M. A. Shafi'I, Dynamic interactive 3D

- mobile navigation aid, *Journal of Theoretical and Applied Information Technology*, Vol. 37, n. 2, pp. 159 – 170, 2012.
- [2] A. Abubakar, T. Mantoro, M. Mahmud, *Exploring end-user preferences of 3D mobile interactive navigation design*, Proceedings of the 9th International Conference on Advances in Mobile Computing and Multimedia (Page: 289 Year of Publication: 2011, ISBN: 978-1-4503-0785-7)
- [3] T.Mantoro, A. Abubakar, M. A. Ayu, *Multi-user navigation: A 3D mobile device interactive support*, Proceedings of IEEE Symposium on Industrial Electronics and Applications (Page: 545 - 549 Year of Publication: 2011, ISBN: 978-1-4577-1418-4).
- [4] A. Nurminen. Mobile 3D City Maps, *IEEE Computer Graphics and Applications*, Vol. 28, no. 4, pp. 20-31, 2008.
- [5] A. Nurminen, *m-LOMA-a mobile 3D city map*, Proceedings of the eleventh international conference on 3D web technology, (Page: 7 - 18 Year of Publication: 2006, ISBN:1-59593-336-0).
- [6] A. Oulasvirta, S. Estlander, A. Nurminen, Embodied interaction with a 3D versus 2D mobile map, *Personal and Ubiquitous Computing*, Vol. 13. No. 4, pp. 303-320, 2009.
- [7] T. Mantoro, A. Abubakar, and H. Chiroma, *Pedestrian position and pathway in the design of 3D mobile interactive navigation aid*, Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia, (Page: 189 - 198 Year of Publication: 2012, ISBN: 978-1-4503-1307-0).
- [8] R.P. Darken, H.Cevik, *Map usage in virtual environments: Orientation issues*, Proceedings of IEEE Virtual Reality (Page: 133 - 140 Year of Publication: 1999, ISBN: 0-7695-0093-5)
- [9] R.P. Darken, B. Peterson, Spatial orientation, wayfinding, and representation." In K. M. Stanney (Ed.), *Handbook of Virtual Environments: Design, Implementation, and Applications*, (California: Erlbaum, 2002, 493-518).
- [10] A. Abubakar, T. Waili, T. Mantoro, *Unveiling the support of 3D representation in mobile devices for pedestrians navigation aid*, Proceedings of IEEE International Conference on Multimedia Computing and Systems, (Page: 384 - 389 Year of Publication: 2012, ISBN: 978-1-4673-1518-0).
- [11] A. Nurminen, A. Oulasvirta, Designing interactions for navigation in 3D mobile maps." In L. Meng et al. (Ed.), *Handbook of Map-based Mobile Services* (Springer- Verlag, Berlin Heidelberg, 2008. 198-227).
- [12] A. Oulasvirta, A.Nurminen, N. Annu-Maaria, *Interacting with 3D and 2D mobile maps: an exploratory study*, Helsinki Institute for Information Technology April 11 (2007).
- [13] A. Abubakar, A. Zeki, H. Chiroma, T. Herawan., Investigating Rendering Speed and Download Rate of Three-Dimension (3D) Mobile Map Intended for Navigation Aid Using Genetic Algorithm, In T. Herawan et al. (Ed), *Handbook of Recent Advances on Soft Computing and Data Mining* (Springer International Publishing Switzerland, 2014. 261-271).
- [14] T. Mantoro, A. Abubakar, *Pragmatic framework of 3D visual navigation for mobile user*, Proceeding of International Conference on Information and Communication Technology for the Muslim World (Page: D19 - D24 Year of Publication: 2010, ISBN: 978-1-4244-7920-7).
- [15] Memon, J., Abd Rozan, M.Z., Uddin, M., Abubakar, A., Chiroma, H., Daud, D., Randomized text encryption: A new dimension in cryptography, (2014) *International Review on Computers and Software (IRECOS)*, 9 (2), pp. 365-373.

Acknowledgments

This research is funded by the Universiti Teknologi Malaysia (UTM) in collaboration with the Malaysian Ministry of Education under the Vot no. 4F238. The authors would like to thank the Research Management Centre of UTM and the Malaysian Ministry of Education for their support and cooperation including students and other individuals who are either directly or indirectly involved in this project.

Authors' information

¹Kulliyyah of Information and Communication Technology, International Islamic University Malaysia.

^{2,3}Advanced Informatics School (AIS), Universiti Teknologi Malaysia.

⁴Faculty of Science and Technology, Universitas Siswa Bangsa International, Jakarta, Indonesia.



Adamu Abubakar is currently an Assistant Professor at the International Islamic University of Malaysia, Kuala Lumpur. His academic qualifications were obtained from Bayero University Kano Nigeria, for bachelor and post-graduate diploma and master degrees, and from the International Islamic University Malaysia for his PhD degree. His research areas of interest include Navigation, Network and Information Security, Machine Learning, Human Computer Interaction (HCI), Information Retrieval Neural Networks, Genetic Algorithms and Fuzzy Logic, Data Mining, Image Processing, Web Design and Security, and Information Systems. He is now working on 3D Mobile Navigation Aids, Cryptography, Web Design and Security, and Digital Watermarking. He is a member of IEEE and ACM.

E-mail: adamu@iiu.edu.my



Sadegh Ameri is currently a PhD candidate in software engineering at AIS (Advanced Informatics School), UTM International Campus, Kuala Lumpur. In 2008, he received his Bachelor of Science (B.Sc.) in software engineering from Islamic Azad University of Shiraz in Iran and his Master's degree in software engineering from UTM, Kuala Lumpur, Malaysia in 2011. He received Best Student Award in 2011 for academic excellence in Master's studies. His research areas of interest include Indoor User Tracking, Cloud Computing, Software Engineering, Mobile Computing, Web Applications, Data Mining and Information Systems.

E-mail: sadegh.ameri@gmail.com



Suhaimi Ibrahim is an associate professor of software engineering currently serving UTM as the Deputy Dean (Research and Development) of UTM-AIS (Advanced Informatics School), UTM International Campus, Kuala Lumpur. He was formerly the Deputy Director of CASE (Centre for Advanced Software Engineering), and has approximately 30 years of experience in teaching and research. He is actively involved in many short-term and long-term national research grants under the research university and government funds. He was awarded an ISTQB certified tester certification and has been a board member of the Malaysian Software Testing Board (MSTB) since 2008. As a board member he is actively involved in promoting professional software testing into the local university syllabus and curriculum via a pilot university programme under the MSTB-government initiatives. He also participated in the design and implementation of the Software Engineering curriculum at Bachelor, Master and Engineering Doctorate levels.

E-mail: suhaimiibrahim@utm.my



Teddy Mantoro is Vice Rector for Academic and Student Affairs (ad-interim) and Dean of the Faculty of Science and Technology, Universitas Siswa Bangsa International, Jakarta, Indonesia. He received his PhD from the School of Computer Science at the Australian National University (ANU), Canberra, Australia. His research interests focus on Information Security, pervasive/ubiquitous computing, context aware computing, mobile computing and intelligent environment. He has published more than 110 conference/journal papers and 4 computing books in Indonesia, USA, and Germany. In 2013 he served on committees and review boards for more than 30 International conferences. He is also a Senior Member of IEEE, the managing editor for the International Journal of Mobile Computing and Multimedia Communications (IJMCMC), and editor for 3 other journals. He is the founder and was the leader of the Integ Lab (Intelligent Environment Research Group) at KICT, International Islamic University Malaysia (IIUM), Kuala Lumpur, Malaysia. Integ Lab has received 43 medals since 2009 from national and international innovation technology competitions. He holds 4 (four) Malaysian patents under his name.

E-mail: tmantoro@gmail.com

Model-Driven Transformation for GWT with Approach by Modeling: from UML Model to MVP Web Applications

R. Esbai¹, M. Erramdani², S. Mbarki³

Abstract – *The continuing evolution of business needs and technology makes Web applications more demanding in terms of development, usability and interactivity of their user interfaces. To cope with this complexity, several frameworks have emerged and a new type of Web applications called RIA (Rich Internet Applications) has recently appeared providing richer and more efficient graphical components similar to desktop applications. Given this diversity of solutions, the generation of a code based on UML models has become important. This paper presents the application of the MDA (Model Driven Architecture) to generate, from the UML model, the Code following the MVP pattern (Model-View-Presenter) for a RIA using the standard MOF 2.0 QVT (Meta-Object Facility 2.0 Query-View-Transformation) as a transformation language. We adopt GWT (Google web Toolkit) for creating a target meta-model to generate an entire GWT-based web application. The transformation rules defined in this paper can generate, from the class diagram, an XML file containing the Views, the Models, and the Presenter. This file can be used to generate the necessary code of a RIA. Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.*

Keywords: *GWT, Model Transformation, MOF 2.0 QVT, Model View Presenter, Transformation Rules*

I. Introduction

In recent years many organizations have begun to consider MDA (Model-View-Presenter) as an approach to design and implement enterprise applications. The key principle of MDA is the use of models at different phases of application development by implementing many transformations.

These changes are present in MDA, and help transform a CIM (Computation Independent Model) into a PIM (Platform Independent Model) or to obtain a PSM (Platform Specific Model) from a PIM.

Rich Internet applications (RIAs) combine the simplicity of the hypertext paradigm with the flexibility of desktop interfaces. Moreover, RIAs provide a new client-server architecture that reduces significantly network traffic using more intelligent asynchronous requests that send only small blocks of data. In fact, the technological advances of RIAs require from the developer to represent a rich user interface based on the composition of Graphical User Interface (GUI) widgets, to define an event-based choreography between these widgets and to establish a fine grained communication between the client and the server layers. Many frameworks that implement the MVP pattern have emerged; for instance: Mvp4g [1], GWT [2], Echo2 [3], JFace [4], Vaadin [5], ZK [6], Nucleo .NET [7].

GWT is an AJAX framework, developed by Google, which permits us to create RIAs by writing the browser-side code in Java, thus gaining all the advantages of Java

(e.g. compiling, debugging, etc.) and generating a generic JavaScript and HTML code that can be executed in any browser. Moreover, GWT makes every attempt to be flexible allowing us to integrate with other client AJAX frameworks (e.g. Script.aculo.us, Dojo, Yahoo! UI) and with server Java frameworks such as Struts [8], EJB, etc. In [9][10], the authors have developed a source and a target meta-model. The first was a PIM meta-model specific to class diagrams. The second was a PSM meta-model for N-tiers web applications (particularly Struts, Spring, DTO, Hibernate) without UI. The purpose of our contribution is to produce and generate an RIA PSM model (particularly GWT), implementing MVP pattern, from the class diagram.

In this case, we elaborate a number of transformation rules using the approach by modeling and MOF 2.0 QVT, as transformation language, to permit the generation of an XML file that can be used to produce the required code of the target application. The advantage of this approach is the bidirectional execution of transformation rules.

This paper is organized as follows: related works are presented in the second section, the third section defines the MDA approach, and the fourth section presents GWT and the MVP model and its implementation as a framework. The transformation language MOF 2.0 QVT is the subject of the fifth section. In the sixth section, we present the UML and MVP meta-models. In the seventh section, we present the transformation rules using MOF 2.0 QVT from UML source model to the MVP target

model. The last section concludes this paper and presents some perspectives.

II. Related Work

Many researches on MDA and generation of code have been conducted in recent years. The most relevant are [11]-[24]. The authors of the work [19] show how to generate JSPs and JavaBeans using the UWE [18], and the ATL transformation language [17]. Among future works cited, the authors considered the integration of AJAX into the engineering process of UWE.

Two other works followed the same logic and have been the subject of two works [15][16]. A meta-model for Ajax was defined using AndroMDA tool.

The generation of Ajax code has been illustrated by an application CRUD (Create, Read, Update, and Delete) that manages people.

Meliá, Pérez and Díaz propose in [25] a new approach called OOH4RIA which proposes a model driven development process that extends OOH methodology.

It introduces new structural and behavioral models in order to represent a complete RIA and to apply transformations that reduce the effort and accelerate its development. In another work [26] they present a tool called OIDE (OOH4RIA Integrated Development Environment) aimed at accelerating the RIAs development through the OOH4RIA approach which establishes a RIA-specific model-driven process. The Web Modeling Language (WebML) [27] is a visual notation for specifying the structure and navigation of legacy web applications. The notation greatly resembles UML class and Entity-Relation diagrams. Presentation in WebML is mainly focused on look and feel and lacks the degree of notation needed for AJAX web user interfaces [28][29]. Nasir, Hamid and Hassan [23] have presented an approach to generate a code for the .Net application Student Nomination Management System. The method used is WebML and the code was generated by applying the MDA approach, but the creation was not done according to the .Net MVC2 logic.

This paper aims to finalize the work presented in [9][10], by applying the standard MOF 2.0 QVT to develop the transformation rules aiming at generating the MVP target model with UI. It is actually the only work for reaching this goal.

III. Model Driven Architecture (MDA)

In November 2000, OMG, a consortium of over 1 000 companies, initiated the MDA approach. The key principle of MDA is the use of models at different phases of application development. Specifically, MDA advocates the development of requirements models (CIM), analysis and design (PIM) and code (PSM).

The MDA architecture [30] is divided into four layers. In the first layer, we find the standard UML (Unified Modelling Language), MOF (Meta-Object Facility) and CWM (Common Warehouse Meta-model).

In the second layer, we find a standard XMI (XML Metadata Interchange), which enables the dialogue between middlewares (Java, CORBA, .NET and web services).

The third layer contains the services that manage events, security, directories and transactions. The last layer provides frameworks which are adaptable to different types of applications namely Finance, Telecommunications, Transport, medicine, E-commerce and Manufacture, etc.).

The major objective of MDA [31] is to develop sustainable models; those models are independent from the technical details of platforms implementation (JavaEE, .Net, PHP or other), in order to enable the automatic generation of all codes and applications leading to a significant gain in productivity. MDA includes the definition of several standards, including UML [32], MOF [33] and XMI [34].

IV. The MVP Pattern

The Model View Presenter is a derivative of the Model View Controller Pattern. Its aim is to provide a cleaner implementation of the Observer connection between Application Model and view.

MVP is a user interface architectural pattern engineered to facilitate automated unit testing and improve the separation of concerns in presentation logic.

Fig. 1 shows the architecture of the MVP pattern. The main feature of this pattern is to be composed of:

- **The model** is an interface defining the data to be displayed or otherwise acted upon in the user interface.
- **The view** is a passive interface that displays data (the model) and routes user commands (events) to the presenter to act upon that data.
- **The presenter** acts upon the model and the view. It retrieves data from repositories (the model), and formats it for display in the view.

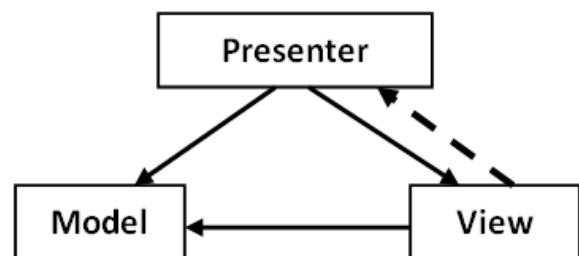


Fig. 1. MVP Architecture

Based on this model many frameworks are designed to help developers build the presentation layer of their user interfaces. In the Java community, many frameworks that implements MVP pattern have emerged, among them: Echo2, JFace, Swing, Vaadin, ZK framework, GWT, etc.

The GWT project is one of the best examples. Implementing MVP in Google Web Toolkit requires only that some component implement the view interface.

IV.1. The GWT Framework

Google Web Toolkit (GWT) [35] is an open source web development framework that allows developers to easily create high-performance AJAX applications using Java. With GWT, you are able to write your front end in Java, and it compiles your source code into highly optimized, browser-compliant JavaScript and HTML.

However, GWT is not the only framework for managing the user interfaces. Indeed, other frameworks have been designed for the same goal, but GWT is the most mature. The main advantage of GWT is the reduced complexity compared to other frameworks of the same degree of power, for instance, JFace, Flex and Vaadin.

V. The Transformations of MDA Models

MDA establishes the links of traceability between the CIM, PIM and PSM models through to the execution of the models' transformations. The models' transformations recommended by MDA are essentially the CIM transformations to PIM and PIM transformations to PSM.

V.1. Approach by Modeling

Currently, the models' transformations can be written according to three approaches: The approach by Programming, the approach by Template and the approach by Modeling. The approach by Modeling is the one used in the present paper. It consists of applying concepts from model engineering to models' transformations themselves. The objective is modeling a transformation, to reach perennial and productive transformation models, and to express their independence towards the platforms of execution. Consequently, OMG elaborated a standard transformation language called MOF 2.0 QVT [36]. The advantage of the approach by modeling is the bidirectional execution of transformation rules. This aspect is useful for the synchronization, the consistency and the models reverse engineering [37].

Figure 2 illustrates the approach by modeling. Models transformation is defined as a model structured according to MOF 2.0 QVT meta-model. The MOF 2.0 QVT meta-model expresses some structural correspondence rules between the source and target meta-model of a transformation. This model is a perennial and productive model that is necessary to transform in order to execute the transformation on an execution platform.

V.2. MOF 2.0 QVT

Transformations models are at the heart of MDA, a standard known as MOF 2.0 QVT being established to model these changes. This standard defines the metamodel for the development of transformation model.

The QVT standard has a hybrid character (declarative / imperative) in the sense that it is composed of three different transformation languages (see Fig. 3).

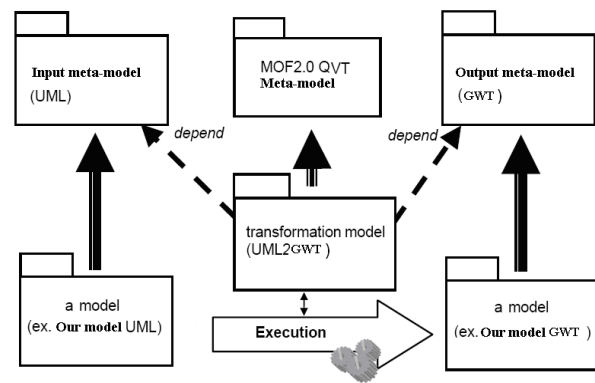


Fig. 2. Approach by Modeling

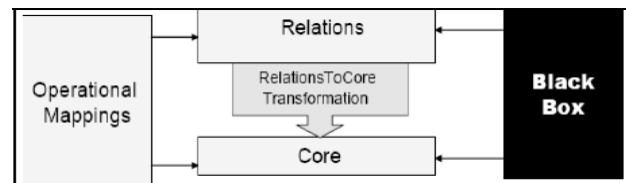


Fig. 3. The QVT Structure

The declarative part of QVT is defined by Relations and Core languages, with different levels of abstraction. Relations are a user-oriented language for defining transformations in a high level of abstraction. It has a syntax text and graphics. Core language forms the basic infrastructure for the declaration part; this is a technical language of lower level determined by textual syntax. It is used to specify the semantics of Relations language in the form of a Relations2Core transformation. The declarative vision comes through a combination of patterns, source and target side to express the transformation. The imperative QVT component is supported by Operational Mappings language. The vision requires an explicit imperative navigation as well as an explicit creation of target model elements. The Operational Mappings language extends the two declarative languages of QVT, adding imperative constructs (sequence, selection, repetition), etc and constructs in OCL edge effect. The imperative style languages are better suited for complex transformations including a significant algorithm component. Compared to the declarative style, they have the advantage of optional case management in a transformation. For this reason, we chose to use an imperative style language in this paper. Finally, QVT suggests a second extension mechanism for specifying transformations invoking the functionality of transformations implemented in an external language Black Box.

This work uses the QVT-Operational mappings language implemented by Eclipse modeling [38].

V.3. OCL (Object Constraint Language)

Object Constraint Language (OCL) is a formal language used to describe expressions on UML models.

These expressions typically specify invariant conditions that must hold for the system being modeled or queries over objects described in a model. Note that when the OCL expressions are evaluated, they do not have side effects. OCL expressions can be used to specify operations / actions that, when executed, do alter the state of the system. UML modelers can use OCL to specify application-specific constraints in their models.

In MOF 2.0 QVT, OCL is extended to Imperative OCL as part of QVT Operational Mappings.

Imperative OCL added services to manipulate the system states (for example, to create and edit objects, links and variables) and some constructions of imperative programming languages (for example, loops and conditional execution). It is used in QVT Operational Mappings to specify the transformations.

QVT defines two ways of expressing model transformations: declarative and operational approaches.

The declarative approach is the Relations language where transformations between models are specified as a set of relationships that must hold for successful transformation.

