In the Name of God

Journal of
Information Systems & Telecommunication

Vol. 2, No. 2, April-June 2014, Serial Number 6

Research Institute for Information and Communication Technology
Iranian Association of Information and Communication Technology

Affiliated to: Academic Center for Education, Culture and Research (ACECR)

Manager-in-charge: Habibollah Asghari, Assistant Professor, ACECR, Iran
Editor-in-chief: Masoud Shafiee, Professor, Amir Kabir University of Technology, Iran

Administrative Manager: Shirin Gilaki
Executive Assistant: Behnoosh Karimi
Website Manager: Maryam sadat Tayebi
Art Designer: Amir Azadi
Print ISSN: 2322-1437
Online ISSN: 2345-2773
Publication License: 91/13216

Editorial Office Address: No.5, Saeedi Alley, Kalej Intersection., Enghelab Ave., Tehran, Iran,
P.O.Box: 13145-799 Tel: (+9821) 88930150 Fax: (+9821) 88930157
Email: info@jist.ir
URL: www.jist.ir

Indexed in:

- Journal of Information Systems and Telecommunication	www.jist.ir
- Islamic World Science Citation Center (ISC)	www.isc.gov.ir
- Scientific Information Database (SID)	www.sid.ir
- Regional Information Center for Science and Technology (RICeST)	www.srlst.com
- Magiran	www.magiran.com
- CIVILICA	www.civilica.com

Publisher:
Regional Information Center for Science and Technology **(RICeST)**
Islamic World Science Citation Center **(ISC)**

This Journal is published under scientific support of
Advanced Information Systems (AIS) Research Group and
Digital Research Group, ICTRC

Acknowledgement

JIST Editorial-Board would like to gratefully appreciate the following distinguished referees for spending their invaluable time and expertise in reviewing the manuscripts and their constructive suggestions, which had a great impact on the enhancement of this issue of the JIST Journal.

(A-Z)

- Abdolvand Neda, Alzahra Univesity, Tehran, Iran
- Abutalebi Hamid Reza, Yazd University, Yazd, Iran
- Akbarizadeh Gholamreza, Shahid Chamran University of Ahvaz, Ahvaz, Iran
- Baghali Khanian Zahra, Amirkabir University of Technology, Tehran, Iran
- Bobarshad Hossein, University of Tehran, Tehran, Iran
- Bohlooli Ali, Isfahan University, Isfahan, Iran
- Charmchi Hossein, Sharif University, Tehran, Iran
- Dastghaibyfard Gholam Hossein, Shiraz University, Shiraz, Iran
- Dehkhoda Parisa, Amirkabir University of Technology, Tehran, Iran
- Dokhanchi Seyed Hossein, Academic Center for Education Culture and Research (ACECR), Tehran, Iran
- Ebrahimi-Moghadam Abbas, Ferdowsi University of Mashhad, Mashhad, Iran
- Emadi Mohammad Javad, Amirkabir University of Technology, Tehran, Iran
- Emami Hossein, Islamic Azad University Majlesi Branch, Isfahan, Iran
- Faghani Farhad, Islamic Azad University, Najafabad Branch, Isfahan, Iran
- Firouzmand Mohammad, Gholamian Mohammad Reza, Tehran, Iran
- Ghassabi Zeinab, Azad University, Science and Research Branch, Tehran, Iran, Tehran, Iran
- Gholamian Mohammad Reza, Iran University of Science and Technology, Tehran, Iran
- Haj Mirzaei Milad, Islamic Azad University, Science and Research Branch, Tehran, Iran
- Heydarian Mohsen, Azarbijan Shahid Madani University, Tabriz, Iran
- Honarvar Mahboobeh, Yazd University, Yazd, Iran
- Hosseini Seno Seyed Amin, Ferdowsi University of Mashhad, Mashhad, Iran
- Hourali Maryam, Maleke Ashtar university of Technology, Tehran, Iran
- Kabiri Peyman, Iran University of Science and Technology, Tehran, Iran
- Kamandar Mehdi, Kerman Graduate University of Technology
- Koofigar Hamid Reza, Isfahan University, Isfahan, Iran
- Mahboobi Esfanjani Reza, Sahand University of Technology, Tabriz, Iran
- Matinfar Mehdi, Tarbiat Modares University, Tehran, Iran
- Mohammad Khanli Leyli, University of Tabriz, Tabriz, Iran
- Montazeri Mohammad Ali, Isfahan University of Technology, Isfahan, Iran
- Montazery Kordy Hossein, Babol Noshirvani University of Technology, Babol, Iran
- Moradi GholamReza, Amirkabir University of Technology, Tehran, Iran
- Naghibzadeh Mahmood, Ferdowsi University of Mashhad, Mashhad, Iran
- Nasersharif Babak, Khaje Nasir-edin Toosi University of Technology, Tehran, Iran
- Nazemi Eslam, Shahid Beheshti University, Tehran, Iran
- Sadeghiyan Babak, Amirkabir University of Technology, Tehran, Iran
- Saeedi Hamid, Tarbiat Modares University, Tehran, Iran
- Safarinejadian Behrouz, Shiraz University of Technology, Shiraz, Iran
- Samimi Hossein, Tarbiat Modares University, Tehran, Iran
- Saniee Abadeh Mohammad, Tarbiat Modares University, Tehran, Iran
- Shirali Shahreza Mohammad Hasan, Amirkabir University of Technology, Tehran, Iran
- Sojoodi Mehdi, Tarbiat Modares University, Tehran, Iran
- Taheri Mohammad, Shiraz University, Shiraz, Iran