The operational approach allows either defining transformations using a complete imperative approach or complementing the relational transformations with imperative operations, by implementing relationships. Imperative OCL adds imperative elements of OCL, which are commonly found in programming languages like Java. Its semantics are defined in [36] by a model of abstract syntax. The complete abstract syntax ImperativeOCL is shown in Fig. 4. The most important aspect of the abstract syntax is that all expression classes must inherit `OclExpression`.

OclExpression is the base class for all the conventional expressions of OCL. Therefore, Imperative Expressions can be used wherever there is OclExpressions.

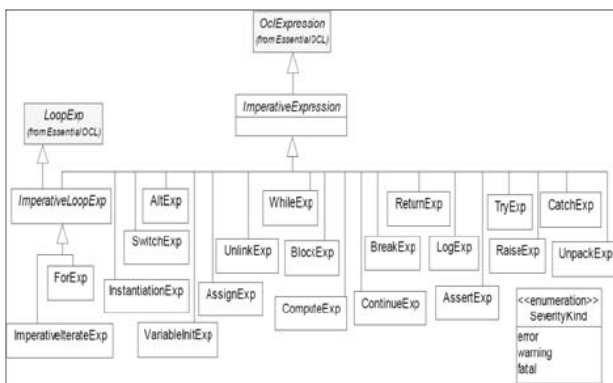


Fig. 4. Imperative Expressions of ImperativeOCL

VI. UML and MVP Meta-models

To develop the transformation algorithm between source and target model, we present in this section, the various meta-classes forming the meta-model UML source and the meta-model MVP target.

VI.1. Meta-model UML Source

The source meta-model structures a simplified UML model based on packages containing data types and classes. Those classes contain typed properties and they are characterized by multiplicities (upper and lower). The classes are composed of operations with typed parameters. Fig. 5 illustrates the source meta-model.

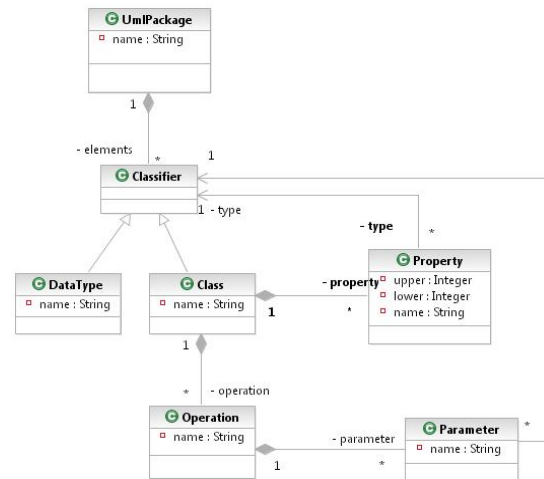


Fig. 5. Simplified UML meta-model

- **UmlPackage:** is the concept of UML package. This meta-class is connected to the meta-class *Classifier*.
- **Classifier:** This is an abstract meta-class representing both the concept of UML class and the concept of data type.
- **Class:** is the concept of UML class.
- **DataType:** represents UML data type.
- **Operation:** is used to express the concept of operations of a UML class.
- **Parameter:** expresses the concept of parameters of an operation. These are of two types, Class or DataType. It explains the link between Parameter meta-class and Classifier meta-class.
- **Property:** expresses the concept of properties of a UML class. These properties are represented by the multiplicity and meta-attributes upper and lower.

The works of Mbarki and Erramdani [20] [21] contains more details related to this section topic.

VI.2. Meta-model GWT MVP Target

Our target meta-model is composed of two essential part. Figure 6 illustrates the first part of the target meta-model. This meta-model represents a simplified version of the MVP pattern. It presents the different meta-classes to express the concept of MVP implementation:

- **UIPackage:** represents the project package. This meta-class is connected to the meta-class *MvpPackage*
- **MvpPackage:** represents the different meta-classes to express the concept of MVP. This meta-class is connected to the meta-class *ClientPackage* and *SharedPackage* which represents respectively View and Model package.

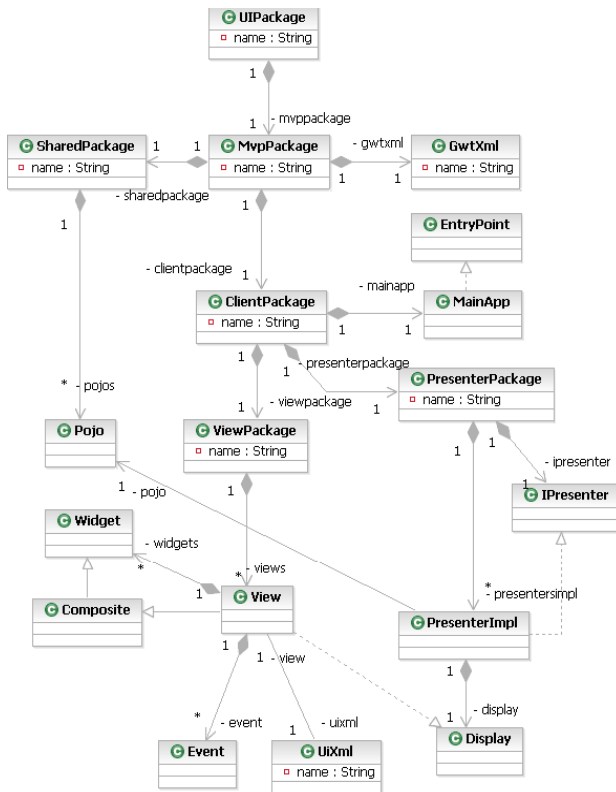


Fig. 6. The proposed MVP metamodel

- **GwtXml:** expresses the concept of GWT module Encapsulates units of GWT configurations (paths, properties, deferred binding etc); defined in an XML module file and stored in the Java package hierarchy.
- **ClientPackage:** represents the client package, in this package, we will typically find, and put, all the code required for the client side part of our application (the part in the browser). This meta-class is connected to the meta-class *PresenterPackage* and *ViewPackage*
- **MainApp:** this meta-class implements EntryPoint interface. When a module is loaded, entry point class is instantiated and its `onModuleLoad()` method gets called.
- **EntryPoint:** represents the concept of entry point interface containing the method *onModuleLoad()*. Implement this interface to allow a class to act as a module entry point.
- **PresenterPackage:** represents the different meta-classes to express the concept of Presenter. This Presenter is Responsible for getting the data, driving the view, listening for GUI events, implements business logic
- **IPresenter:** represents the concept of basic presenter interface that all of our presenters will implement and containing the methods *bind()* and *go()*
- **PresenterImpl:** expresses the concept of specific Presenter implementation all methods to bind and go are implemented in this meta-class.
- **Display:** represents the concept of the inner interface type of the view is determined by the *getView()* method.

- **View:** expresses the concept of the view contains all of the UI components that make up our application.
- **SharedPackage:** represents package which contains the different meta-classes to express the concept of model.
- **Pojo:** represents the concept of pojo. The latter extends the meta-class *Class*. The pojoes represents objects in the area of application.
- **Widget:** expresses the concept of the GWT Widget.

Fig. 7 illustrates the second part of target meta-model. Like the Abstract Window Toolkit (AWT) and Swing, GWT is based on widgets. To create a user interface, you instantiate widgets, add them to panels, and then add your panels to the application's root panel, which is a top-level container that contains all of the widgets in a particular view.

GWT contains many widgets whose classes are described by an inheritance hierarchy. An illustration of some of those widgets is shown in Fig. 7.

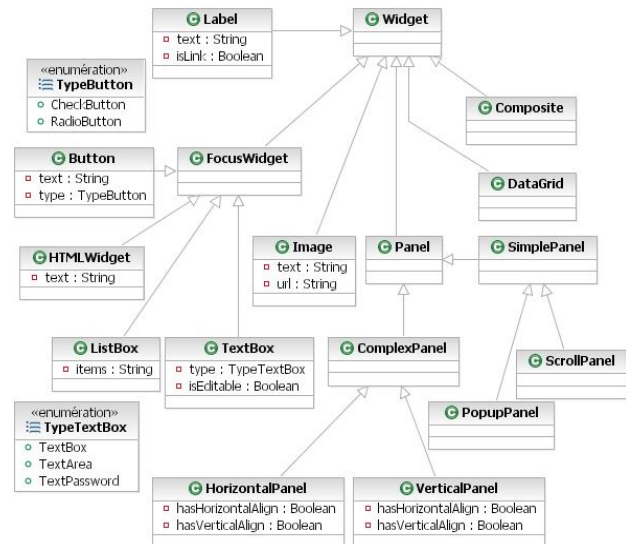


Fig. 7. Simplified GWT metamodel

- **Panel:** A panel that lets you place widgets a pixel locations.
- **Button:** A button that the user can click.
- **Composite:** An opaque wrapper for a set of widgets.
- **DataGrid:** A table that arranges its widgets in a grid.
- **HorizontalPanel:** A panel that arranges its widgets horizontally.
- **VerticalPanel:** A panel that arranges its widgets vertically.
- **Image:** An image that can fire load events when it loads its corresponding image file.
- **Label:** Text that supports word wrap and horizontal alignment.
- **PopupPanel:** A panel that pops up when it's shown.
- **ScrollPane:** A panel that automatically adds scrollbars to itself on demand.
- **TextBox:** A single-line text widget.
- **ListBox:** A list of choices that the user can select. Can be either a list box or a drop-down list.

VII. The Process of Transforming UML Source Model to MVP Target Model

CRUD operations (Create, Read, Update, and Delete) are most commonly implemented in all systems. That is why we have taken into account in our transformation rules these types of transactions.

We first developed ECORE models corresponding to our source and target meta-models, and then we implemented the algorithm (see sub-section 7.1) using the transformation language QVT Operational Mappings.

To validate our transformation rules, we conducted several tests. For example, we considered the class diagram (see Figure 8). After applying the transformation on the UML model, composed by the class Employee, we generated the target model (see Fig. 11).



Fig. 8. UML instance model

VII.1. The Transformation Rules

By source model, we mean model containing the various classes of our business model. The elements of this model are primarily classes.

Main algorithm:

```
input umlModel:UmlPackage
output gwtModel:UIPackage
begin
  create UIPackage crudProjectPackage
  create MvpPackage mvpPackage
  create ClientPackage clientPackage
  create MainApp mainapp
  link mainapp to clientPackage
  create PresenterPackage presenterPackage
  create IPresenter ipresenter
  ipresenter.name = 'IPresenter'
  ipresenter.methods = declaration of {do,bind}
  link ipresenter to presenterPackage
  for each e ∈ source model
    x = transformationRuleOne(e)
    link x to presenterPackage
  end for
  create ViewPackage viewPackage;
  for each e ∈ source model
    x = transformationRuleTwo(e)
    link x to viewPackage
  end for
  create SharedPackage sharedPackage;
  for each e ∈ source model
    x = transformationRuleThree(e)
    link x to sharedPackage
  end for
```

```
create GwtXml gwtxml;
link presenterPackage to clientPackage
link viewPackage to clientPackage
link clientPackage to mvpPackage
link mvpPackage to crudProjectPackage
link sharedPackage to crudProjectPackage
link gwtxml to crudProjectPackage
return crud
end
```

```
function
transformationRuleOne(e:Class):PresenterImpl
begin
  create PresenterImpl presenterImpl
  presenterImpl.name = e.name+
  'PresenterImpl'
  for each el ∈ PresenterPackage
    if el.name = 'I'+e.name+ 'Presenter'
      put el in interfaces
    end if
  end for
  link interfaces to presenterImpl
  return presenterImpl
end
```

```
function
transformationRuleTwo(e:Class):ViewPackage
begin
  create ViewPackage vp
  for each e ∈ source model
    if e.methods.name ≠ 'remove'
      create View page
      link page to vp
    end if
  end for
  return vp
end
```

```
function
transformationRuleThree(e:Class):Pojo
begin
  create Pojo pj
  pj.name = e.name
  pj.attributes = e.properties
  return pj
end
```

Fig. 9 illustrates the first part of the transformation code of UML source model to the MVP target model.

The transformation uses as input a UML type model, named umlModel, and as output a GWT type model named gwtModel. The entry point of the transformation is the main method. This method makes the correspondence between all elements of type UmlPackage of the input model and the elements of type UIPackage output model.

The objective of the second part of this code is to transform a UML package to GWT package, by creating the elements of type package 'Presenter', 'View' and 'Shared'. It is a question of transforming each class of package UML, to IPresenter and PresenterImpl in the Presenter package, and to Pojo, in the Shared package, to Display contains widgets in the View Package without forgetting to give names to the different packages.

The methods presented in Fig. 10 means that each operation in a class corresponds to View.

The codes and models are publicly available online <http://sites.google.com/site/uml2mvp/>.

```
uml2gwt.qtdo
modeltype UMLMM uses "http://umlmm.ecore";
modeltype gwtMM uses "http://gwtmm.ecore";

transformation uml2gwt(in umlModel:UMLMM, out gwtModel:gwtMM);
main() {
    umlModel.objects()[UmlPackage]->map UmlPackage2CrudProjectPackage();
}

mapping UmlPackage::UmlPackage2CrudProjectPackage () : UIPackage {
    name:= 'crud'+self.name;
    mvppackage:= object MvpPackage {
        name:= 'mvpPackage';
        clientpackage:= object ClientPackage {
            name:= 'clientPackage';
            mainapp:= object MainApp {
                name:= 'MainApp';
            };
        };
        presenterpackage:= object PresenterPackage {
            name:= 'presenter';
            ipresenter:= map class2IPresenter();
            presentersimpl:= umlModel.objects()[Class]->map class2PresenterImpl();
        };
        viewpackage:= object ViewPackage {
            name:= 'view';
            views:= umlModel.objects()[Class].map class2view();
        };
    };
    sharedpackage:= object SharedPackage {
        name:= 'shared';
        pojoes:= umlModel.objects()[Class]->map class2Pojo();
    };
    gwtxml:= object GwtXml {
        name:= result.name;
    };
};
};
```

Fig. 9. The transformation code UML2Gwt

```
@mapping Operation::Op2View (cl:Class) : View{
    name:=self.name+cl.name+'View';
    widgets:= Sequence {
        object Panel {
            name:=self.name+'Panel';
        }
    };
    if(self.name<>'display')then
        widgets+= cl._property.map property2widget(self.name)
    endif;
    if(self.name='display')then
        widgets+= object DataGrid {
            name:=self.name+cl.name+'s';
        }
    endif;
    if(self.name='create')then
        widgets+=object Button {
            name:=self.name+'Button';
        }
    endif;
    if(self.name='update')then
        widgets+=object Button {
            name:=self.name+'Button';
        }
    endif;
    widgets+=object Button {
        name:= 'cancelButton';
    }
};
```

Fig. 10. The mapping Op2View

VII.2. Result

Fig. 11 shows the result after applying the transformation rules. The first element in the generated PSM model is UIPackage which includes MvpPackage that contains gwt.xml file, Client Package and Shared Package.

The Client Package contains the main application, the Presenter Package and the View Package that contains the Three Views, namely CreateEmployeeView, DisplayEmployeeView and UpdateEmployeeView.

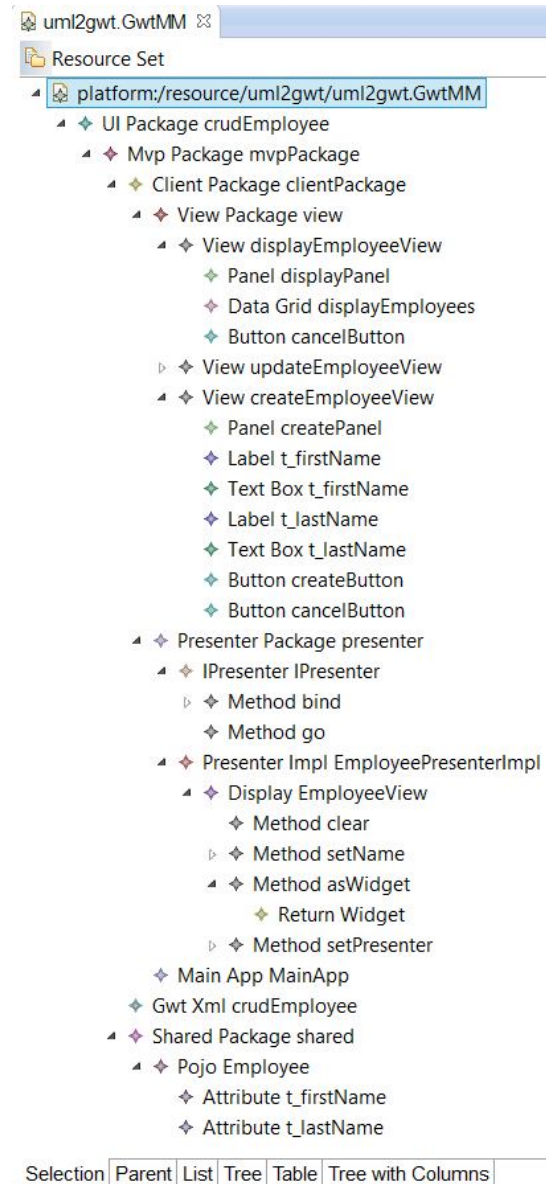


Fig. 11. Generated PSM MVP model

Since the operation of the removal requires any view, we'll go to every view element, which contains a multiple element widget like Panel, firstNameTextBox, lastNameTextBox, actionButton and cancelButton. Since the view Display contains the DataGrid widget that contains removal button. The Presenter Package includes one presenter' interface, one presenter' implementation that contains methods with their parameters and their implementations and the last package element in the generated PSM model is Shared Package which contains one Pojo' object that contains their attributes correspond to the object 'Employee'.

VIII. Conclusion and Perspectives

In this paper, we applied the MDA approach to generate the MVP web application based on UML class diagram.

The purpose of our contribution is to finalize the works presented in [9] [10]. This involves developing all meta-classes needed to be able to generate a GWT application respecting a MVP pattern and then we applied the approach by modeling and used the MOF 2.0 QVT standard as a transformation language. The transformation rules defined allow browsing the source model instance class diagram, and generating, through these rules, an XML file containing layers of MVP architecture according to our target model.

This file can be used to produce the necessary code of the target application. The algorithm of transformation manages all CRUD operations. Moreover, it can be re-used with any kind of methods represented in the UML class diagram. In the future, this work should be extended to allow the generation of other components of Web application besides the configuration files. Afterward we can consider integrating other frameworks like Flex and JFace.

References

- [1] Mvp4g A framework to build a GWT application the right way (<https://code.google.com/p/mvp4g/>)
- [2] GWT source web site (<https://code.google.com/p/google-web-toolkit/>)
- [3] Echo2 source web site (<http://echopoint.sourceforge.net/>)
- [4] Harris, Robert; Warner, Rob, *The Definitive Guide to SWT and JFACE (1st ed.)*, (Apress, 2004).
- [5] Vaadin Framework web site (<https://vaadin.com/home>)
- [6] ZK framework web site (<http://www.zkoss.org>)
- [7] Nucleo .NET framework web site (<http://nucleo.codeplex.com/>)
- [8] Apache Software Foundation: The Apache Struts Web Application Software Framework (<http://struts.apache.org>).
- [9] Esbai, R, Erramdani, M., Mbarki, S., Arrassen, I, Meziane, A. and Moussaoui, M., Model-Driven transformation with approach by modeling: From UML to N-tiers Web Model, *International Journal of Computer Science Issues (IJCSI)*, Vol. 8, Issue 3, May 2011, ISSN (Online): 1694-0814
- [10] Esbai, R, Erramdani, M., Mbarki, S., Arrassen, I, Meziane, A. and Moussaoui, M., Transformation by Modeling MOF 2.0 QVT: From UML to MVC2 Web model, *InfoComp - Journal of Computer Science*, vol. 10, no. 3, p. 01-11, September of 2011, ISSN 1807-4545.
- [11] AndroMDA web site (<http://www.andromda.org/>).
- [12] Bezivin, J., Busse, S., Leicher, A., Suss, J.G, *Platform Independent Model Transformation Based on TRIPLE*. Proceedings of the 5th ACM/IFIP/USENIX International Conference on Middleware, (Page: 493, Year of publication: 2004).
- [13] Bezivin, J., Hammoudi, S., Lopes, D., Jouault, F., *Applying MDA approach for web service platform*. Proceedings of the 8th IEEE International Enterprise Distributed Object Computing Conference, (Page: 58, Year of publication: 2004).
- [14] Cong, X., Zhang, H., Zhou, D., Lu, P., Qin, L., A Model-Driven Architecture Approach for Developing E-Learning Platform , Entertainment for Education, *Digital Techniques and Systems Lecture Notes in Computer Science*, Volume 6249/2010, 2010.
- [15] Distant, D., Rossi, G., Canfora, G., *Modeling Business Processes in Web Applications: An Analysis Framework*. In Proceedings of the The 22nd Annual ACM Symposium on Applied Computing (Page: 1677, Year of publication: 2007, ISBN: 1-59593-480-4).
- [16] Gharavi, V., Mesbah, A., Deursen, A. V., *Modelling and Generating AJAX Applications: A Model-Driven Approach*, Proceeding of the 7th International Workshop on Web-Oriented Software Technologies, New York, USA (Page: 38, Year of publication: 2008, ISBN: 978-80-227-2899-7)
- [17] Jouault, F., Allilaire, F., Bézivin, J., Kurtev, I., ATL: A model transformation tool. *Science of Computer Programming-Elsevier* Vol. 72, n. 1-2: pp. 31-39, 2008.
- [18] Koch, N., *Transformations Techniques in the Model-Driven Development Process of UWE*, Proceeding of the 2nd International Workshop Model-Driven Web Engineering, Palo Alto (Page: 3 Year of publication: 2006 ISBN: 1-59593-435-9).
- [19] Kraus, A., Knapp, A., Koch N., *Model-Driven Generation of Web Applications in UWE*. Proceeding of the 3rd International Workshop on Model-Driven Web Engineering, CEUR-WS, Vol. 261, 2007
- [20] Mbarki, S., Erramdani, M., Toward automatic generation of mvc2 web applications, *InfoComp - Journal of Computer Science*, Vol.7 n.4, pp. 84-91, December 2008, ISSN: 1807-4545.
- [21] Mbarki, S., Erramdani, M., Model-driven transformations: From analysis to MVC 2 web model, (2009) *International Review on Computers and Software (IRECOS)*, 4 (5), pp. 612-620.
- [22] Mbarki, S., Rahmouni, M., Erramdani, M., *Transformation ATL pour la génération de modèles Web MVC 2*, Proceeding of the 10e Colloque Africain sur la Recherche en Informatique et en Mathématiques Appliquées, Theme5:Information Systems, CARI (Year of publication: 2006).
- [23] Nasir, M.H.N.M., Hamid, S.H., Hassan, H., WebML and .NET Architecture for Developing Students Appointment Management System, *Journal of applied science*, Vol. 9, n. 8, pp. 1432-1440, 2009
- [24] Ndie, T. D., Tangha1, C., Ekwoge, F. E., MDA (Model-Driven Architecture) as a Software Industrialization Pattern: An Approach for a Pragmatic Software Factories. *J. Software Engineering & Applications*, pages 561-571, 2010
- [25] Meliá S., Gómez J., Pérez P., Díaz O., *A Model-Driven Development for GWT-Based Rich Internet Applications with OOH4RIA*, Proceedings of ICWE '08. Eighth International Conference on, Yorktown Heights, NJ, (Page: 13, Year of publication: 2008, ISBN: 978-0-7695-3261-5).
- [26] Meliá S., Gómez J., Pérez S., Diaz O. Facing Architectural and Technological Variability of Rich Internet Applications. *IEEE Internet Computing*, vol. 99, pp.30-38, 2010.
- [27] S. Ceri, P. Fraternali, and A. Bongio. Web modeling language (WebML): a modeling language for designing web sites. *Computer Networks*, vol. 33(1-6) pp137-157, 2000.
- [28] Preciado J. Carlos, M. Linaje, S. Comai, and F. Sanchez-Figueroa. *Designing Rich Internet Applications with Web engineering methodologies*. Proceedings of the 9th IEEE International Symposium on Web Site Evolution (WSE'07)(Page: 23 Year of publication: 2007).
- [29] Trigueros M. L., J. C. Preciado, and F. S'anchez-Figueroa. *A method for model based design of Rich Internet Application interactive user interfaces*. In ICWE'07: Proceedings of the 7th International Conference Web Engineering (page: 226 Year of publication: 2007).
- [30] Miller, J., Mukerji, J., al. *MDA Guide Version 1.0.1* (OMG, 2003).
- [31] Pastor, O., Molina J.C, *Model-Driven Architecture in Practice: A Software Production Environment Based on Conceptual Modeling* (New York: Springer-Verlag, 2007).
- [32] UML Infrastructure Final Adopted Specification, version 2.0, September 2003, <http://www.omg.org/cgi-bin/doc?ptc/03-09-15.pdf>
- [33] *Meta Object Facility (MOF), version 2.0* (OMG, 2006)
- [34] *XML Metadata Interchange (XMI), version 2.1.1* (OMG, 2007),
- [35] GWT project web site <http://www.gwtproject.org/>
- [36] *Meta Object Facility (MOF) 2.0 Query/View/Transformation (QVT), Version 1.1* (OMG, 2009).
- [37] Czarnecki, K., Helsen, S., *Classification of Model Transformation Approaches*, Proceedings of the 2nd OOPSLA'03 Workshop on Generative Techniques in the Context of MDA. Anaheim (Year of publication: 2003).
- [38] Eclipse modeling, <http://www.eclipse.org/modeling/>.

Authors' information

¹Department of Commerce, ENCGO, Mohammed 1 University, Oujda, Morocco.

²Department of Management, EST, Mohammed 1 University, Oujda, Morocco.

³Department of Mathematics and Computer Science, Faculty of Science, Ibn Tofail University, Kenitra, BP 133, Morocco.



Redouane Esbai teaches the concept of Information System at Mohammed 1 University,. He got his thesis of national doctorate in 2012. He got a degree of an engineer in Computer Sciences from the National School of Applied Sciences at Oujda. He received his M.Sc. degree in New Information and Communication Technologies from the faculty of sciences and Techniques at Sidi Mohamed Ben Abdellah University. His activities of research in the MATSI Laboratory (Applied Mathematics, Signal Processing and Computer Science) focusing on MDA (Model Driven Architecture) integrating new technologies XML, Spring, Struts, GWT, etc.

E-mail: es.redouane@gmail.com



S. Mbarki received the B.S. degree in applied mathematics from Mohammed V University, Morocco, 1992, and Doctorat of High Graduate Studies degrees in Computer Sciences from Mohammed V University, Morocco, 1997. In 1995 he joined the faculty of science Ibn Tofail University, Morocco where he is currently a Professor in Department of mathematics and computer science. His research interests include software engineering and model driven architecture.

E-mail: mramdani69@yahoo.co.uk



Mohammed Erramdani teaches the concept of Information System at Med I University. He got his thesis of national doctorate in 2001. His activities of research in the MATSI Laboratory (Applied Mathematics, Signal Processing and Computer Science) focusing on MDA (Model Driven Architecture) integrating new technologies XML, EJB, MVC, Web Services, etc.

E-mail: mbarkisamir@hotmail.com

An Adaptive Bilateral Filter for Noise Removal and Novel Hybrid Fuzzy Cognitive Map-FNN Edge Detection Method for Images

T. Karthikeyan¹, N. P. Revathy²

Abstract – The important process in image processing is edge detection and it is considered as the basis for the pattern recognition owing to its effect on the subsequent processes of image processing. Edge detection is highly needed for processing of noisy images as noise is the main shortcoming of the images. There are many optimization algorithms such as SVM, WSVM classifiers so far reported for performing edge detection. Inspire of rendering better results, these existing methods suffer from lack of transparency thus affecting the overall performance of the system. This paper presents a novel method for edge detection which operates on the hybrid fuzzy cognitive map based fuzzy neural network (HFCM-FNN). This method employs a facile way of processing images using HFCM-FNN by reducing the execution time and upgrading the visual quality. The experimentation results shows the overall performance results of the proposed HFCM-FNN edge detection method and it can be compared with existing SVM, WSVM methods. **Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords: Edge Detection, Weighted SVM (WSVM), Fuzzy Cognitive Map (FCM), Hybrid Fuzzy Cognitive Map Based Fuzzy Neural Network (HFCM-FNN), Adaptive Bilateral Filtering (ABF), Noise Removal

Nomenclature

α	Small Constant
σ_n^2	Noise Variance
$\tilde{\sigma}_s$	Local Standard Deviations
\tilde{W}_l	Intensity Weight
J	Jacobian Matrix
e_f	Error
I	Identity Matrix

I. Introduction

Edge detection is an indispensable step in the computer vision and object recognition, because the most fundamental characteristic for image recognition is the edges of an image. Edges are boundaries between different textures. Edge also can be defined as discontinuities in image intensity from one pixel to another [1].

The edges for an image are always the important characteristics that offer an indication for a higher frequency. The goal of edge detection is to convert a 2D image into a set of curves. The salient features are expected to be the boundaries of objects that tend to produce sudden changes in the image intensity. They can show where shadows fall in an image or any other distinct change in the intensity of an image.

The quality of edge detection is highly dependent on lighting conditions, the presence of objects of similar intensities, density of edges in the scene.

Since different edge detectors work better under different conditions, the objective of this paper is to identify the suitable edge detector for the IR images with the noise present in it [2]. Moreover, the aim is to find the good edge detector for the filtered images and the respective filter used for the identification of the same.

The reason for choosing IR image is that quality of the IR images differs from the normal images.

The edge detection operators such as the Sobel, Prewitt, Log and Canny are commonly used in the process of the edge detection. The noise taken for this study is Salt and Pepper noise a type of Impulsive noise, where the noise value may be either the minimum or maximum of the dynamic gray scale range of the image.

Common traditional edge extraction algorithms such as Canny use a constant window that can be mixed with some smoothing filters. They need high quality values for their parameters in order to reach extraction efficiency [3], [4]. Despite simplicity, low computational cost and the fact that these parameters are known in high degree of quality based on experiences during last years, but they are still dependent to lightening conditions, noise etc. Lack of any of these dependencies could result in fail of these methods. In addition, using a constant parameter all over the image can result as discontinuity in edges and this discontinuity in edges is one of the most important weaknesses in such algorithms. Some methods try to extract special edges by applying transformations such as Hough transform but all edges don't meet required conditions. Due to lack of information, using hybrid techniques usually leads the process to fail.

The study begins by taking images, and then to add salt and pepper noise with probability 0.1 [5], [6]. Edge detector operators are applied to all the noisy images.

Then, the PSNR value is for these edge detected noisy images. The values show that the Sobel edge detection operator is best for detecting the edge for the noisy images. But it produces less edge detection results for images. In order to overcome the problem of the existing edge detection method and remove noise from image samples in this work initially noise in the image are removed by using the adaptive bilateral filtering (ABF) to remove noise and features of the pixel to detect edges are extracted. Then finally apply edge detection method based on the hybrid fuzzy cognitive map with fuzzy neural network (HFCM-FNN).

The results show that after applying HFCM-FNN produces best edge detection result with high PSNR, less MSE and number of edges detected results.

The organization of this paper is as follows: Section 2 Reviews existing edge detection methods. Section 3 explains the procedure of the proposed edge detection method in detail and Section 4 shows the results of the edge detection method, Section 5 finally concludes the paper.

II. Related Work

This paper proposes a novel edge detection algorithm for noise corrupted images. The algorithm detects the edges by eliminating the noise from the image so that only the correct edges are determined [7].

For making the image noise free, the algorithm calculates closeness parameters and based on those parameter the noisy pixel is replaced by the most appropriate value. The edges of the noise free image are usually determined using morphological operators erosion and dilation. The proposed algorithm uses a combination of these operators to find the edges. Two different types of structuring elements are used in this algorithm to efficiently determine the edges of the images.

Morphological operator [8] is mainly used in this method of edge detection. Morphological edge detector majorly relies on the choice of the structuring element (SE) and therefore the results will vary from one SE to other. Hence, the novel approach to find the SE directly from the image using freeman chain code is adopted which is then followed by the Morphological Gradient method to detect edges. The novel approach based on the shearlet transform is also performed in which a greater ability multiscale directional transform is used to localize distributed discontinuities called edges [9]. Indeed, unlike traditional wavelets, shearlets have the ability to fully capture directional and other geometrical features thus the optimal and theoretical representation of images with edges could be possible.

There are some numerical examples to demonstrate the effectiveness of the shearlet approach for the detection of both the location and orientation of edges

which ultimately implies the significance of shearlet methods as compared to wavelet and other standard methods. Furthermore, the shearlet approach is useful to design simple and effective algorithms for the detection of corners and junctions.

ACO is also introduced in this study to overcome the image edge detection problem [10]. The proposed ACO-based edge detection approach enables establishment of a pheromone matrix for representing the edge information presented at each pixel position of the image based on the movements of a number of ants which are dispatched on the image. The movements of these ants are driven by the local variation of the intensity values of the images. Experimental results are derived to demonstrate the superior performance of the proposed approach.

A multiscale method has also been already proposed to minimize least-squares reconstruction errors and discriminative cost functions under regularization constraints [11]. It helps for edge detection, category-based edge selection and image classification tasks. Experiments on the Berkeley edge detection benchmark and the PASCAL VOC'05 and VOC'07 datasets demonstrated the computational efficiency of the proposed algorithm and its ability to learn local image descriptions that effectively support demanding computer vision tasks.

The task of edge detection has been carried out by assuming the input image to be corrupted by additive zero mean Gaussian noise [12]. A simple edge detection algorithm using first-order gradients is initially applied to exclude structures or details from contributing to the noise variance estimation. Then the use of Laplacian operator followed by an averaging over the whole image will provide very accurate noise variance estimation.

There is only one parameter which is self-determined and adaptive to the image contents whereas others require specific operators. Simulation results show that the proposed algorithm performs well for different types of images over a large range of noise variances. The comparison of performances against all other existing approaches is also done.

This paper proposes combination of Sobel edge detection operator and soft-threshold wavelet de-noising approach for edge detection on images with white Gaussian noises. In many edge detection methods [13], the combination of mean de-noising and Sobel operator based median filtering has been performed. However, Sobel operator cannot remove salt and pepper noise very well. Here, soft-threshold wavelet is been initially used to remove noise followed by the usage of Sobel edge detection operator for edge detection on the image.

This method is mainly used on the images having white Gaussian noises. From the pictures resulted from the experiment, comparison of the images with the traditional edge detection methods has been done and found that the method proposed in this paper has a more obvious effect on edge detection. Also, canny arithmetic operator has been proved to have good detective effect in the common usage of edge detection [14].

Based on the drawbacks of the traditional canny algorithm, an improved canny algorithm is proposed in this paper. In this approach, self-adaptive filter is used to replace the Gaussian filter and morphological thinning is adopted to thin the edge and refining the edge point's detection and the single pixel level edge. The results of experiment showed the marked efficacy of the improved Canny algorithm

III. Proposed Methodology

The work proposes a novel edge detection algorithm for noisy images which constitutes two phases, one for making noise free image noise free and the other for finding edges. The identification of edges is based on the occurrence of changes in the grey level in the regions. In particularly, the salt and pepper noise has been considered in this research for proposed edge detection method is represented in Fig. 1.

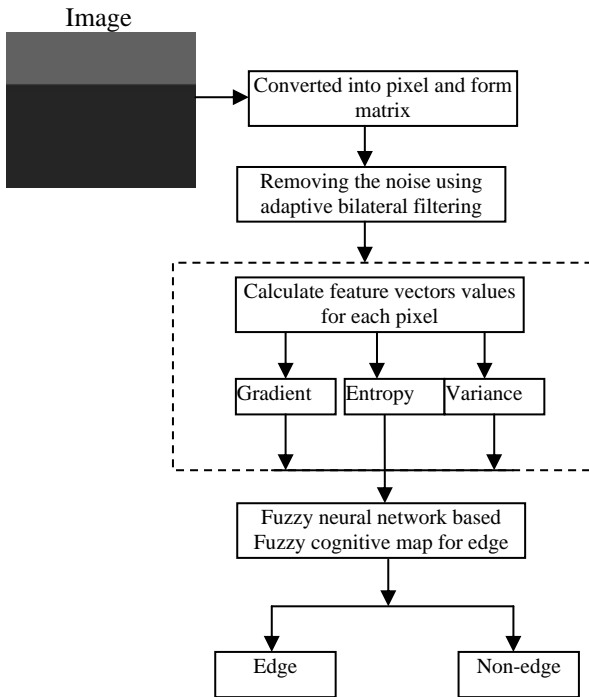


Fig. 1. Flow diagram of proposed edge detection

III.1. Adaptive Bilateral Filtering (ABF)

The main advantage of the classical bilateral filtering method is that it allows for consideration of both the spatial locality and neighboring points with similar amplitudes and time thus helps for preserving the image edges and textures better than the conventional linear filtering algorithms [15].

The Eq. (1) implies the dependence of behavior of the classical bilateral filtering method on the selection of the two parameters $\{\sigma_s, \sigma_I\}$.

However, these two parameters used for controlling the spatial and intensity weight, are constant for the entire image signal.

In other words, the local filters used in classical bilateral filtering method used to change depending on the underlying image signal. The changing degree is usually constrained by a fixed set of parameters $\{\sigma_s, \sigma_I\}$.

There is not much theoretical research reports available on the method of selection of the optimal values for $\{\sigma_s, \sigma_I\}$ in the filtering process.

Therefore, the two empirical parameters are usually selected. The analysis of the two parameters as a function of noise variance for image denoising applications has been studied already [16] and two conclusions about selection of parameters were given: one is the value σ_s should be approximately between $[1.5 \sim 2.1]$ and the other is that the optimal value of σ_I is linearly proportional to the standard deviation of the noise under the sense of mean square error (MSE). The two parameters determined will remain unchanged in the filtering process for the entire image signal.

A typical non-stationary signal which consisting of smooth, edge and texture regions represents the natural image. As each region will have different statistics characteristics, the best image filtering performance can be achieved by either the optimal values of $\{\sigma_s, \sigma_I\}$ or the optimal weights for those that changes based on the local statistics characteristics of the natural image change.

Therefore, spatially adaptive bilateral filtering method is proposed in this study to change the values of $\{\sigma_s, \sigma_I\}$ as per changes in the spatial location during the filtering process. The spatially adaptive bilateral filtering method is represented as follows:

$$\hat{f}(x) = \frac{1}{\hat{C}} \sum_{\xi \in N} W_s(x, \xi) \hat{W}_I(x, \xi, \sigma_n) I(\xi) \quad (1)$$

where the new intensity weight is given by:

$$\hat{W}_I(x, \xi, \sigma_n) = \exp\{-\alpha \tilde{\sigma}_n |I(\xi) - I(x)|^2 / \sigma_n^2\} \quad (2)$$

The new normalization constant:

$$\hat{C} = \sum_{\xi \in N} W_s(x, \xi) \hat{W}_I(x, \xi) \quad (3)$$

The parameter α, σ_n^2 and $\tilde{\sigma}_s$ shown above represents a small constant, the noise variance and the local standard deviations of the image signal respectively. The intensity parameter σ_I in classical bilateral filtering plays a crucial role in the filtering process. For the given spatial parameter σ_s , the largest parameter σ_I will lead to more flatness of Gaussian shape of intensity weight whereas smaller the parameter σ_I will lead to sharpness of the shape of intensity weight. For the natural image signal in the smooth regions, the shape of intensity weight will be expected to be more flat as it could effectively smooth the noise. For the edge and texture regions, the shape of intensity weight is expected to be sharper for the well preservation of the edges and textures.

Beyond the adaptation of the local characteristics of the image signal, the intensity weight also depends on the noise level. When the signal to noise ratio is low, the shape of intensity weight should be more flatness to better suppress the noise.

Therefore, the optimal parameter σ_I should adaptive with the local characteristics of the image signal and the noise level. Considering the computational complexity, we use the local deviation to reflect the local statistical characteristics of the image. Comparing the new intensity weight with the classical intensity weight, it can be seen that the classical intensity parameter σ_I is replaced by:

$$\tilde{\sigma}_I = \sigma_n^2 / \alpha \tilde{\sigma}_s \quad (4)$$

III.2. Features Selected From Image

In this section, we will describe the four different features for image pixels that we use.

Pixel Features

The information of an image pixel can be best obtained by comparing the pixel's features with those of its neighboring pixels. This can be done by extracting 3×3 matrix of the neighboring pixels surrounding the pixel in question. (Other odd-number-sized matrices like 5×5 or 7×7 are also possible.) For any pixel $[x, y]$, its neighborhood matrix contains 9 pixels: $[x-1, y-1], \dots, [x+1, y+1]$.

We use a 3×3 neighborhood matrix for extracting of the features of variance, entropy, and gradient, for each pixel in the image. These attributes hold special properties to determine edge and non-edge pixels of the image. For a grayscale image, the color intensity or grayness of a pixel $[x, y]$ will be denoted as $f(x, y)$. the value of $f(x, y)$ ranges from 0 to 255.