Table of Contents

Assessment of Performance Improvement in Hyperspectral Image Classification Based on Adaptive Expansion of Training Samples

Maryam Imani
Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran
maryam.imani@modares.ac.ir
Hasan Ghasemian*
Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran
ghassemi@modares.ac.ir

Received: 07/Jul/2013 Accepted: 12/Apr/2014

Abstract
High dimensional images in remote sensing applications allow us to analysis the surface of the earth with more details. A relevant problem for supervised classification of hyperspectral image is the limited availability of labeled training samples, since their collection is generally expensive, difficult and time consuming. In this paper, we propose an adaptive method for improving the classification of hyperspectral images through expansion of training samples size. The represented approach utilizes high-confidence labeled pixels as training samples to re-estimate classifier parameters. Semi-labeled samples are samples whose class labels are determined by GML classifier. Samples whose discriminator function values are large enough are selected in an adaptive process and considered as semi-labeled (pseudo-training) samples added to the training samples to train the classifier sequentially. The results of experiments show that proposed method can solve the limitation of training samples in hyperspectral images and improve the classification performance.

Keywords: Classification, Hyperspectral Image, Limited Training Data, Pseudo-Training Samples.

1. Introduction

With the development of the remote sensing imaging systems and hyperspectral sensors, the use of hyperspectral image is becoming more interesting. The objective of analysis is to associate each pixel in a hyperspectral image with a proper label. The basis of a classification system is illustrated in Fig. 1 [1]. The increasing of spectral resolution provided by the new sensor technology has brought about new potentials and challenges to data analysis. It is possible to identify more details about classes with higher accuracy than would possible with the data from earlier sensors using a large number of available spectral bands. One the other hand, in order to fully utilize the information contained in the new features, a large number of training samples are needed for using a large number of interesting classes and a large number of available spectral bands. Unfortunately, obtaining of training samples is generally expensive and difficult. When the number of available training samples is relatively small with respect to the number of features, curse of dimensionality problem, Hughes phenomenon occurs [2]. When the given training set is fixed and the dimension of the space grows, the classification accuracy reaches a maximum point and then decreases. When a new feature is added to the data (and the number of training samples is as before) the Bayes error decreased, but at the same time the bias of the classification error increases. This increase is due to the fact that more parameters must be estimated from the same number of available training samples. If the increase in the bias of the classification error is more than the decrease in the Bayes error, then, using of the additional feature degrades the performance of the decision rule. This effect is called the Hughs phenomenon. Therefore, design of the hyperspectral image classifiers that can deal with the small training set has become interesting recently. The suggested solutions can be divided into four categories: 1) fusion of spatial and spectral information [3]-[4];2) feature reduction [5]-[7];3) low complexity classifiers, such as support vector machine (SVM) [8]-[9];4) enlarging the training set by semi-supervised learning. Estimates with smaller covariance matrices can be found by using additional semi-labeled samples. Therefore, better performance can be acquired without the extra cost for selection of more training samples. We focus on the fourth solution in this paper. While the collection of labeled samples is generally time consuming and difficult, unlabeled samples can be generated in a much easier way. Then, the idea of using semi-supervised learning techniques is formed. The main assumption of such methods is that the new training samples can be obtained from a limited set of available labeled samples. A survey of this algorithm is represented in [10]. An adaptive classifier for mitigating the small training sample problem is proposed in [11]. This adaptive classifier that is based on decision fusion, enhances estimation and thus improves the classification accuracy by utilizing the classified samples (referred as semi-labeled samples), in addition to the original training samples. In this classifier, learning of classifier is performed at two steps. At the

* Corresponding Author

beginning of this method, partitioning of observation space is done and several groups of bands are produced. After providing the primary decisions, several rules are used in decision fusion to determine the final label of pixels. In [12], authors used some contextual information such as correlation between a sample and its neighbors for deletion of outlier samples. Then, the semi-labeled pixels are selected from appropriate region. An ensemble algorithm which combines generative model (mixture of Gaussians) and discriminative classifier (support cluster machine) is proposed in [13]. In [14], authors defined a novel composite semi-supervised classifier based on SVMs specifically designed for addressing spectral–spatial categorization of hyperspectral data. In this paper, we use an adaptive classifier for increase the training samples size. Among unlabeled samples whose labels are determined after classification, just samples that their discriminator functions are large enough in Gaussian maximum likelihood (GML) decision rule are selected as semi-labeled samples because they have higher confidence than other classified samples.

The reminder of this paper is organized as follows: some related works about semi-supervised methods are reviewed in section 2. Section 3 describes the suggested approach for increase the training samples size. Section 4 presents the experimental results. Finally, conclusion of this paper is given in section 5.

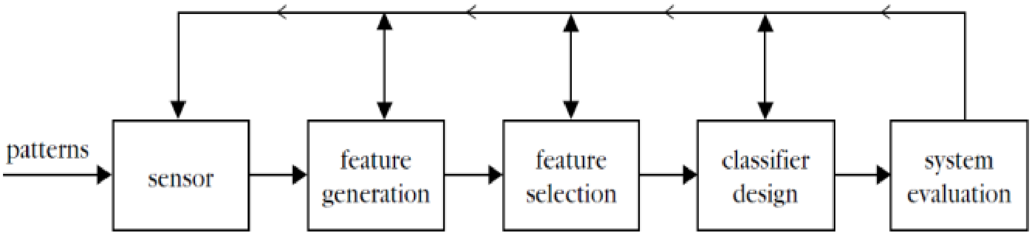


Fig. 1: Basis of classification system [1]

2. Some Semi-supervised Approaches

The much information contained in hyperspectral images, allows to characterize and classify the land-covers with more details and improved accuracy and reliability. But the high number of spectral features and the low number of training samples pose the Hughes phenomenon. In the remote sensing applications, many supervised and unsupervised classifiers have been developed to tackle the hyperspectral image classification problem. The main difficulty with supervised method is that performance is basically depends on the quality of training samples. The enough training samples is not available and this is another difficulty. On the other hand, unsupervised methods are not sensitive to the number of training samples, but the relation between clusters and classes is not ensured. Because of represented reasons, it is natural that we use semi-supervised methods for improving of performance. Authors in [15] introduce a semi-supervised graph-based method. Their proposed method has the following characteristics: 1- this method is kernel based and thus the high dimensionality problem is mitigated. 2- The huge number of unlabeled samples is exploited to improve performance. 3- Using graph-based methodology, the relative importance to the labeled samples is considered naturally. 4- The contextual information is incorporated using a family of composite kernels. The graph-based method can be interpreted as a graph $G = (V, E)$ defined on $\mathcal{X} \in \mathbb{R}^N$ ($\mathcal{X}$ denotes a dada set of pixels in a N dimensional space) where the vertex set V is just $\mathcal{X}$ and the edges E are weighted by W. The weight matrix W is constructed among all labeled and unlabeled samples and matrix S is defined as follows:

$$S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{1}$$

where D is a diagonal matrix with its (i, i) element equal to the sum of the ith row of W. Given n unlabeled samples, A $n \times c$ matrix F corresponds to a classification on the dada set $\mathcal{X}$ by labeling each point $\mathrm{x}_i \in \mathcal{X}$ with a label $y_i = arg\,\max_{j \le c} F_{ij}$ (c is the number of classes). F can be understood as a vectorial function which assigns a vector F_i to each point x_i. A $n \times c$ matrix Y is defined as follows:

$$\mathrm{Y_{ij}} = \begin{cases} 1 & \mathrm{y_j = j} \\ 0 & \mathrm{y_j \neq j} \end{cases} \tag{2}$$

In the proposed graph-based method in [14], following spreading function is iterated until convergence:

$$F(t+1) = \alpha S F(t) + (1 - \alpha) Y \tag{3}$$

where $0 < \alpha < 1$ specifies the relative amount of the information from its neighbors and its initial label information.

In semi-supervised methods, if the unlabeled samples are not properly selected, those may confuse the classifier and reduce the classification accuracy. Thus, it is important that the most highly informative unlabeled samples are identified. In [16], a semi-supervised method is proposed that uses self-learning framework. The proposed method in [16] is based on two steps. In the first step, a candidate set, consist of labeled and unlabeled samples, is selected using a self-learning strategy based on spatial information. In the second step, the standard active learning algorithms on the previously derived candidate set is run to automatically select the most informative samples from the candidate set. Spatial information can be adopted as a reasonable criterion to select unlabeled samples in the proposed method in [16]. First a probabilistic classifier is used to produce a global classification map. Then, the neighbors of the labeled training samples based on a local similarity assumption, are identified and the by analyzing the spectral similarity of spatial neighbors with regard to the original labeled samples, the candidate set is computed. In this method, candidate set is obtained based on spectral and spatial information and its samples are highly reliable. After obtaining candidate set, the most informative unlabeled samples are selected automatically using active learning algorithms and then newly obtained labeled and unlabeled training samples are finally used to retain the classifier. This procedure is repeated iteratively until a convergence criterion.

A semi-labeled bagging technique is proposed by authors in [17]. The novelties of the bagging technique in [17] lie in the definition of a general classification strategy for ill-posed problems by the joint use of training and semi-labeled samples and the design of an effective bagging method (driven by semi-labeled samples) for a proper exploitation of different classifiers based on bootstrapped hybrid training sets. In the bagging algorithms, subsets of bootstrapped samples are generated, and a classifier is built from each subset. The final classification map is obtained by an ensemble rule to achieve a better classification result than the single classifier. The proposed architecture in [17] includes an initial classifier which only exploits the training set to generate the initial classification map. In this way, unlabeled samples become semi-labeled. Then, the generic *b*th classifier of the architecture is defined by selecting semi-labeled samples obtained from the previous classifier. This process is iterated until the desired number of classifiers included in the ensemble of classifiers.

Another semi-supervised method is proposed in [18]. An iterative procedure to produce accurate classification map using a hierarchical segmentation is done. The proposed approach uses the active learning strategies to select the most informative pixels to be labeled. The proposed method in [18] exploits simultaneously the data structures obtained by unsupervised segmentation and information contained in labeled samples. This method uses the available labeled information directly to find the most probable classification map in a hierarchical clustering structure.