Variance

Statistical properties like mean and variance contain important information about pixels. Variance is a common measure of how far the numbers lie from the mean. Low variance indicates small variation in grayness and high variance means large variation in grayness. So pixel with high variance is candidate to be an edge pixel.

The mean $\mu(x, y)$ of grayness for 3×3 neighborhood matrix centered at the pixel $[x, y]$ is computed as:

$$\mu(x, y) = (1/9) \sum_{i,j=-1}^1 f(x+i, y+j) \quad (5)$$

$$var(x, y) = (1/9) \sum_{i,j=-1}^1 (f(x+i, y+j) - \mu(x, y))^2 \quad (6)$$

Entropy

From information theory, we know that the smaller the local entropy is, the bigger the information

gain and the rate of change in the intensity is. Thus, we can conjecture that the smaller the local entropy is, the bigger the dispersion is.

So, the pixel with big local entropy is more likely to be an edge pixel [17]. For a given pixel $[x, y]$, we take a 3×3 neighborhood matrix with that pixel at the center.

The entropy for the pixel $[x, y]$ is calculated as:

$$entropy(x, y) = - \sum_{i=x-1}^{x+1} \sum_{j=y-1}^{y+1} p_{ij} \log p_{ij} \quad (7)$$

$$p_{ij} = f(x, y) / \sum_{i=x-1}^{x+1} \sum_{j=y-1}^{y+1} f(i, j) \quad (8)$$

Gradient

The gradient is the directional change in grayness of an image. The magnitude of the gradient tells us how quickly the image is changing, while the direction of the gradient tells us the direction in which the image is changing most rapidly. We use the similar gradient measure as the one used in the Sobel method [18].

The gradient $G(x, y)$ for a 3×3 neighborhood matrix centered at the pixel $[x, y]$ is computed as:

$$gradient(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \quad (9)$$

$G_x(x, y)$ is the mask in X direction and $G_y(x, y)$ is the mask in Y direction respectively.

III.3. Edge Detection using Hybrid Fuzzy Cognitive Map based on Fuzzy Neural Network Classification

In this section we present a new way to detect edges by using the hybrid fuzzy cognitive map based on fuzzy neural network [19]. The decision needed in this case is between "the pixel is part of an edge" or "the pixel is not part of an edge". In order to obtain this decision we must extract the information from the images since the entire image is not useful as the input to the hybrid fuzzy cognitive map based on fuzzy neural network (HFCM).

In order to obtain this decision we must extract the information needed from the images. In this work a vector is formed for each pixel given the difference between this one and the pixels in a 3×3 neighborhood around it. This way a components vector such as entropy, gradient, and variance are calculated at each pixel except for the border of the image, because in this case the differences cannot be calculated. This vector is used as input to the HFCM both in the training process and when we use the trained HFCM over real images.

The existing WSVM differs from the new approach in detecting the edge without considering the denoising method or by removing the noise from the image, this work remove the noise from image after feature extraction and edge detection for image processing.

An FCM based FNN method is used which helps in improved performance of edge detection in image processing and further reduce the execution time of the existing methods.

Before performing the edge detection task first important features in the images are extracted as mentioned above, then perform edge detection task with improved FCM with FNN methods. So first study the basic FCM method how it is applicable to the edge detection, the major issue of the FCM method how it is solved by using the FNN for edge detection. FCM methods are a signed directed graph which consists of a set of features which corresponds to same feature for the image represented by nodes and a set of weights represented by directed arcs for features.

The active degree of features f_i is described by state value $s_i \in [0, 1]$, either edge or non edge features which changes over time in inference process.

The value of the weight E_{ij} specifies how strongly the causal feature f_j affects the effect feature f_i . While a positive value of E_{ij} represents a proportional edge detection effect, a negative value represents an inversely proportional to edge detection results effect and zero E_{ij} represents the absence of an edge detection results effect.

A graphical representation of FCM for edge detection is depicted in Fig. 2.

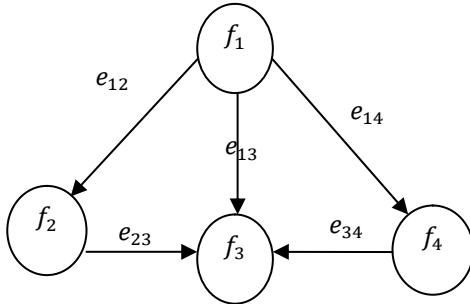


Fig. 2. Flow diagram of proposed edge detection

The inference mechanism of FCM is described by the following formula:

$$\begin{cases} S(t) = (s_i(t))_{1 \times N} \quad i, j = 1, 2, \dots, N \\ s_i(t+1) = f\left(\sum_{j=1}^N s_j(t) \times e_{ji}\right) \quad t = 0, 1, 2, \dots, T \end{cases} \quad (10)$$

where t is the iteration step, and $s_i(t)$ indicates the state value of feature f_i at iteration t . $S(t)$ indicates the system is edge or non edge detection state at iteration, and f is a threshold function. From the Eq. (10), it is inferred that the process of an FCM is an iterative process and will be applied to the scalar product and threshold function.

The discrete-time series of system state could be generated only until the FCM converges yields based on the following:

A fixed point: In this case $S(t_0 + 1) = S(t_0)$ ($t_0 \in T$) where $S(t_0)$ is the state of detection of final edge or non edge. A limit cycle of this case is $S(t_0 + \Delta T) = S(t_0)$ ($t_0 \in T$) in which $S(t_0)$ is the final state. It is meant that the system falls in a loop of a specific period and after a certain number of inference steps ΔT , the same final state $S(t_0)$ will be reached. A chaotic attractor in this case is the change in FCM state vector based on iteration. The states are repeated until there is no detection of edge or non-edge detection states.

FCM Based on OWA Operators

FCM uses simple weighted sum Σ which is very less to represent the edge or non edge detection results for image. Hence, this paper introduces OWA operators instead of FCM, to replace weighted sum Σ and threshold functions f . An n -dimensional OWA operator is a mapping $\phi: R^n \rightarrow R$ that has an associated weighting vector $W = (\omega_1, \omega_2, \dots, \omega_n)$ of having the properties [20]-[21]:

$$\begin{aligned} \omega_1 + \omega_2 + \dots + \omega_n &= 1 \\ 1 \leq \omega_n \leq 1, i &= 1, 2, \dots, n \end{aligned} \quad (11)$$

and such that:

$$\phi(f_1, f_2, \dots, f_n) = \sum_{j=1}^n \omega_j b_j \quad (12)$$

where b_j is the j^{th} largest element of the collection of the aggregated features $\{f_1, f_2, \dots, f_n\}$. The measure of orness of the aggregation is defined as:

$$\text{orness}(W) = \frac{1}{n-1} \sum_{i=1}^n (n-i) \omega_i \quad (13a)$$

By introducing OWA operators to FCM the following reasoning formula were represented as:

$$S_i(t+1) = \phi[e_{1j}(t)f_1(t), \dots, e_{nj}(t)f_n(t)] \quad (13b)$$

From this formula, it is found that OWA-FCM contains two types of weights information: weights e_{ji} for measuring causal link strength and weight vector W for measuring And/Or relationships among data source.

Obtainment of OWA-FCM Weights

The weights of OWA operator ω are determined by the maximum entropy of the edge detection result which formulates the problem of weights of OWA operators based on the solution of the following mathematical programming problem [22]:

$$\max f(\omega) = -\omega_i \ln \omega_i \quad (14)$$

$$\text{s.t. } \frac{1}{n} \sum_{i=1}^n (n-i) \omega_i = \alpha, 0 \leq \alpha \leq 1 \quad (15)$$

From the above method, the weight value is assigned without consideration of the edge result values. In order to overcome this problem and to consider the result based on weight assignment, the weight coefficient e_{ji} is obtained by fuzzy neural network training as described in the following sections.

Fuzzy Neural Network for Improved FCM Learning The Structure of the Proposed Fuzzy Neural Network

To solve the problem of the FCM random assignment of the weight value, this work proposes a fourlayer fuzzy neural network for realizing the automatic identification of weight value for each feature in the images for edge detection as shown in Fig. 3.

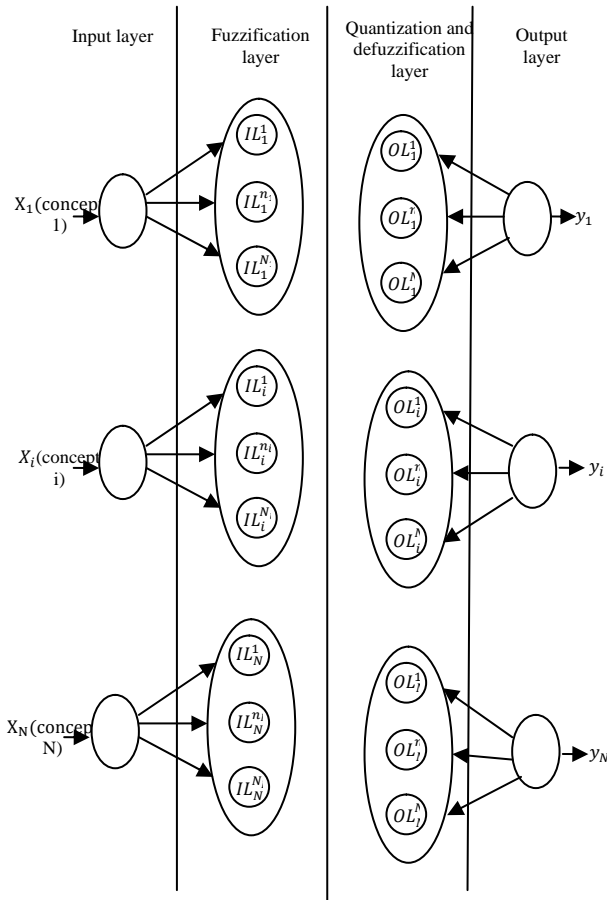


Fig. 3. The representation of the structure of the HFCM –FNN

Input Layer

In this layer, each input node x_i represents a feature f_i for each pixel in the image. Input values $X^T = [x_1, x_2, \dots, x_i, \dots, x_N]$ are directly transmitted to the next layer. So, the net input $f_i^{(1)}$ and net output $x_i^{(1)}$ of the i^{th} node (weight value of the feature for edge detection):

$$f_i^{(1)} = x_i, x_i^{(1)} = f_i^{(1)} \quad (16)$$

Fuzzification Layer

Nodes in this layer represent the linguistic-term set of feature inputs. The n_i th linguistic term of the input

feature variable x_i for pixel it is denoted as $IL_i^{n_i}$ ($n_i = (1, 2, \dots, N_i)$), which is expressed as “small (S),” “medium (M),” or “large (L),” etc. In the proposed FNN, the fuzzy set $IL_i^{n_i}$ is modeled by a symmetric Gaussian-membership with center $C_{IL_i}^{n_i}$ and spread $\sigma_{IL_i}^{n_i}$, which is represented by $IL_i = (C_{IL_i}^{n_i}, \sigma_{IL_i}^{n_i})$. Therefore, the net output $x_{IL_i^{n_i}}^{(2)}$ of the $IL_i^{n_i}$ th linguistic node are given by:

$$x_{IL_i^{n_i}}^{(2)} = e^{f_{IL_i^{n_i}}^{(2)}} = e^{\frac{-(x_i^{(1)} - C_{IL_i}^{n_i})}{\sigma_{IL_i}^{n_i}}} \quad (17)$$

Quantification and Defuzzification Layer

In this layer, the causalities among features and defuzzification are realized. The linguistic term $OL_i^{n_i}$ of output variable y_i is represented in the layer which indicates the same semantic symbol expressed by $IL_i^{n_i}$. $\varepsilon(IL_i^{n_i}, OL_i^{m_j})$ is mutual between $IL_i^{n_i}$ and $OL_i^{m_j}$ to measure the similarity between them. Fuzzy weights $1 - \varepsilon(IL_i^{n_i}, OL_i^{m_j})$ represent the causal effect relationship from input linguistic term $IL_i^{n_i}$ to output linguistic term $OL_i^{m_j}$. Therefore, the net input $f_{OL_i^{m_j}, IL_i^{n_i}}^{(3)}$ and net output $x_{OL_i^{m_j}}^{(3)}$ of the output linguistic term $OL_i^{m_j}$ are expressed:

$$f_{OL_i^{m_j}, IL_i^{n_i}}^{(3)} = x_{IL_i^{n_i}, OL_i^{m_j}}^{(2)} (1 - \varepsilon(IL_i^{n_i}, OL_i^{m_j})) \quad (18)$$

$$x_{OL_i^{m_j}}^{(3)} = \frac{\sum_{i=1}^N (f_{OL_i^{m_j}, IL_i^{n_i}}^{(3)} C_{IL_i^{n_i}} \sigma_{IL_i^{n_i}})}{\sum_{i=1}^N (f_{OL_i^{m_j}, IL_i^{n_i}}^{(3)} \sigma_{IL_i^{n_i}})}, i \neq j, n_i = 1, 2, \dots, N_i \quad (19)$$

Output Layer

In this layer, the net input $f_j^{(4)}$ and net output $x_j^{(4)}$ of the j^{th} output are calculated by:

$$f_j^{(4)} = \sum_{m_j}^{M_j} \xi_{j, m_j} m_j x_{OL_i^{m_j}}^{(3)} \quad (20)$$

$$x_j^{(4)} = y_j = f_j^{(4)} \quad (21)$$

where ξ_{j, m_j} is the crisp weight from the output linguistic term $OL_i^{m_j}$ to the output variable y_i . In order to improve the speed of the FNN in the FCM, in this work presents a Levenberg- Marquart (L-M) Training Algorithm, which combines the gradient falling algorithm with Newton method.

The iterative formula of L-M algorithm is as follows [23]:

$$W_{n+1} = W_n - (J^T J + \mu I)^{-1} J^T e_f \quad (22)$$

in which J is Jacobian matrix of derivatives of network function error, e_f is the error that is calculated as the difference between desired edge detection and estimated edge detection outputs, I is the identity matrix with proper dimensions. The variation of network weights and deviation:

$$\Delta\omega_n = -(J^T J + \mu I)^{-1} J^T e_f \quad (23)$$

IV. Experimental Results

The proposed edge detection method is simulated using MATLAB for different images. It is observed that this proposed method provides much more distinct marked edges as compared to other edge detection algorithms such as SVM and weighted SVM.

The performance metrics used for analyzing the proposed method are Mean square error (MSE), Peak signal to noise ratio (PSNR).

IV.1. Peak Signal to Noise Ratio

The peak signal to noise ratio is represented by the ratio between the maximum possible powers to the power of corrupting noise.

It is also referred as the logarithmic function of peak value of image and mean square error and hence represented as:

$$PSNR = 10 \log_{10}(MAX_i^2 / MSE) \quad (24)$$

IV.2. Mean Square Error

Mean square error (MSE) of an estimator is to quantify the difference between an estimator and the true value of the quantity being estimated:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2 \quad (25)$$

TABLE I
PERFORMANCE COMPARISON

Methodology	MSE	PSNR	Number of edges detected
SVM	0.3259	34.258	989
WSVM	0.2369	37.259	2045
HFCM-FNN	0.2145	40.256	3125

The graphical representation of comparison of MSE and PSNR of the proposed method with the existing methodologies such as cluster algorithms, SVM and WSVM has been represented graphically in Fig. 4 and Fig. 5. The results inferred suggest that the proposed method removes noise from image samples using ABF.

Hence, the performance of the proposed methodology using HFCM-FNN is found to be far better than the existing edge detection algorithms such as clustering, WSVM and SVM.

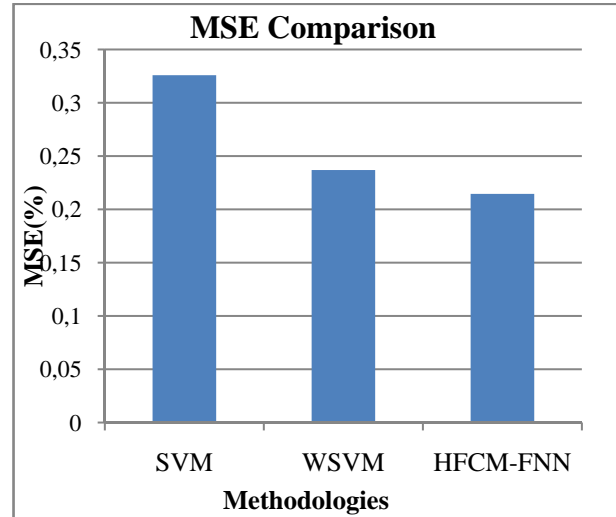


Fig. 4. Comparison of MSE of various edge detection methods with HFCM-FNN

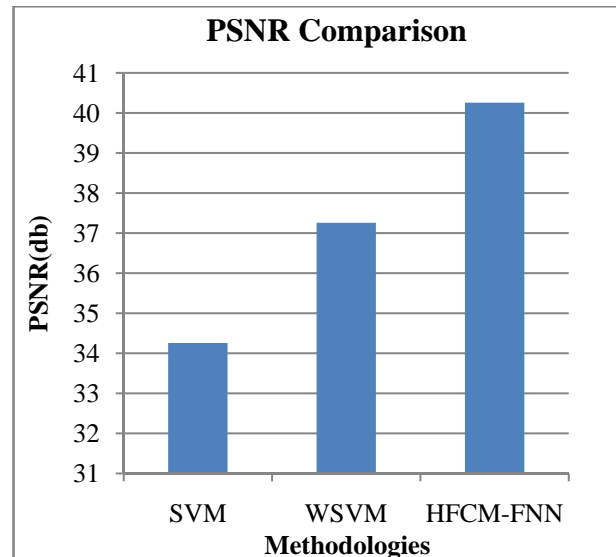


Fig. 5. Comparison of PSNR of various edge detection methods with HFCM-FNN

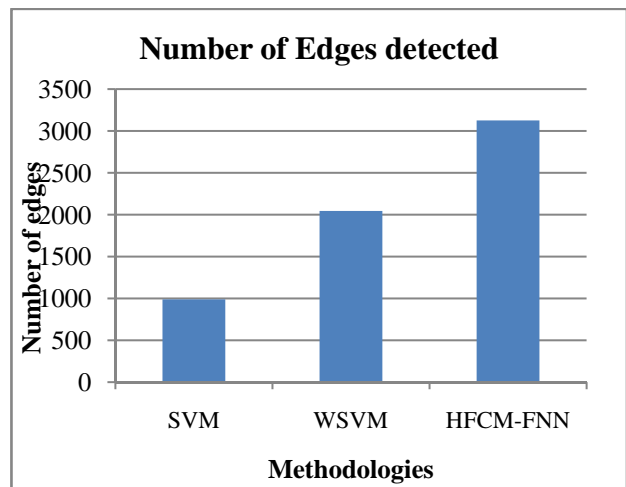


Fig. 6. Comparison of number of edges between various edge detection methods

The number of edges detected by proposed method has also been compared with existing methodologies such as cluster algorithms, WSVM, SVM and represented graphically in Fig. 6. The performance of the proposed methodology using HFCM-FNN is better than the existing edge detection algorithms such as clustering, SVM and WSVM.

A ground-truth image helps for deciding whether an edge is truly present or absent. The results of the edge detection algorithm will further decide the extent of detection of edge. The results of the low level pixel based comparison between the ground truth and the number of images of the images based on the following values is mentioned in the Table II.

TABLE II
VARIABLE DEFINITION

Parameters	Edge	Non edge
Edge	True Positive (TP)	False Positive (FP)
Non edge	False Negative (FN)	True Negative(TN)

From these measures the following measures were used to measure the result of the edge detection methods:

$$\text{Sensitivity} = \frac{TP}{TP + TN} \quad (26)$$

$$\text{Specificity} = \frac{FN}{FN + FP} \quad (27)$$

The sensitivity results of the proposed HFCM-FNN have also been compared with the edge detection methods such as WSVM and SVM. The graphical representation of comparison of sensitivity results is shown in Fig. 7. The results showed that the proposed HFCM-FNN is less sensitive than the existing edge detection methods such as WSVM and SVM. This is because of the ABF applied to this edge detection method thus resulting in high edge detection results.

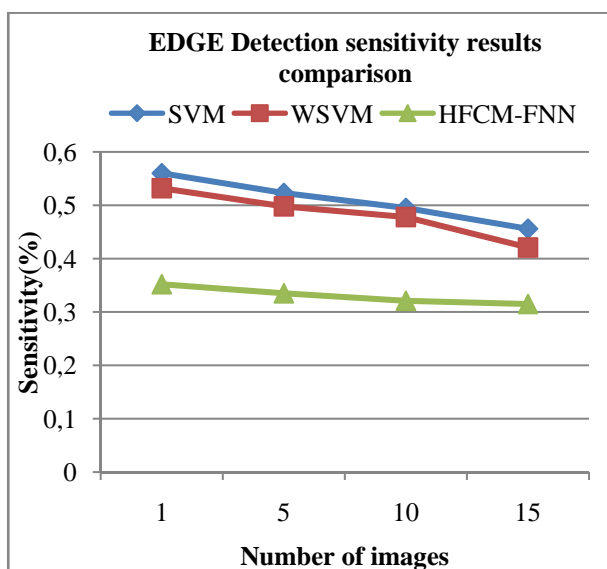


Fig. 7. Sensitivity results comparison for edge detection methods

The specificity results of the proposed HFCM-FNN are higher than the existing edge detection methods such as WSVM and SVM. The reason is also attributed to the application of ABF and the higher performance of the edge detection results of the proposed system. Fig. 8 shows the graphical representation of comparison of specificity results with other edge detection methods such as HFCM-FNN, WSVM and SVM.

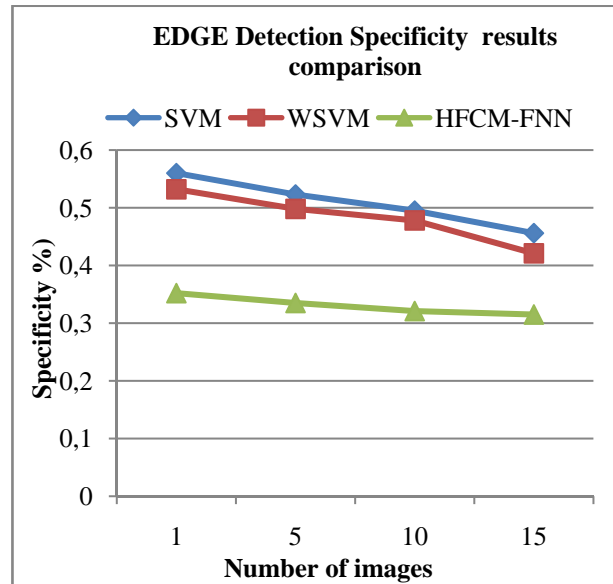


Fig. 8. Specificity results comparison for edge detection methods

V. Conclusion

The improved edge detection method operating on Hybrid Fuzzy Cognitive Map based Fuzzy Neural Network(HFCM-FNN) was presented in this work to perform the pixel classification between edge and no edge. The disadvantages of the edge detection based on SVM and WSVM such as low efficiency, lack of transparency of results have been overcome by this method. The experimental results obtained by Hybrid Fuzzy Cognitive Map based Fuzzy Neural Network (HFCM-FNN) method of edge detection were proved to be efficient. The comparison of the proposed HFCM-FNN edge detection with existing edge detection algorithms such as SVM and WSVM confirmed the better performance of HFCM-FNN algorithm.

References

- [1] Chongyang hao, Min qi ,U. Heute and C. Moraga ,New method for fast image edge detection based on Subband decomposition, *Image Anal Stereol* ,Vol.20, pp. 53-57,2001.
- [2] G. Padmavathi, P. Subashini, P. K. Lavanya, *Performance evaluation of the various edge detectors and filters for the noisy IR images*, Proceedings of the 2nd WSEAS International Conference on Sensors, and Signals and Visualization (Imaging and Simulation and Materials Science, ISSN: 1790-5117 Pages 199-203).
- [3] M. Davoodianidaliki , A. Abedini , M. Shankayi Adaptive Edge Detection Using Adjusted Ant Colony Optimization, *International Archives of the Photogrammetry, Remote Sensing and Spatial*

- Information Sciences*, Vol. XL-1/W3, pp.123-126, 2013.
- [4] Ehsan Nadernejad, Sara Sharifzadeh, Hamid Hassanpour, Edge Detection Techniques: Evaluations and Comparisons, *Applied Mathematical Sciences*, Vol. 2, No. 31, pp.1507 – 1520, 2008.
 - [5] Peter Meer, and Bogdan Georgescu, Edge Detection with Embedded Confidence, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 12, pp.1351-1365, December 2001.
 - [6] Mamta Juneja , Parvinder Singh Sandhu ,Performance Evaluation of Edge Detection Techniques for Images in Spatial Domain *International Journal of Computer Theory and Engineering*, Vol. 1, No. 5, December, 2009 pp.1793-8201.
 - [7] Akansha Mehrotra, Krishna Kant Singh and M.J.Nigam. Article: A Novel Algorithm for Impulse Noise Removal and Edge Detection. *International Journal of Computer Applications* Vol.38, No.7, pp.30-34, January 2012. Published by Foundation of Computer Science, New York, USA.
 - [8] Ratika Pradhan, Prasanna Pradhan, Reshmi Bhattacharjee, Divyendra Singh ,Edge Detection using Morphological Operator: A New Approach , *International Journal of Advanced Research in Computer Science and Software Engineering* ,Vol.4, No.2, pp. pp. 84 -88 ,February 2014.
 - [9] Sheng Yi, Labate, D. ; Easley, G.R. ; Krim, H ,A Shearlet Approach to Edge Analysis and Detection, *Image Processing, IEEE Transactions on* , Vol.18 , No.5 , pp. 929 – 941,2009.
 - [10] Jing Tian ; Weiyu Yu ; Shengli Xie ,An ant colony optimization algorithm for image edge detection, *IEEE Congress on Evolutionary Computation*, pp. 751 – 756, 2008.
 - [11] Julien Mairal, Marius Leordeanu, Francis Bach, Martial Hebert, Jean Ponce ,*Discriminative Sparse Image Models for Class-Specific Edge Detection and Image Interpretation* (Computer Vision – ECCV 2008 Lecture Notes in Computer Science Vol.5304, pp 43-56,2008).
 - [12] Shen-Chuan Tai ; Shih-Ming Yang ,*A fast method for image noise estimation using Laplacian operator and adaptive edge detection*, 3rd International Symposium on Communications, Control and Signal Processing (pp. 1077 – 1081,2008. ISCCSP 2008).
 - [13] Wenshuo Gao ; Xiaoguang Zhang ; Lei Yang ; Huizhong Liu ,*An improved Sobel edge detection*, 3rd IEEE International Conference on Computer Science and Information Technology (Vol.5 ,pp. 67 - 71 ,2010).
 - [14] Bing Wang ; DeptShaosheng Fan ,*An Improved CANNY Edge Detection Algorithm* ,Second International Workshop on Computer Science and Engineering (Vol. 1, pp. 497 – 500, 2009. WCSE '09).
 - [15] Qi Min, Zhou Zuofeng , Liu Jing3,c , Cao Jianzhong , Wang Hao, Yan Aqi , Wu Dengshan , Zhang Hui and Tang Linao, *Image denoising algorithm via spatially adaptive bilateral filtering* ,Proceedings of the 2nd International Conference On Systems Engineering and Modeling (ICSEM-13),pp.0431-435.
 - [16] M. Zhang, B.K. Gunturk, Multiresolution Bilateral Filtering for Image Denoising, *IEEE Trans. Image Process.* Vol.17, , pp. 2324–2333,2008.
 - [17] W. Dai and K. Wang, An image edge detection algorithm based on local entropy, in Proceedings of the IEEE International Conference on Integration Technology (ICIT'07) (pp. 418–420,2007).
 - [18] I. E. Sobel, Camera Models and Machine Perception, Ph.D. Thesis, Electrical Engineering Department, Stanford University, California, United States, 1970.
 - [19] Huaibin Wang , Li Wang ,Application of Improved Fuzzy Cognitive Map Based on Fuzzy Neural Network in Intrusion Detection , *Journal of Information & Computational Science*, Vol.10,No. 1,pp.271–278, 2013.
 - [20] JianWu, Qingwei Cao, An OWA operator based approach to aggregate group opinion by similarity degree, Fourth International Conference on Business Intelligence and Financial Engineering (pp.665-667,2011).
 - [21] Liu, H., Liang, L., Zhang, L., Submarine hydraulic rudder control system based on fuzzy logic algorithm, (2012) *International Review on Computers and Software (IRECOS)*, 7 (5), pp. 2367-2371.
 - [22] R. Fuller, *On obtaining OWA operator weights: A sort survey of recent developments*, International Conference on Computational Cybernetics (pp.241-244, 2007).
 - [23] Wenqiang Xiang, Hua Zhang, Heng Wang, Application of BP neural network with L-M algorithm in power transformer fault diagnosis, *Applied Mechanics and Materials*, Power System Protection and Control, Vol.39 , pp.100-104,2011.

Authors' information

¹Associate Professor, P.S.G. College of Arts and Science, Coimbatore.
E-mail: t.karthikeyan.gasc@gmail.com

²Research scholar, Karpagam University, Coimbatore.
E-mail: revathynp@yahoo.com



N. P. Revathy, She is a Phd Research Scholar in Karpagam University, Coimbatore, Tamilnadu, India. She is working in Karpagam University, Coimbatore. Interest areas are Data Mining, Image Processing.



Mr. Thirunavu Karthikeyan received his graduate degree in Mathematics from Madras University in 1982. Post graduate degree in Applied Mathematics from Bharathidasan University in 1984. Received Ph.D., in Computer Science from Bharathiar University in 2009. Presently he is working as a Associate Professor in Computer Science Department of P.S.G. College of Arts and Science, Coimbatore. His research interests are Image Coding, Medical Image Processing and Datamining. He has published many papers in national and international conferences and journals. He has completed many funded projects with excellent comments. He has contributed as a program committee member for a number of international conferences. He is the review board member of various reputed journals. He is a board of studies member for various autonomous institutions and universities.
E-mail: t.karthikeyan.gasc@gmail.com

Adaptive Resource Allocation Mechanism (ARM) for Efficient Load Balancing in WiMAX Network

P. Kavitha¹, R. Uma Rani²

Abstract – Today in the wireless network field, WiMAX (Worldwide Interoperability for Microwave access) has emerged out as one of the promising networking technologies. In order to compete with the existing wireless technologies like Wi-Fi and Bluetooth (IEEE 802.15), WiMAX has to promise cost efficiency and better quality of service (QoS). The Adaptive Resource Allocation Mechanism (ARM) is proposed to control traffic rate and ensure load balancing (LB) in the WiMAX network. The proposed approach considers relay station (RS) in the network. When there is an arrival of new users in the network, its data rate is computed and then compared with the data rate of RS. When the data rate of new user is less than that of RS, then the data rate of the corresponding base station (BS) is compared. The connections are switched from congested stations to non-congested stations to minimize network load. The handover mechanism is used by BSs for optimally balancing the traffic load within the network. LB-based handover mechanism guarantee users are QoS and evenly distribute the traffic load. The experimental analysis showed that the proposed method achieves better traffic management and load balancing when compared with the existing approaches. **Copyright** © 2014 Praise Worthy Prize S.r.l. - All rights reserved.

Keywords: Adaptive Resource Allocation, Handover Mechanism, IEEE 802.16J, Load Balancing, Traffic Control, WiMAX

I. Introduction

WiMAX supports the technologies that make triple-play service offerings possible such as multicasting and quality of service [1]-[25]. These are basic to the WiMAX standard. As a standard determined to satisfy the requirements of next-generation data networks (4G), WiMAX is differentiated by its dynamic burst algorithm modulation adaptive to the physical environment the RF signal travels through. One of the important benefits of advanced wireless systems like WiMAX is the spectral efficiency. The significant advantage of WiMAX comes from integrating SOFDM (orthogonal frequency division multiple access) with smart antenna technologies. This enhances the efficient spectral efficiency through smart network deployment topologies and multiple reuse. Smart antennas also known as adaptive array antennas are antenna arrays with smart signal processing techniques utilized to determine spatial signal signature, such as DOA (direction of arrival) of the signal. Like all wireless technologies, WiMAX can work at higher bit rates or over longer distances but not both. The bit error rate is increased when operated at the maximum range of 50 km. This results in a much lower bitrate. WiMAX is a long range system, covering many kilometers. The licensed or unlicensed spectrum is used to deliver connection to a network. The QoS mechanism is used depends on connections between the base station (BS) and the user device. Each connection depends on specific scheduling algorithms.

WiMAX provides mobile or at-home internet access across whole cities or countries. Last-mile broadband internet access can be provided in remote locations.

Multiple levels of QoS and flexible channel bandwidths support allows WiMAX to be used by service providers for high-bandwidth and low-latency entertainment applications. For e.g. WiMAX could be entrenched into a portable gaming device for usage in a fixed environment for interactive gaming. The advantages of WiMAX include:

- Coverage of large areas
- High data rate support
- Flexible and dynamic QoS support
- Multiple services support with different QoS policies
- Better support for NLOS (Non-line-of-sight) technologies
- Good spectral efficiency
- Inexpensive and fast deployment of “last mile access” to public networking
- Corporate-grade security
- Cost-effective alternative to 3G/4G cellular networks and WiFi

High data rate is one of the desired features of the WiMAX network. It needs a highly effective usage of the available spectrum. Fixed RS has fewer functionalities than base stations. RS is used to overcome poor channel conditions and to maintain low infrastructure cost. RSs always uses the same spectrum as (mobile station) MSs and BSs.

Too many users accessing one station (RS or BS) leads to load imbalance. To guarantee the user's quality of service requirements, LB should be adopted.

The paper proposes an ARM method for better traffic management and load balancing in WiMAX network. The rest of the paper is organized as follows. Section II presents a description about the previous research in resource allocation schemes. Section III involves the detailed description about the proposed ARM technique. Section IV presents the experimental results of the proposed system. Section V involves conclusion and future enhancements.

II. Related Work

The low-overhead scheduling algorithms for WiMAX uplink scheduling [1] performs the adaptive prediction of user's load and choose a small set of active users to be served. The novel techniques like piggybacking are used to minimize the MAC (media access control) layer overhead. The approach provides QoS guarantees and also reduce the scheduling overhead. An effective ALA (Adaptive Load-balancing Association handoff approach) [2] is used for enhancing utilization and improving QoS in WiMAX networks.

The approach consists of two phases: AAH (Adaptive Association Handoff phase) and PDLB (Predictive Direction-based Load Balancing phase). A re-association mechanism is used to minimize lost synchronizations and to enhance the grade of service. The moving direction of mobile nodes is correctly predicted by the polynomial regression-based RSS prediction algorithm. The On Demand Bandwidth Allocation (ODBA) approach is used for distribution of bandwidth [3] in WiMAX network.

The approach consists of a management module in SS (Subscriber Station) for managing UL (uplink) bandwidth. A new module is present in BS. The ODBA approach minimizes the queuing delay and enhances the throughput of a WiMAX network. The end-to-end QoS adjustment algorithm [4] is used for dynamically adjusting the L2 bandwidth allocated in the WiMAX mesh and PMP networks and the L3 data rate in the internet after the handoff.

A CBQ + RED QoS adaptation algorithm provides the effective queue and traffic management mechanisms as each packet arrives. The results show that the approach supports lower packet loss ratios, higher throughputs and smaller delays on the end-to-end path under distinct parameter combinations and two handoff scenarios. A two-stage mechanism known as PSBA (packet scheduling and bandwidth allocation) [5] depends on the channel quality information.

The approach enhances system throughput and reduces packet delay in WiMAX network. A Markov chain with bulk service was built for analyzing the WiMAX Point-to-Multipoint (PMP) network based on the channel quality by utilizing BW-REQ (bandwidth request message).

The results showed that PSBA achieved higher throughput and lower delay time among multiple users. An adaptive call admission control [6] is used to effectively enhance the system capacity. The control mechanism can admit an incoming VOD (video on demand) connection establishment without adequate idle bandwidth/slots to meet the QoS requirement. Each connection maintains its fluency of real-time video performance over dynamic wireless channel status by employing AMC (adaptive modulation and coding) scheme.

The concurrent stream connection number and connection blocking probability are evaluated with distinct resource management strategies. An adaptive scheduling packets algorithm [7] for the uplink traffic in WiMAX network is designed to be fully dynamic. A new deadline-based scheme aims at limiting the maximum delay to the real-time applications by using the states of the uplink virtual queues at the base station and a cross-layer approach. The approach was evaluated by modeling and simulation in environments where different MCSs (modulation and coding schemes) were utilized. A flow admission control and scheduling scheme for multihop WiMAX networks [8] assures that the different QoS parameters for the 802.16 service classes are accomplished. The results showed that the approach ensures the maximum allowable packet delay, minimum bandwidth and maximum allowable packet jitter. The multiple downlink fair packet scheduling scheme [9] satisfies both throughput and delay guarantee to different real and non-real applications corresponding to various scheduling schemes.

The technique achieves tight QoS guarantee in terms of delay for all traffic types as described in WiMAX standards. The process maintains the fairness of the allocation and helps to avoid starvation of lower priority class services. A new effective and generalized scheduling schemes for IEEE 802.16 broadband wireless access system is proposed that reflects the delay requirements. The greedy scheduling algorithm (GSA) [10] enhances the utilization of the available WiMAX radio resources at the minimum computational cost. The approach also fulfills a wide range of requirements depending on network environment specifics and operator's preferences.

The results show that GSA accomplishes better performance in terms of efficiency, computational load and interference mitigation. The WiMAX mesh network can operate in a multihop mode in which SS communicate with the base station without any direct link between them. The allocation of channel for SSs is a major issue in such type of networks. A dynamic programming (DP) algorithm [11] determines the conflict-free set of nodes that can be activated to accomplish optimality in throughput.

A genetic algorithm that is more scalable than the DP approach is also proposed. The performance evaluation show that the algorithms are more efficient than the existing approaches.

Two allocation methods: Adaptive Slot Allocation (ASA) and Reservation-Based Slot Allocation (RSA) [12] performs fair resource allocation among various types of service flows (SFs) by considering their channel qualities and QoS characteristics.

The main aim of the approach is to enhance the capacity of the system subject to QoS constraints for each type of SFs in terms of BER (Bit Error Rate) and data rate. The joint routing and scheduling in WiMAX-based mesh network [13] determines minimum schedule period satisfying a given uplink and downlink traffic demand. The proposed maximum spatial reuse (MSR) model assumes centralized scheduling at BS and attempt to maximize the system throughput through suitable routing tree selection. The approach also accomplishes effective spectrum reuse by opportunistic link scheduling. The problem is decomposed using a CG (column generation) approach. Two formulations for modeling MSR namely path-based (CGPath) and link-based (CGLink) formulation are presented. The experimental results show that the path-based formulation requires minimum computational time than the link-based for determining the optimal solution.

The swapping min-max (SWIM) technique [14] is used for UGS (unsolicited grant service) scheduling. The approach meets the delay constraint with optimal throughput and reduces burst overhead and delay jitter. Call admission control (CAC) and packet scheduling are the two significant issues to be considered in assuring QoS requirement. The link aware dual partitioning, call admission control (LADP CAC) approach [15] includes dual partition of the bandwidth for call admission control and priority earliest deadline for packet scheduling.

The approach achieves high throughput with maximum link utilization. A new framework [16] solves the QoS issues for fixed PMP (point to multipoint) 802.16 systems. The approach consists of a CAC module and uplink scheduler.

The CAC module interact with the uplink scheduler status and makes its decision depending on the scheduler's queues status. The fractional frequency reuse (FFR) approach [17] is used for hierarchical resource allocation in WiMAX networks. The approach guarantees quality of service of distinct service flows in the system.

The architecture coordinates the resource allocation in terms of slots between RRA (Radio Resource Agent) and RRC (Radio Resource Controller). The three types of diversity, namely, traffic diversity, mutual interference diversity, and selective fading channel diversity are captured.

The approach enhances the overall throughput of the system and guarantees QoS requirements for a mixture of real-time and non-real-time service flows under various diversity configurations.

The cross-layer paradigm [18] is used for improving the performance of WiMAX. A novel routing-scheduling scheme to demonstrate the application and realization of the cross-layer paradigm is presented.

The relay station (RS) placement [19] on IEEE 802.16j network performance is analyzed.

An effective near-optimal placement solution for IEEE 802.16j WiMAX networks is proposed. The throughput performance shows that the approach tremendously enhance the IEEE 802.16j network capacity. A novel network design and optimization model for WiMAX networks [20] uses multi-hop relays. The approach determines the optimal locations of BSs and RSs so that the network can guarantee QoS in terms of access data rate.

The results show that the technique achieves better network service coverage when compared with the existing models.