3. Proposed Adaptive Method for Classification

In this paper, original training samples are samples whose class labels are correctly known and used for training of classifier, i.e. for estimate of mean vectors and covariance matrices in discriminator function of classifier for all classes. Semi-labeled samples that we call them pseudo-training samples are samples whose class labels are determined by a decision rule. They are unlabeled samples before implementation of classification and their class label information partially obtained after classification. The label for a semi-labeled sample can be either right or wrong.

Consider two different estimators $\hat{\theta}, \breve{\theta}$ with negligible biases, and assume that $cov(\hat{\theta}) \geq cov(\breve{\theta})$. The expected error by using $\hat{\theta}$ is greater than the expected error by using $\breve{\theta}$ in the decision rule [1], i.e.

$$E\{\hat{e}\} \geq E\{\breve{e}\} \tag{4}$$

By using additional semi-labeled samples, estimates with smaller covariance matrices can be found. If we know which samples have been correctly classified and use them accordingly to re-estimate statistics in addition to original training samples, the estimated statistics should be more precise because the training samples set has been enlarged. In this section, we propose an adaptive method for classification of hyperspectral image using limited training samples. Our used classification method is Gaussian maximum likelihood (GML) that its discriminator function is:

$$g_i(X) = -\ln(|\Sigma_i|) - (X - M_i)^T \Sigma_i{}^{-1}(X - M_i) \tag{5}$$

where $(M_i)_{n_b \times 1}$ and $(\Sigma_i)_{n_b \times n_b}$ are respectively the mean vector and covariance matrix for class i which are calculated using training samples and $X_{n_b \times 1}$ denotes the unlabeled vector. Also the number of features is denoted by n_b. The decision rule of GML is represented by:

$$if\ g_i(X) > g_j(X) \quad for\ all\ j \neq i \quad ,i,j \in \{1,2,\dots,N_c\}$$
$$then\ X \in w_i \tag{6}$$

Where N_c is the number of classes and $X \in w_i$ denotes that X belongs to class i. In the proposed method, at first, we classify the pixels of hyperspectral image using limited available training samples. Then, we find pixels that are labeled with high-confidence. These pixels were unlabeled before classification. We call these samples as pseudo- training samples and add them to original training set and use the new extended training set for re-classification in a sequential process.

Our criterion for selection of pseudo-training samples is high-confidence of semi-labeled samples. Assume, two unlabeled samples are labeled after classification and both of them get class label $(x_1, x_2 \in w_i)$. In this conditions, which sample is more appropriate to be considered as pseudo-training sample? The respond is as follows:

$$if\ g_i(x_1) > g_i(x_2),\ \ i \in \{1,2,\dots,N_c\}$$
$$then\quad prob\{x_1 \in w_i\} > prob\{x_2 \in w_i\}$$

where $prob\{\cdot\}$ denotes the probability of $\{\cdot\}$. Therefore, it is reasonable that x_1 is selected as a pseudo-training sample. But we want to obtain superlative candidates (high-confidence) for selection of pseudo-training samples. Therefore we define a threshold for discriminator function value. The proposed algorithm for classification of hyperspectral image is illustrated in Fig. 2.

The adaptive proposed method is described as follows:

Step 1- The hyperspectral image is classified using just original training samples according Equation (6).

Step 2- An appropriate threshold is obtained for selection of pseudo-training samples. We calculate the discriminator function for all training samples in all classes and locate their values in a matrix:

$$G = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1N_t} \\ g_{21} & g_{22} & \cdots & g_{2N_t} \\ \vdots & \vdots & \vdots & \vdots \\ g_{N_c1} & g_{N_c2} & \cdots & g_{N_cN_t} \end{bmatrix} \tag{7}$$

Where N_t is the number of training samples per class. An appropriate threshold is selected as follows:

$$Th = \min_{i=1:N_c}\left[\max_{n=1:N_t} g_{in}\right] \tag{8}$$

Step 3- selection of pseudo-training samples is done using obtained threshold in step 2 as follows:

$$if \ \ (\, g_i(X) > g_j(X) \ \ \& \ \ g_i(X) > Th)$$
$$for\ all\ j \neq i \quad , i,j \in \{1,2,\dots,N_c\}$$
$$then\ X\ is\ selected\ as\ high\text{-}confidence\ labeled$$
$$sample\ or\ pseudo\ training\ sample\ for\ class\ i$$

Step 4- classification is repeated using the new extended training set (original training samples plus the obtained pseudo-training samples in step 3).