III. Proposed Method

The proposed ARM approach focuses on switching the connections from congested stations to non-congested stations. The available frequency resource is improved for the congested stations to accomplish LB. The link qualities between RSs and MSs as well as the traffic load information of RSs are reported to BS by RSs. The BS is responsible for accomplishing handover mechanisms in each sector.

Handover mechanism maintains uninterrupted user communication session during a user's movement from one location to another. The flow of the proposed approach is shown in Fig. 1.

When there is a new arrival in the network, its data rate is checked and compared with the data rate of RS. If the data rate of the new user is less than the data rate of RS, traffic load is high at RS. When an RS is overloaded, it does not have adequate frequency resource for the users nearby.

The request then goes to BS. The RS and BS data rate is compared. When the data rate of BS is high, some users originally associated with the RS switch their serving station from the congested RS to BS in order to balance the traffic load and minimize the blocking probability. The low transmission power at RS limits the coverage area of RS. So, the users associated with one RS cannot establish connections with other RS in the same sector. The main benefit of the handover mechanism is the replacement of two-hop transmission with one-hop transmission. This conserves more resource for the rest of the users associated with RSs.

But, BS cannot always cover the place where a new arrival arises, and the optimal value of the throughput cannot be acquired by the handover of the new user either. So, the BS computes the expected data rate if a user is switched to BS. The user that can accomplish the largest benefit by switching the serving station is selected. If the RS is still overloaded, the next acceptable user will be chosen to connect to the BS till the RS is no longer overloaded. The objective of the handover mechanism is the redistribution of partial traffic load from RS to BS to minimize the heavy traffic load of the RS. When the data rate of BS is low when compared

with RS, traffic is overloaded at BS. So, some users are switched from the current BS to non-congested BS.

Let U_{min} be the user's minimum traffic rate under QoS requirements. $U_l^m(T_W)$ denote the average data rate for user m over fixed-length time window T_W . The user's QoS cannot be satisfied and new user's will be blocked on link l if $U_l^m(T_W)$ is less than U_{min} . In this case, there is no more available resource to guarantee QoS and traffic is overloaded.

The transmission power for users in current BS is denoted as S_B^{cur} . The transmission power for users in the next non-congested BS is S_B^{nx} .

Algorithm: Handover mechanism

Initialization: $p = 0$, $U_{cur} = 0$, $U_{nx} = 0$, $I[\]$, $O[\]$

Handover Execution:

```

while  $U_l^m(T_W) < U_{min}$  do
   $p = p + 0.1$ ,  $S_B^{nx} = S_B$ ,  $S_B^{cur} = p \cdot S_B^{nx}$ 
  for  $m \in M_l$  do
    if  $\Gamma_{bz-an}^{cur} \geq Z_{th} [64QAM (5/6)]$  then
       $i(1, m) = 1$ ,  $o(1, m) = 0$ 
    else
       $i(1, m) = 0$ ,  $o(1, m) = 1$ 
    end if
  end for
  for  $u = 1 : U_Z$  do
    if  $\max_{m \in N_{cur}} \{ N_l^m(u) \} > \max_{m \in N_{nx}} \{ N_l^m(u) \}$ 
       $U_{cur} = U_{cur} + 1$ , and update  $N_l^m(u)$ ,  $m \in N_{cur}$ 
    else
       $U_{nx} = U_{nx} + 1$ , and update  $N_l^m(u)$ ,  $m \in N_{nx}$ 
    end if
    if  $U_{cur} + 3U_{nx} > U_Z$  then
      break
    end if
  end for
  if  $T(p) < T(p - 0.1)$  and  $S_{cov} < 95\%$  then
     $p = p - 0.1$ 
    break
  end if
end while

```

The power ratio of the transmission power for the current BS users to the transmission power for non-congested BS users is defined as $p = S_B^{cur} / S_B^{nx}$.

The power ratio p varies from 0 to 1. The BS periodically measures and computes $U_l^m(T_W)$ of the users associated with BS. Whenever new user accesses BS, $U_l^m(T_W)$ is compared with U_{min} by BS to determine whether it is overloaded or not. If the BS is overloaded due to the arrival of new user, the handover technique is carried out to prevent traffic load. The handover algorithm is shown above.

When the BS is overloaded, p is increased according to actual load status. The modulation and coding scheme (MCS) level is determined by the signal to interference plus noise ratio (SINR) of a link. Γ denotes SINR of a link. S_{cov} is the percentage of area that received SINR above the threshold.

S_B is the transmission power of BS. $T(p)$ is the throughput with power ratio at p . Let M_l be the set of users on link l . $O[\]$ is the M_l column of vector of users in non-congested BS and $o(1, m)$ is the m th element of $O[\]$. $I[\]$ is the M_l column of vector of users in the congested BS and $i(1, m)$ is the m th element of $I[\]$.

When BS adopts transmission power S_B^{nx} , the users associated with BS with high MCS are switched to non-congested BS.

Let Γ_{bz-an}^{cur} denote SINR when the transmission power of BS is S_B^{cur} . The user m is allocated to the non-congested BS if Γ_{bz-an}^{cur} conforms to Eq. (1):

$$\Gamma_{bz-an}^{cur} \geq Z_{th} [64QAM (5/6)] \quad (1)$$

where $Z_{th} [64QAM (5/6)]$ denotes the SINR threshold of the 64QAM (5/6) modulation. U_{cur} and U_{nx} be the number of slots allocated to congested BS and non-congested BS, respectively. U_{cur} and U_{nx} should conform to total slots constraint U_Z , i.e., $U_{cur} + 3U_{nx} = U_Z$. N_{cur} and N_{nx} represent sets of the users in the current congested BS and the next non-congested BS, respectively. $N_l^m(u)$ is the PF (proportional fair) metric.

The round robin (RR) scheduling and analysis is then performed.

IV. Performance Analysis

The performance evaluation of the proposed adaptive resource allocation mechanism (ARM) is discussed in this section. The proposed algorithm is compared with the greedy approach, Bounded Greedy Weighted Algorithm (BGWA), Call Admission Control (CAC), Joint Scheduling and Resource Allocation (JSRA) and Coverage Based Cell Selection (CBCS).

Fig. 2 shows the comparison graph of network throughput for different amount of bandwidth for the greedy approach, BGWA, CAC, JSRA, CBCS and the proposed ARM approach. When the amount of bandwidth increases, the network throughput also increases and the curves of ARM are much higher than all other existing approaches.

Fig. 3 shows that the comparison graph of network throughput with bandwidth consumption ratio for different amount of bandwidth. The results show that the proposed approach achieves optimal n/w throughput to BW consumption ratio for various amount of bandwidth.

Fig. 4 shows the comparison graph of network throughput for different number of subscriber station. The results show that the proposed ARM approach achieves higher network throughput than all other existing approaches for various number of subscriber stations. The ratio of network throughput to bandwidth consumption as a function of the number of SSs for existing approaches and the proposed ARM method in shown in the Fig. 5. The results show that the ARM approach attains higher optimal network throughput to BW consumption ratio when compared with other existing methods.

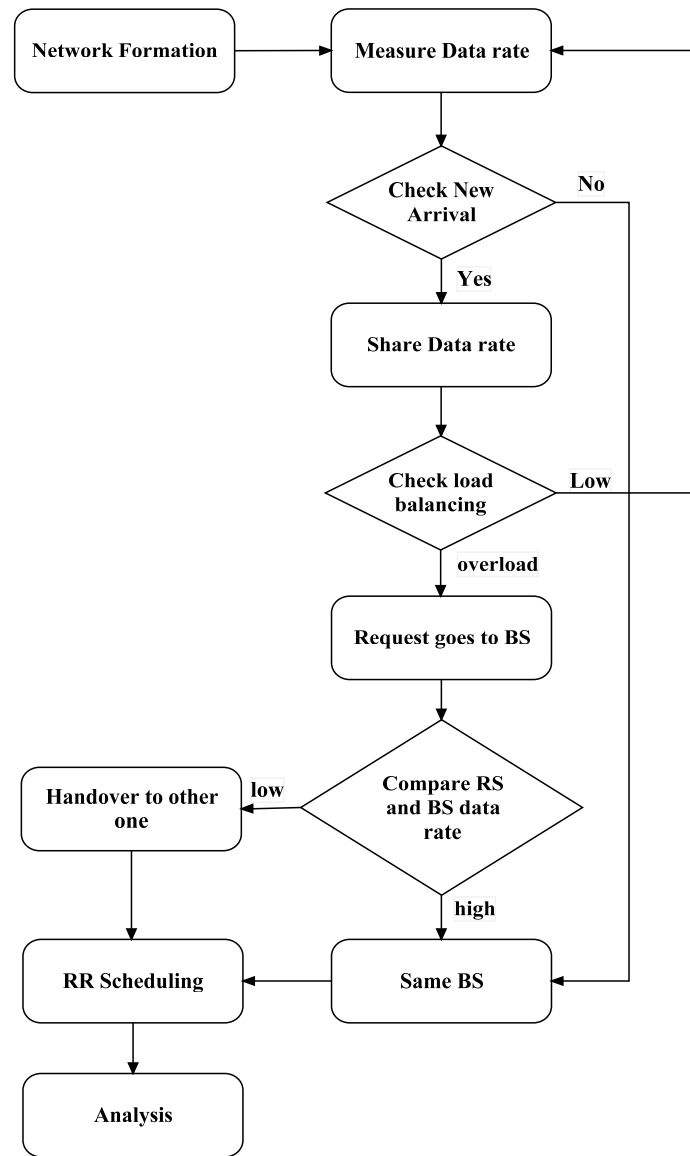


Fig. 1. Flow diagram for proposed adaptive resource allocation mechanism (ARM)

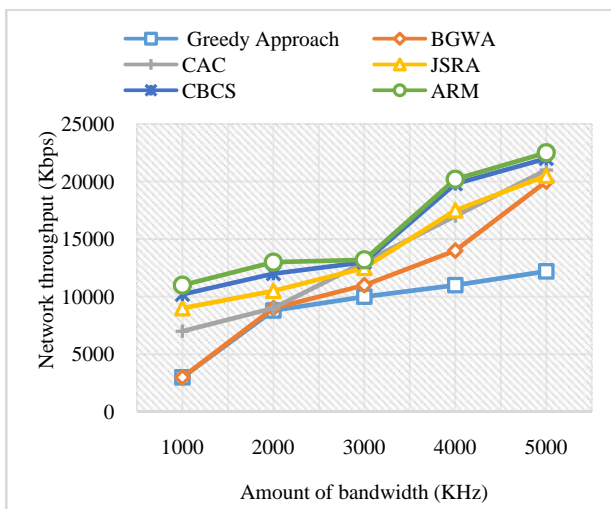


Fig. 2. Network throughput for different amount of bandwidth

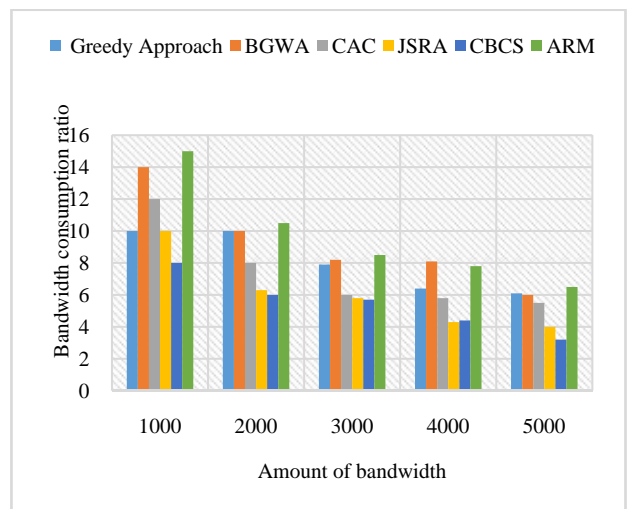


Fig. 3. Network throughput to bandwidth consumption ratio for various amount of bandwidth

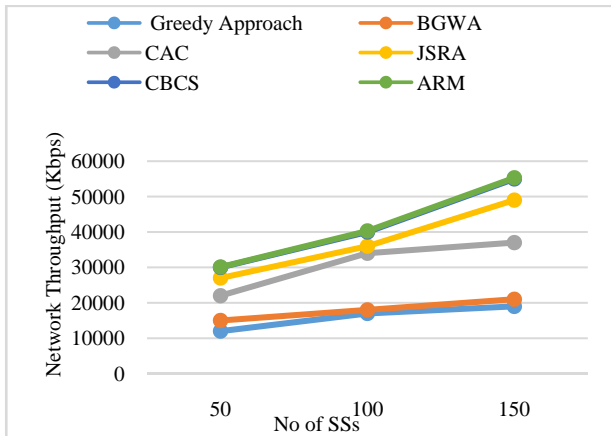


Fig. 4. Network throughput for different number of SSs

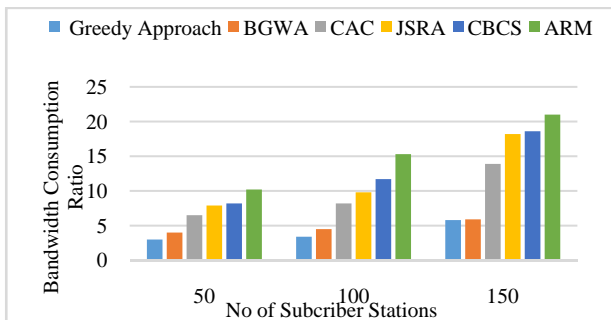


Fig. 5. Network throughput-to-bandwidth consumption ratio for different number of SSs.

V. Conclusion and Future Work

This paper proposes Adaptive Resource Allocation Mechanism (ARM) to accomplish load balancing and to control traffic in the WiMAX network. In the proposed approach, the data rate of the user is considered. The connections are switched from congested stations to non-congested stations to achieve load balancing. The proposed approach introduces RS in the network.

The experimental results showed that the proposed approach achieves optimal network throughput for different number of subscriber station and better traffic management when compared with the existing approaches such as greedy, BGWA (Bounded Greedy Weighted Algorithm), Call Admission Control (CAC), Joint Scheduling and Resource Allocation (JSRA), and Coverage Based Cell Selection (CBCS). As a future work, channel estimation is considered using the Cell-Degree based Resource Allocation (CBRA) scheme to ensure user fairness.

References

- [1] W. Nie, H. Wang, and N. Xiong, "Low-overhead uplink scheduling through load prediction for WiMAX real-time services," *IET communications*, vol. 5, pp. 1060-1067, 2011.
- [2] R. H. Hwang, B. J. Chang, Y. M. Lin, and Y. H. Liang, "Adaptive load-balancing association handoff approach for increasing utilization and improving GoS in mobile WiMAX networks," *Wireless Communications and Mobile Computing*, vol. 12, pp. 1251-1265, 2012.

- [3] Z. Sun and A. Gani, "Evaluating of on demand bandwidth allocation mechanism for point-to-multipoint mode in WiMAX," in *Information Computing and Applications*, ed: Springer., pp. 16-23, 2010.
- [4] I.-C. Chang and Y.-T. Mai, "The end-to-end QoS guarantee framework for interworking WiMAX PMP and mesh networks with Internet," *Computers & Electrical Engineering*, vol. 39, pp. 1905-1934, 2013.
- [5] F.-M. Yang, W.-M. Chen, and J.-L. C. Wu, "A dynamic strategy for packet scheduling and bandwidth allocation based on channel quality in IEEE 802.16 e OFDMA system," *Journal of Network and Computer Applications*, 2013.
- [6] M.-H. Tsai, J.-T. Sung, and Y.-M. Huang, "Resource management to increase connection capacity of real-time streaming in mobile WiMAX," *IET communications*, vol. 4, pp. 1108-1115, 2010.
- [7] M. A. Teixeira and P. R. Guardieiro, "A new and efficient adaptive scheduling packets for the uplink traffic in WiMAX networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2011, pp. 1-11, 2011.
- [8] K. C. Chilukuri and P. Reddy, "Admission Control and Flow Scheduling for IEEE 802.16 WiMAX Networks with QoS Requirements," *International Journal of Computer Applications*, vol. 42, 2012.
- [9] P. Gupta, B. Kumar, and B. Raina, "Multiple Downlink Fair Packet Scheduling Scheme in Wi-Max," *International Journal of Computer Applications Technology and Research*, vol. 2, p. 389, 2012.
- [10] A. Zubow, D. C. Mur, X. P. Costa, and P. Favaro, "Greedy scheduling algorithm (GSA)—Design and evaluation of an efficient and flexible WiMAX OFDMA scheduling solution," *Computer Networks*, vol. 54, pp. 1584-1606, 2010.
- [11] R. Gunasekaran, S. Siddharth, P. Krishnaraj, M. Kalaiarasan, and V. Rhymend Uthariaraj, "Efficient algorithms to solve broadcast scheduling problem in WiMAX mesh networks," *Computer Communications*, vol. 33, pp. 1325-1333, 2010.
- [12] T. Ali-Yahiya, A.-L. Beylot, and G. Pujolle, "Downlink resource allocation strategies for OFDMA based mobile WiMAX," *Telecommunication Systems*, vol. 44, pp. 29-37, 2010.
- [13] J. El-Najjar, C. Assi, and B. Jaumard, "Joint routing and scheduling in WiMAX-based mesh networks," *Wireless Communications, IEEE Transactions on*, vol. 9, pp. 2371-2381, 2010.
- [14] C. So-In, R. Jain, and A.-K. Al-Tamimi, "A scheduler for unsolicited grant service (UGS) in IEEE 802.16 e mobile WiMAX networks," *Systems Journal, IEEE*, vol. 4, pp. 487-494, 2010.
- [15] D. Shu'aibu and S. S. Yusof, "Link aware call admission and packet scheduling for best effort and UGS traffics in mobile WiMAX," *Int. J. Phys. Sci.*, vol. 6, pp. 1694-1701, 2011.
- [16] E. Laia and I. Awan, "An interactive QoS framework for fixed WiMAX networks," *Simulation Modelling Practice and Theory*, vol. 18, pp. 291-303, 2010.
- [17] T. Ali-Yahiya and H. Chaouchi, "Fractional frequency reuse for hierarchical resource allocation in mobile WiMAX networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2010, p. 7, 2010.
- [18] M. S. Kuran, G. Gur, T. Tugcu, and F. Alagoz, "Applications of the cross-layer paradigm for improving the performance of WiMax [Accepted from Open Call]," *Wireless Communications, IEEE*, vol. 17, pp. 86-95, 2010.
- [19] H.-C. Lu, W. Liao, and F.-S. Lin, "Relay station placement strategy in IEEE 802.16 j WiMAX networks," *Communications, IEEE Transactions on*, vol. 59, pp. 151-158, 2011.
- [20] C. Prommak and C. Wechtaison, "WiMAX network design for cost minimization and access data rate guarantee using multi-hop relay stations," *International Journal of Communications*, vol. 4, pp. 39-46, 2010.
- [21] Alla, S.B., Ezzati, A., A QoS-guaranteed coverage and connectivity preservation routing protocol for heterogeneous wireless sensor networks, (2012) *International Journal on Communications Antenna and Propagation (IRECAP)*, 2 (6), pp. 363-371.

- [22] David Neels Pon Kumar, D., Murugesan, K., Arun Kumar, K., Raj, J., Performance analysis of fuzzy neural based QoS scheduler for mobile WiMAX, (2012) *International Journal on Communications Antenna and Propagation (IRECAP)*, 2 (6), pp. 377-385.
- [23] Mohades, Z., Vakili, V.T., Razavizadeh, S.M., Enhanced dynamic fractional frequency reuse in WiMAX systems using AMC and power control, (2011) *International Journal on Communications Antenna and Propagation (IRECAP)*, 1 (4), pp. 338-341.
- [24] Ismael, F.E., Yusof, S.K.S., Faisal, N., Bandwidth grant algorithm for delay reduction in IEEE 802.16j MMR WiMAX networks, (2013) *International Journal on Communications Antenna and Propagation (IRECAP)*, 3 (2), pp. 140-145.
- [25] Zmezm, H., Hashim, S., Sali, A., Alezabi, K., Seamless and Secure Design for Subsequent Handover in Mobile WiMAX Networks, (2014) *International Review on Computers and Software (IRECOS)*, 9(8), pp. 1399-1407.
doi: <http://dx.doi.org/10.15866/irecos.v9i8.2942>

Authors' information

¹Assistant Professor, Bharathiyar Arts & Science College for Women, Deviyakurichi, Attur (Tk), Salem (Dt)-636112.

²Associate Professor, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu, India.



Mrs. **P. Kavitha** received her Master Degree in Computer Applications from Bharathidasan University, Trichy, Tamil Nadu, India in the year 1999 and received M.Phil degree in Computer Science from Periyar University, Salem, Tamil Nadu, India in the year 2006. Currently she is working as Guest Lecturer, Department of Computer Science, Arignar

Anna Govt.Arts College, Attur -Salem, TamilNadu . She has 12 years of experience in teaching in Bharathiyar Arts & Science college (w),Attur,TamilNadu . She is doing Ph.D in Computer science at Periyar University, Salem, under the supervision of Dr. R.Umarani, Associate Professor, Department of Computer Science , Sri Saradha SCollege For Women (Autonomous),Salem,TamilNadu. Her research interests include Networking, Mobile Computing.



Dr. **R. Uma Rani** received the Ph.D. degree in Computer Science from Periyar University, Salem, Tamilnadu, India, in 2005. She has 20 years of experience in teaching. She is currently working as the Associate Professor, Dept. of Computer Science in Sri Sarada College for Women, Salem, Tamilnadu, India. She has published many papers in International Journals.

She is an editorial board member in a few international journals. She has also acted as chair person in National and International Conferences. She has acted as resource Person in many National Conferences. Her area of interest includes Data Mining, Grid Computing and Data Warehousing.

Adaptive Algorithm for Beacon and Superframe Values in IEEE802.15.4 Based Networks

Wail Mardini¹, Abdulaziz Alraddadi²

Abstract – The IEEE 802.15.4 standard was proposed by IEEE TG4 and has been commercially adopted to specify protocols for the physical and MAC layers. Despite the diverse nature and goals intended from WSN applications, they all must work in a way that enables them to benefit from features provided by the standard. This is directly related to the chosen Beacon and Superframe orders (BO, SO) combination values. The studies of Marwa Salaymeh et al. (2013) investigated the standard behavior of the protocols as it is applied on CBR applications implemented on a seven star topology scenarios and reveals the optimal range of combinations for Bo and So. Moreover, an adaptive algorithm that converges to the network current performance has been proposed in order to adaptively change Bo and So to improve the overall performance. For different topologies and different application behaviors the optimal combination values may significantly vary. Thus for VBR-based applications, different topologies will be investigated and analyzed in order to reveal the optimal (Bo,So) combinations. A new modified approach based on our analysis will be presented. The new approach will operate in two modes; Evaluating Options Mode (EOM) and non-EOM mode in which the approach will adapt and move to different values in order to gain better performance, if possible. The analysis phase and the implementation of the new approach have been tested in QualNet version 5.2 simulation environment. The new approach could gain performance improvements of up to 10% of the previous performance based on the defined performance metric. **Copyright © 2014 Praise Worthy Prize S.r.l. - All rights reserved.**

Keywords: Wireless Sensor Networks, IEEE802.15.4, Beacon, Superframe, Energy Consumption

Nomenclature

WSN	Wireless Sensor Network
LR-WPAN	Low Rate Wireless Personal Area Network
MAC	Media Access Control
FFD	Full Functional Devices
RFD	Reduced Functional Devices
BI	Beacon Interval
SD	Superframe Duration
aBaseslotDuration	Slot duration
aBaseSuperframeSuratio	Frame duration
CSMA/CA	Carrier Sense Multiple Access/Collision Avoidance
GTS	Guaranteed Time Slots
CFP	Contention Free period
CBR	Constant Bit Rate
BO	Beacon Order
SO	Superframe Order
Bo _{pan}	BO for personal area network coordinator
So _{pan}	SO for personal area network coordinator
Bo _{Coor}	BO for area coordinator
So _{Coor}	SO for area coordinator
Bo _{Reg}	BO for regular nodes
So _{Reg}	SO for regular nodes
EOM	Evaluating Options Mode

I. Introduction

Wireless Sensor Network (WSN) applications are diverse, and each one is characterized with special needs that may not be required by others. However, regardless of WSN application type, all WSN applications foresee the same goal, that is, energy conservation. Many algorithms have been proposed for this purpose ([1]-[19]). The Institute of Electrical and Electronic Engineers Task Group 4 (IEEE TG4) put the standard for IEEE 802.15.4. This standard is proposed to support Low Rate (LR) applications which is referred to as LR-WPAN (Wireless Personal Area Network). This is usually designed for use in applications in homes, buildings and some industrial automated applications [10].

In [10], the standard has been approved for the physical and the Media Access Control (MAC) layers. Two types of devices were assumed; Full Functional Devices (FFDs) and Reduced Functional Devices (RFDs). RFD is used for regular sensor nodes which sense and send data to its parent node. The other nodes such as pan coordinator, area coordinator, sink or intermediate nodes. Both FFDs and RFDs communicate with each other forming two types of topologies: star and peer to peer topologies. Peer to peer topology can be classified to either a mesh or a cluster tree topology [6][13]-[15]. There are two modes in which IEEE802.15.4 MAC operates; beacon or beaconless

modes. The beacons are a kind of advertising frames broadcasted by FFD regularly.

These frames are used to synchronize with nodes in its area of transmission [16]. The Beacon Interval (BI) is the time between two successive beacons. This time is divided into 16 equal sized slots and can be calculated using (1) [15]. Nodes can use the channel during the whole BI period or can sleep for some time outside the Superframe Duration (SD) active session. The SD is calculated using (2) [15]:

$$BI = aBaseSuperframeDuration * 2^{BO} \quad (1)$$

$$SD = aBaseSuperframeDuration * 2^{SO} \quad (2)$$

where $0 \leq SO \leq BO \leq 14$.

aBaseSuperframeDuration is calculated using (3):

$$\begin{aligned} aBaseSuperframeDuration &= \\ &= aBaseSlotDuration * totalNumberOfSlots \end{aligned} \quad (3)$$

In 2.4 Ghz RF band there are 62500 symbol/s where each symbol equals 16 μ s. Since each slot duration (aBaseSlotDuration) equals about 60 symbols, then the 16 slots will contribute in 960 aBaseSuperframeDuration symbols which can sum to about 15.36 ms.

Therefor each node can infer the duration of its sleep period. All those concepts can be indicated through one concept which is the duty cycle (D) which can be defined as the percentage of time the node is awake from the whole time between the two successive beacons. D is mathematically expressed in (4) [15][16]:

$$D = \frac{SD}{BI} \times 100\% \quad (4)$$

The node should locate the beginning of the next time slot in order to be able to access the medium. To do so it uses the known contention based algorithm; slotted Carrier Sense Multiple Access/Collision Avoidance algorithm (CSMA/CA). Therefor this time portion is referred to as Contention Access Period (CAP) [15][16].

Furthermore, the standard empowers PANc with the authority to assign some slots excessively for some nodes during which they can utilize the channel alone. This is why such time slots are referred to as Guaranteed Time Slots (GTS).

The optional period which includes those slots is referred to as the Contention Free period (CFP) and it include maximum of seven GTS which are preserved optionally after the CAP period. CAP and the optional CFP together are referred to as the Active Period. Active period is the time during which nodes can be active and are able to use the medium. The duration of this period is often referred to as the SuperFrame Duration [15] [16]. More precisely, every time the node needs to access the channel, it needs to locate the boundary of what is called the slotted/un-slotted CSMA backoff period. The Backoff period unit is indicated through the aUnitBackoffPeriod which equals to 20 symbols or 0.32 ms [16]. The lengths

of the discussed periods are assigned through the beacon frame which is transmitted in the first time slot (slot 0) [10].

Obviously, improving the beacon enabled standard performance is directly related to the chosen BO and SO values. How to decide the optimal BO and SO values that achieve the best performance is an application related issue. For example, an application may have packets ready for transmission every second but needs to be active for 30 minutes and sleep for 30 minutes.

Some applications spend most of the time inactive, thus, i.e., low duty cycles; others need to work through full duty cycles, while many of applications need to sleep for some time portions. Hence, each application has its own special case that has much to do in the decision of BO and SO values, keeping in mind that the basic building block of any network topology consists of seven nodes (piconet). Therefore, we need to find a mechanism that is general enough for beacon enabled MAC coordinator to regularly examine PAN status and performance to adapt the superframe parameters and durations.

IEEE 802.15.4/ZigBee standard is proposed to achieve specific relaxed goals that other standards cannot provide. To achieve goals intended, IEEE 802.15.4 MAC layer is characterized with features that allow it support extra low cost, low power, low data rate and low complexity for wireless short range networks. However, those features worth nothing if they are not harnessed in a way that achieves the optimal standard performance especially when it is related to energy consumption.

Improving the beacon enabled standard performance is directly related to the chosen MAC superframe structure parameters, mainly, BO and SO. Therefore, we need to find a mechanism that is general enough through which a beacon enabled MAC coordinator regularly examines PAN status and performance to adaptively change superframe parameters and durations for the aim of achieving near optimal network performance irrespective of application type, network size and load. This is considered as a preliminary step to apply the proposed algorithm on all the standard supported topologies.

Other than CBR applications which are fully investigated in [1] [2] study, the same work shall be conducted for VBR applications with different mean values. Investigation shall take place on star and mesh topologies. Such applications are needed to operate through the nine possible duty cycles in order to decide the optimal (BO, SO) combination that achieves optimal or near optimal performance. Moreover, the cut-off combination after which the standard performance drops dramatically is needed to be revealed besides any other combinations that achieve bad performance and which are needed to be excluded from our consideration. After deciding the cutoff combination, the new adaptive algorithm performance can be investigated on such VBR applications. The new adaptive algorithm proposed in [1] estimates average end to end delay at each nodes

randomly, thus, it is needed to estimate average end to end delay on each node in the PAN and the total overall result from all nodes is needed to be averaged and according to which PAN evaluates the current PAN performance.

This rest of this paper is organized as follows. Section Two presents a summary for some of literature work which is closely related to the paper topic. In section three we present the methodology that will be used to analyze the problem with some performance results. Section four discusses the suggested approach and the results of this approach are presented in section five.

Finally a concluding remarks and some future work will be presented in section six.

II. Literature Review

Recently, almost all wired sensors are replaced with wireless ones forming a wireless network of sensors from which the Wireless Sensor Networks (WSNs) era has been emerged. WSN applications are divers, and each one is characterized with special needs that may not be required by others. However, regardless of WSN application type, all WSN applications foresee the same goal, that is, energy conservation. This is because thousands of sensor nodes are often scattered in areas range from tenth of meters to thousands of kilometers where they cannot be easily recharged, so it is reasonable to force sensor nodes track and monitor phenomena for months or even years. Finding ways for energy conservation has been the basic interest for all WSN recent and literature studies which all focus on designing WSN energy efficient algorithms and standards one of which is the IEEE802.15.4 standard.

The IEEE 802.15.4 standard was proposed by the Institute of Electrical and Electronics Engineers Task Group 4 (IEEE TG4) and has been commercially adopted to specify protocols for the physical and MAC layers. The standard is basically proposed to support Wireless Personal Area Network (WPAN) communication and is considered promising for low cost, low power and low data rate (LR) networks such as WSNs, thus, it is often referred to as LR-WPAN IEEE 802.15.4 standard and it best suits home, building and industrial automated application.

IEEE 802.15.4 MAC supports Full Function Devices (FFDs) and Reduced Function Devices (RFDs). FFDs provide the full services of MAC IEEE 802.15.4 while RFDs perform reduced IEEE 802.15.4 MAC services. FFDs can be classified to a PAN coordinator (only one PAN coordinator), local coordinators like routers and end devices that perform the regular application operations, while RFDs can act only as an ordinary end device. Nodes with different types that follow the standard can communicate with each other's forming two types of topologies which are star and peer to peer topologies. Peer to peer topology can be classified as either a mesh or cluster tree topologies [6]-[8].

IEEE802.15.4 standard can work through three

different Radio Frequency (RF) bands which work through different data rates and distributed over world continents. The most popular RF band is the 2.4 GHz.

As it is stated previously, one of the WSN MAC critical issues is the power dissipation which may result due to overhearing, idle listening, channel access collisions and packet massaging overhead. This will increase power consumption and hence will reduce network lifetime. IEEE802.15.4 MAC overcomes these factors through operating either asynchronous beaconless mode or through the beacon enabled mode which schedules nodes transmission through sending regular beacon frames between transmitter and receiver. This will reduce power consumption because nodes need not be awake all the time and can sleep between coordinated transmissions. Those beacon frames can be sent only by FFDs.

The superframe is the time between two successive beacons which is divided into 16 active period time slots. There is an optional inactive period which can be used when the sleep mode is used for all nodes. The active period consists of a Contention Access Period (CAP) and an optional Contention Free Period (CFP). The time between those two frames is referred as the Beacon Interval (BI). BI duration can be specified through Beacon Order parameter (BO), nodes can use the channel during the whole BI period or can sleep for some time portions, and the parameter which decides that is the superframe order (SO) where $0 \leq SO \leq BO \leq 14$. All those concepts can be indicated through one concept which is the duty cycle (D) [1]. It is the percentage of time the node is awake from the whole time between the two successive beacons, thus, the standard can support up to nine duty cycles. In order to avoid all nodes transmit at the same time, a commonly followed approach in CAP is the slotted Carrier Sense Multiple Access/ Collision Avoidance (CSMA/CA) through which nodes can compete with each others when accessing a channel. Every time the node needs to access the channel, it has to locate the boundary of what is called the slotted/un-slotted CSMA backoff period. Thus, the standard deals with time in term of backoff period unit which in beacon enabled mode is aligned between the slot boundaries and indicated through aUnitBackoffPeriod which equals to 20 symbols or 0.32 ms. the length of each period discussed previously is assigned through beacon frame which is transmitted in the first time slot (slot 0). Despite the diverse nature and goals intended from WSN applications, they all must work in a way that enables them benefit from features provided by the standard, however, this is directly related to the chosen (BO, SO) combination values. How to decide the optimal BO and SO values that achieve the best performance is an application related issue. WSN behaviour can be characterized according to the duration of time they need to be active; which naturally indicates their sleep period and which refers to the concept of duty cycle. Moreover, wireless applications can be classified according to the rate at which the surrounding

phenomena can be captured which is determined through packets arrival rate. For example, an application may work at a Constant Bit Rate (CBR), and may has packets ready for transmission every second but need to be active for 30 minutes and hence shall sleep for the other 30 minutes in an hour. Some applications spend most of the time inactive, thus, work through low duty cycles; others need to work in full duty cycles, while many of them need to sleep in some time portions.

Furthermore, after the enhancements in IEEE802.15.4 standard, more divers WSN applications continue to arise which work through Variable Bit Rates (VBR). Hence, each application has its own special case that has much to do in the decision of BO and SO, keeping in mind that the basic building block of any network topology consists of seven nodes (piconet).

In [3], the IEEE 802.15.4 standard performance is evaluated in term of throughput and packet delivery ratio.

The study interested in the quality of service QoS for real time sensor applications and provides an enhancement to the current IEEE 802.15.4 beacon enabled standard by dynamically allocating the already existed GTS. The standard performance metrics were evaluated through varying both BO and SO while preserving dynamically allocated one GTS. However, the study considered only 100% and 50% duty cycles and the maximum SO and BO values tested were 6 due to association latency that may result from choosing higher values which is not sufficient for WSN applications. Other QoS property examined was the collision probability which was evaluated through varying number of nodes. Simulation run through NS2 simulator and applied on a star topology. Results showed that higher values of BO increases throughput due to the decreased possibility of packets drops. Moreover, results revealed that collision probability increases as the number of nodes increases which will degrade the successful use of the channel and hence achieving poor throughput. In [4], the performance of beacon enabled mode IEEE 802.15.4 is evaluated in term of energy consumption in large scale peer to peer based WSN.

A clustered tree network in which data aggregation and load distribution among cluster heads was mathematically analyzed. Analysis of the IEEE 802.15.4 MAC were performed on a real Zigbee nodes applied on home network areas by varying BO values between 6 and 10 while fixing SO value to 0. High fraction of packets transmitted is sacrificed for the aim of minimizing the power consumed by allowing nodes active for only 15.36 ms and turn the transceiver off else after. Results revealed that power consumption keeps on decreasing by increasing BO to some value (approximately 10) after which it is started to increase again. In [4], the coordinator function was synchronized with cluster heads through receiving regular beacons in order to decrease energy consumption even though for high values of Bos.

However, the study considered only very low duty cycles due to the small fraction of CAP and did not consider the effect of SO on the standard performance at

all. In [5], performance of the slotted CSMA/CA is investigated by studying the effects of some IEEE 802.15.4 standard configuration parameters on the network behaviour. Those parameters were SO, BO and Backoff Exponent (BE). However, the same study took place in [3] but with other criteria's considered such as number of nodes and data frame size. Simulation experiments were done for 13 different values of BO and SO all with a duty cycle of 100%. Those experiments intended to reach up the best range of traffic load offered that achieves the optimal performance metrics values. Metrics which were evaluated are the throughput, average delay and network reliability.

In [5], a utility concept is defined which combines two or more metrics. The best range of offered load that achieved the optimal trade-off between throughput and average delay utility was found to be between 35% and 60%. This study did not concern parameters behaviour with sleep period enabled.

However, [6] proposes an optimization problem in order to achieve the minimum energy consumed under the constraints of packet delivery reliability, that is, number of packets received successfully and the maximum delay a packet takes in its journey from the moment it is generated by the application layer until the moment it reaches the receiver. The objective function was achieved after finding optimal values for the two decision variables which are BO and SO.

In [6], traffic, total energy and power consumption minimization probabilistic models were proposed. Experiments run through a C implemented simulator and applied on a star topology. Simulation results revealed that for a network where packets generated under Poisson processes and where the number of nodes varying from 5 to 35, the optimal value for BO was found to be 7 while that for SO was found to be 1 in case that number of nodes is less than 15 and 2 otherwise. However, choosing optimal values for BO and SO depends much on the quality of service constraints chosen.

The studies in [1][2] investigates the standard behaviour as it is applied on a 1s CBR application implemented on a seven star topology scenarios. It reveals the optimal range of combinations through which such an application can work through while achieving its optimality in terms of energy consumption, throughput and average end to end delay. However, the same study should be followed on other types of WSN applications such as VBR applications in order to find optimal (BO, SO) combinations. Moreover, this work proposes an adaptive algorithm that converges to the network current performance and improves network performance accordingly irrespective of the duty cycle. This algorithm foresees generality through deciding how to adaptively change superframe parameters and durations irrespective of application type, network size and load. Therefore, this algorithm needs more investigation to be applied on other WSN applications and standard supported topologies in order to achieve general optimality and accuracy.

III. Methodology and Performance Analysis

This work investigates the IEEE 802.15.4 standards performance behaviour as it is applied on a VBR WSN application. Investigation shall take place on a basic building blocks star and mesh network topologies. Different VBR means shall be considered comprising very low, low, normal, and high traffic loads.

Investigation will consider the following nine possible MAC duty cycles:

D {100%, 50%, 25%, 12.5%, 6.24%, 3.13%, 1.56%, 0.78%, 0.39%}

For each VBR mean, the standard performance shall be evaluated according to energy consumption, average end to end delay and throughput. Consequently, the optimal (BO, SO) combination or range of combinations shall be decided. Moreover, the cut-off combination after which the standard performance drops dramatically shall be revealed along with any other combinations that achieve bad performance. Such combinations must then be excluded from our consideration when we parameterize the standard in case it is allowed to work following VBR applications. Finding the cut-off (BO, SO) combination, enables the study of the new adaptive algorithm similar to the one proposed in [1]. Thus, both the new adaptive algorithm and the standard original MAC algorithm performance shall be reevaluated in terms of the three metrics mentioned above. Investigation shall take place on the same stated basic building blocks. This needs a cross layer investigation between both the MAC and VBR application layers.

The idea is to decide PAN performance according to all performance metrics combined and according to which it decides whether to manipulate (BO, SO) or not.

For example, if current energy behaviour is found to be worse than the previously estimated one, then BO value may increase. On the other hand, if current throughput is found to be worse than the previous one then we should adapt to new SO value. This work shall consider the physical layer for energy consumption estimation, and application layer for throughput estimation, thus, shall comprise three layers which are, physical, MAC and application layers. All such investigation works stated will be achieved using the QualNet 5.2 simulator.

The following specific assumptions were made:

- Two topologies were considered
 - o Topology 1: Mesh and star nodes
 - o Topology 2: Multi Stars (three stars; with 3 nodes+1 coordinator each)
- Three types of devices were used
 - o Regular Devices: which are the sensors who generate packets toward sink node.
 - o Coordinators Devices: which are the devices that can route packets.
 - o Pan-coordinator: which is the main device controls the network and it will be considered the sink node.

- Five traffic averages were considered:
 - o 0.5, 1, 2, 4 seconds interarrival packet averages with exponential distribution
- performance metrics considered are
 - o End-to-end delay
 - o Throughput
 - o Power consumption
- The following (Bo,So)=(x,y) combinations where considered:

Scenario 1={ (Bo_{pan}, So_{pan}) = (Bo_{cor}, So_{cor}) = (Bo_{reg}, So_{reg}) = (x, x) :x=1,2,..14 }

Scenario 2={ (Bo_{pan}, So_{pan}) = (Bo_{cor}, So_{cor}) = (Bo_{reg}, 1+So_{reg}) = (x, x) :x=1,2,..13 }

Scenario 3={ (Bo_{pan}, So_{pan}) = (Bo_{cor}, So_{cor}) = (1+Bo_{reg}, 1+So_{reg}) = (x, x) :x=1,2,..13 }

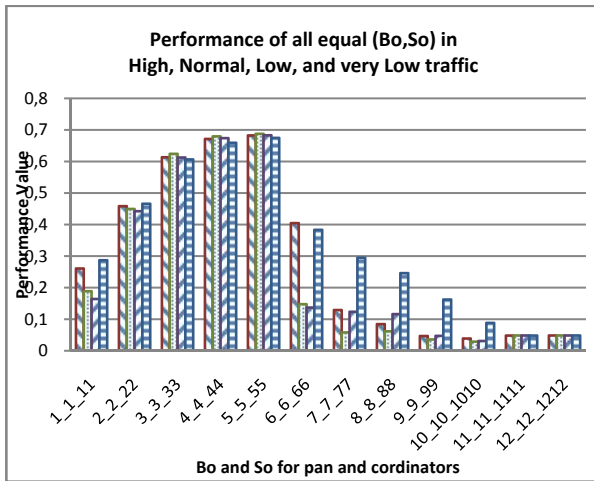
Scenario 4={ (Bo_{pan}, 1+So_{pan}) = (Bo_{cor}, 1+So_{cor}) = (1+Bo_{reg}, 1+So_{reg}) = (x, x) :x=1,2,..13 }

Many other combinations have been tested and they all give very low performance which is not comparable with these cases presented above. The following results are scaled performance results based on delay, throughput, and energy consumed in four different traffic scenarios; High, Normal, Low and Very Low packet interarrival times for all scenarios above. There four interarrival times correspond to the four columns at each point, respectively. In all cases presented in Figs. 1 to 4, the best performance is achieved around values from 2 to 5 for either Bo or So. Based on which the algorithm in the next section will be developed.

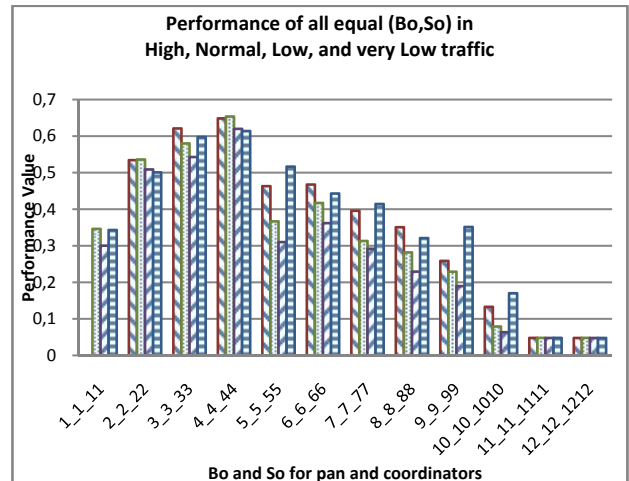
IV. Proposed Approach

Based on the analysis made on the previous section (see Figs. 1, 2, 3, and 4) we can decide on set of values around which the performance achieves optimal or near optimal values. The main idea of our approach is to periodically monitor and collect the three performance measures at the application layer and calculate a Normalized Performance Value (NPV) based on these measured values. The NPV value is passed to the MAC layer which will adjust the So and Bo according to the following two modes:

- If Evaluating Options Mode (EOM) is on, then three options will be tried one after another for some short-enough time and based on that the decision will be made:
 - The first option is to set So_{pan}, Bo_{cor}, So_{cor}, Bo_{reg} and So_{reg} to the same value of Bo_{pan}.
 - The second option is to set the So_{pan}, Bo_{cor} and So_{cor} to the same value of Bo_{pan} while Bo_{reg} and So_{reg} to one value lower.
 - The third option is to set the Bo_{cor} to the same value of Bo_{pan} while So_{cor}, Bo_{reg} and So_{reg} to one value lower.

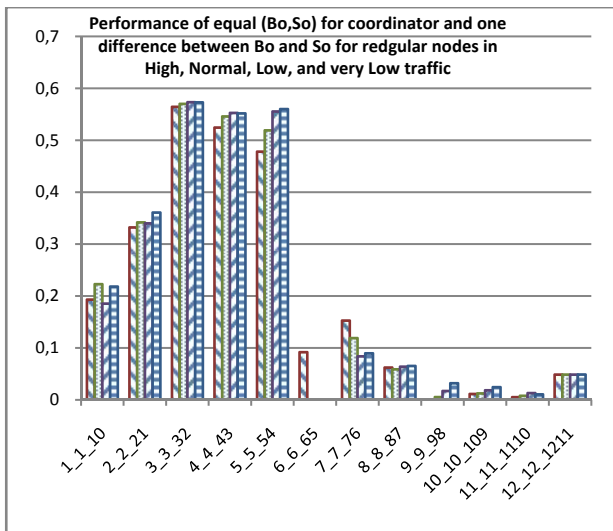


(a)

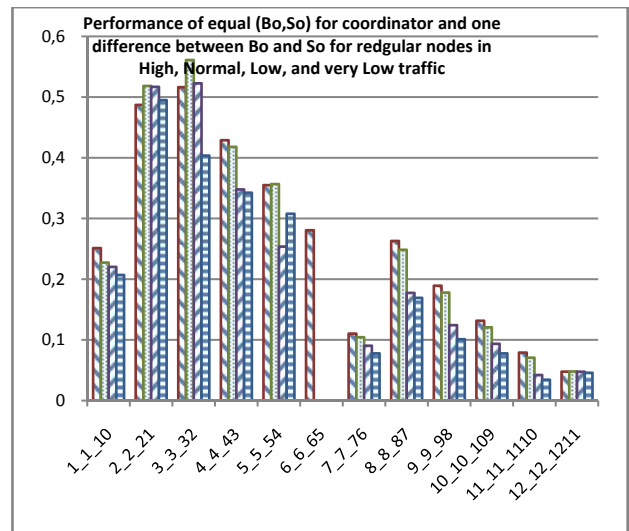


(b)

Figs. 1. Performance of Topology 1 (a) and Topology 2 (b) under scenario 1 with four types of traffic loads

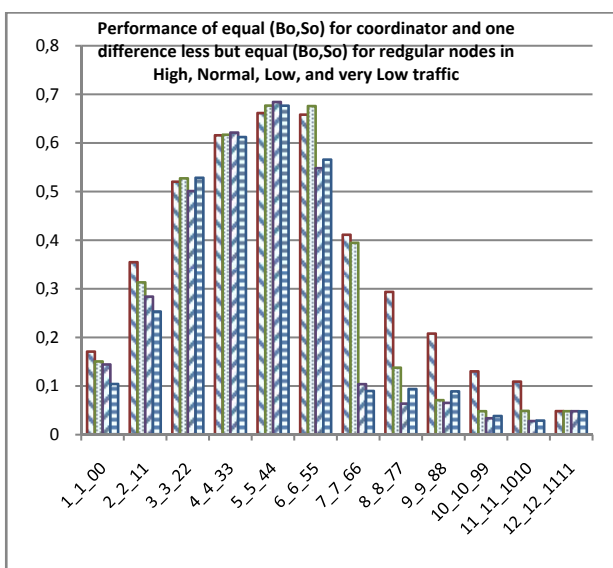


(a)

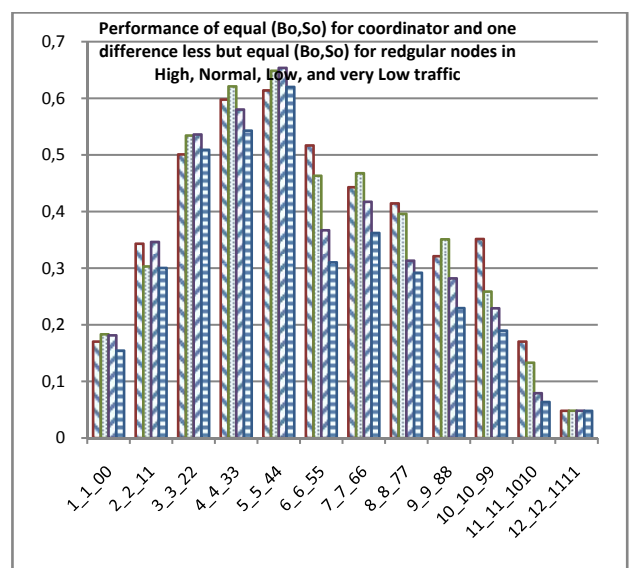


(b)

Figs. 2. Performance of Topology 1 (a) and Topology 2 (b) under scenario 2 with four types of traffic loads

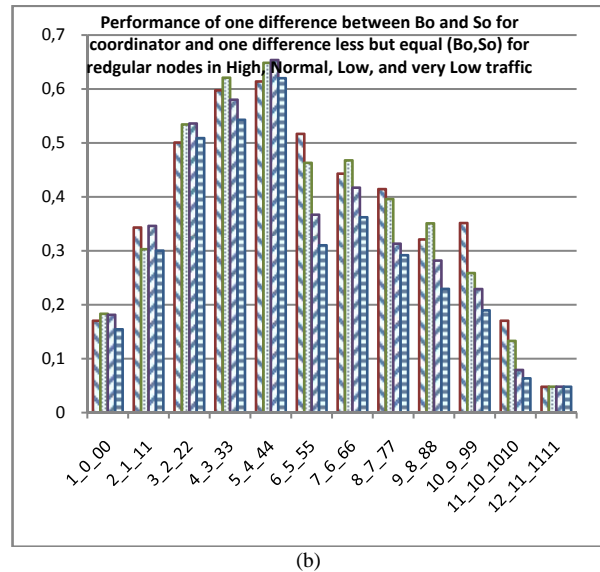
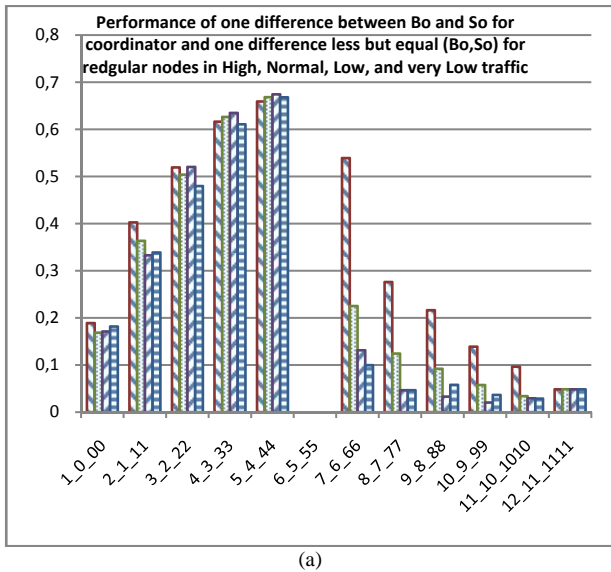


(a)



(b)

Figs. 3. Performance of Topology 1 (a) and Topology 2 (b) under scenario 3 with four types of traffic loads



Figs. 4. Performance of Topology 1 (a) and Topology 2 (b) under scenario 4 with four types of traffic loads

If any of the three options gave better NPV, then these values will be used. Otherwise, the previous values will be kept and the mode will be changed to the non-EOM mode:

- If non-EOM is on, then if the difference between the new value and the previous value is:
 - Non-negative difference (i.e. the new value is the same or better than the old one), then the values will be kept for some time and then it switch to the EOM mode.
 - Negative difference then all the value will be set to equal values to (3,3), (4, 4) or (5,5) one after another when placed on this case.

V. Results and Analysis

The approach described in the previous section is implemented using the same environment discussed in section 4. The following tables present the obtained results for the two topologies under test. Table I shows the results obtained for topology 1 and Table II shows the results for Topology 2.

We can see that the improvement in low traffic cases is much better than the cases of very low and low traffic.

This can be justified as follows. The algorithm scans for the current performance and then adjust the values of Bo and So based on the result. In the high traffic case, there is no enough time to do so because the traffic changes faster than the other cases.

Moreover, the algorithm worked better and gave better improvements in the case of topology 1 rather than topology 2 since more learning time is needed for larger networks as the case in topology 2.

VI. Conclusions and Future Work

A new approach is proposed to improve the performance in IEEE 802.15.4 standard.

TABLE I
IMPROVEMENTS IN THE RESULTS FOR TOPOLOGY 1 USING THE BEST ORIGINAL VALUES AND THE NEW PROPOSED APPROACH

Traffic load	Performance value (Original)	Performance value (New)
Very Low	0.67	0.78
Low	0.69	0.75
Normal	0.69	0.71
High	0.68	0.69

TABLE II
IMPROVEMENTS IN THE RESULTS FOR TOPOLOGY 2 USING THE BEST ORIGINAL VALUES AND THE NEW PROPOSED APPROACH

Traffic load	Performance value (Original)	Performance value (New)
Very Low	0.62	0.68
Low	0.65	0.70
Normal	0.65	0.66
High	0.61	0.62

A various VBR-based applications and different topologies were investigated and analyzed in order to reveal the optimal (Bo,So) combinations.

The new approach will operate in two modes; Evaluating Options Mode (EOM) and non-EOM mode in which the approach will adapt and move to different values in order to gain better performance, if possible. The new approach could gain performance improvements of up to 10% of the pervious performance based on the used performance metric which is a combination of the three typical performance metrics; end-to-end delay, throughput, and energy consumption.

A future work would be to study the effect of the frequency at which the performance is calculated.

This should be related to the traffic rate as more frequent packets would affect the performance fast and new calculations should be carried on.

Acknowledgements

This work was supported by Jordan University of Science and Technology as part of a sabbatical leave

work for the first author, Wail Mardini.

References

- [1] Marwa Salaymeh, "Power Efficiency Model for ZigBee Networks", Jordan University of Science and Technology Master Thesis supervised by Wail Mardini and Yaser Khamayseh, 2013.
- [2] Marwa Salaymeh, Wail Mardini, Yaser Khamayseh, Muneer Bani Yassein, "Optimal Beacon and Superframe Orders in WSNs", The Fifth International Conference on Future Computational Technologies and Applications, FUTURE COMPUTING 2013, May 27 - June 1, 2013 -Valencia, Spain.
- [3] Charfi F, Bouyahi M. "Performance evaluation of beacon enabled IEEE 802.15. 4 under NS2", arXiv preprint arXiv 2012:1204.1495.
- [4] Khan SA, Khan FA, "Performance analysis of a zigbee beacon enabled cluster tree network", In Third International Conference on Electrical Engineering (ICEE'09) 2009 April: 1-6.
- [5] Koubaa A, Alves M, Tovar E. "A comprehensive simulation study of slotted CSMA/CA for IEEE 802.15. 4 wireless sensor networks", IEEE WFCS 2006: 63-70.
- [6] Shu F, Sakurai T, Vu HL, Zukerman M. Optimizing the IEEE 802.15. 4 MAC. In IEEE Region 10 Conference (TENCON) 2006 November: 1-4.
- [7] IF. Akyildiz, W. Su, Y. Sankarasubramaniam, and A. Cayirci; "A survey on sensor networks", Communications Magazine 2002. Atlanta, GA, USA, vol. 40(8), pp. 102-114, 2002.
- [8] L. Selavo, A. Wood, Q. Cao, T. Sookoor, H. Liu, A. Srinivasan, and J. Porter, "Wireless sensor network for environmental research", Proc. the 5th international conference on Embedded networked sensor systems. Sydney, Australia Nov. 2007, pp. 103-116.
- [9] A. Koubaa, "Promoting Quality of Service in Wireless Sensor Networks", (Submitted for receiving Habilitation Qualification in Computer Science) National School of Engineering, Sfax, Tunisia, 2011.
- [10] SC. Ergen, "ZigBee/IEEE 802.15. 4 (Summary)", [Online][accessed April 2013], Available from URL <http://pages.cs.wisc.edu/~suman/courses/838/papers/zigbee.pdf>.
- [11] P. Park, C. Fischione, and KH. Johansson, "Adaptive IEEE 802.15. 4 protocol for energy efficient, reliable and timely communications", Proc. the 9th ACM/IEEE international conference on information processing in sensor networks. Stockholm, April 2010, pp. 327-338.
- [12] J. Hoffert, K. Klues, and O. Orjih "Configuring the IEEE 802.15. 4 MAC Layer for Single-sink Wireless Sensor", Washington University in St. Louis, 2005.
- [13] P. Patro, M. Raina, V. Ganapathy, M. Shamaiah, and C. Thejaswi, "Analysis and improvement of contention access protocol in IEEE 802.15. 4 star network", Proc. Mobile Adhoc and Sensor Systems (MASS 07), IEEE International Conference. Piza, Italy, Oct. 2007, pp. 1-8.
- [14] H. Deng, J. Shen, B. Zhang, J. Zheng, J. Ma, and H. Liu, "Performance Analysis for Optimal Hybrid Medium Access Control in Wireless Sensor Networks". Proc. Global Telecommunications Conference (GLOBECOM 08). LA, USA, Nov. 2008, pp 1-5.
- [15] E. Casilari and J.M. Cano-García, "Impact of the Parameterization of IEEE 802.15. 4 Medium Access Layer on the Consumption of ZigBee Sensor Motes", Proc. The Fourth International Conference on Mobile Ubiquitous Computing Systems, Services and Technologies (UBICOMM 2010). Florence, Italy, Oct. 2010, pp. 117-123.
- [16] X. Li, CJ. Bleakley, and W. Bober, "Enhanced Beacon-Enabled Mode for improved IEEE 802.15. 4 low data rate performance", Wireless Networks 2012, vol. 18, pp. 59-74.
- [17] F. Charfi, and M. Bouyahi, "Performance evaluation of beacon enabled IEEE 802.15.4 under NS2", arXiv preprint arXiv 2012, pp. 1204.1495.
- [18] SA. Khan and FA. Khan, "Performance analysis of a zigbee beacon enabled cluster tree network", Proc. Third International Conference on Electrical Engineering (ICEE'09). Lahore April 2009, pp. 1-6.
- [19] M. Neugebauer, J. Plonnigs, and K. Kabitzsch, "A new beacon order adaptation algorithm for IEEE 802.15. 4 networks", Proc. The Second European Workshop on Wireless Sensor Networks. Ghent, Belgium 2005, pp. 302-311.

Authors' information

¹Taibah University/ Kingdom of Saudi Arabia and Jordan University of Science and Technology/Jordan.

²Taibah University/ Kingdom of Saudi Arabia.



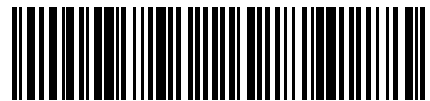
Wail Mardini is an associate professor of Computer Science at the Taibah University/Yanbu branch since September 2013. Dr. Mardini received his Ph.D. degree in Computer Science from University of Ottawa/Canada at 2006 and his Masters degree from University of New Brunswick/Canada at 2001. Dr. Mardini worked before at the Computer Science department at Jordan University of Science and Technology (JUST) from 2006 until 2013. Dr. Mardini was the CS department chair during the academic years 2010/2011 and 2011/2012, and the faculty vice dean during the academic year 2012/2013. Dr. Mardini have many publications in the area of network survivability, wireless and wireless sensor networks, and optical-wireless networks. He is currently working on Wireless Mesh Networks, Wireless Sensor Networks, Optical Network Survivability, WiMax Technology, Scheduling in Parallel Computing and Intrusion Detection in database Techniques.



Abdulaziz Alraddadi is the supervisor of Taibah University branch in Yanbu. Dr. Alraddadi received his bachelor and master degree in Electrical Engineering from USA in 1995 and 1997, respectively. He received his Ph.D. degree in Computer Engineering from University Wayne State University-Michigan in 2003. Dr. Alraddadi have many academic and administrative experiences. He was the dean for the collage of science and computer engineering/Taibah University for the last three years, before that he was the dean for the technical collage of Yanbu for another 3 years. Dr. Alraddadi have many publication and participation in international journal and conferences. His main field is the integration between the computer engineering and the information technology to improve the electronic learning quality.



Praise Worthy Prize



1828-6011(201409)9:9;1-E