We continue this sequential process to converge to the final values of accuracy and reliability.

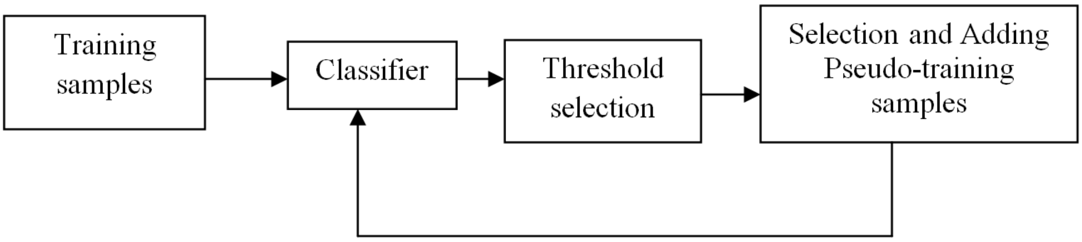


Fig. 2: Proposed algorithm for classification

4. Experiments

In order to evaluation of proposed method, several experiments are done. We use three datasets in our experiments. The first data is a synthetic image with size of 80×120. This synthetic scene comprises eight classes and 12 spectral bands which selected from a digital spectral library compiled by the U.S. Geological Survey (USGS) and available online [19]. The false-color image and class map of test data is shown in Fig. 3. The second data is a real multispectral image which is an agricultural segment of Indiana State (F210 dataset) [20]. This image contains 8 different farm classes and is provided in 12 bands with 256 grey levels. The third data is a real hyperspectral image. The hyperspectral data is Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) Indian pines image [20]. This image has 145×145 pixels and contains 16 classes that most of them are different types of crops. The AVIRIS sensor generates 220 spectral bands that we reduced the number of them to 190 by removing 30 absorption and noisy bands. In all experiments 16 random pixels per class are used as original training samples and GML classifier is applied for classification of images. For evaluation of classification performance, we should use some measures. The used measures in this paper are average accuracy, average reliability and kappa coefficient (KC). These measures can be represented as follows:

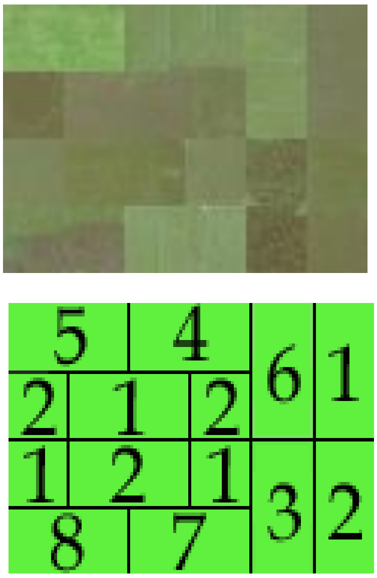


Fig. 3: Test data: (up to down) false -color image, class map.

$$average\ ccuracy = mean(P_1, P_2, \dots, P_{N_c}) \tag{9}$$

$$average\ reliability = mean(R_1, R_2, \dots, R_{N_c}) \tag{10}$$

where $mean$ return the mean value of elements. P_i and R_i $i \in \{1,2,\dots,N_c\}$ are defined as follows:

$$P_i = \frac{N_i}{A_{i1}} \ , R_i = \frac{N_i}{A_{i2}} \tag{11}$$

where N_i is the number of test samples of class i that are correctly classified. A_{i1} denotes the total number of test samples that belongs to class i and A_{i2} is the total number of test samples that classified in class i. The KC is computed as follows [21]:

$$KC = \frac{N\sum_{c=1}^{n_c} t_{cc} - \sum_{c=1}^{n_c} t_{c+}t_{+c}}{N^2 - \sum_{c=1}^{n_c} t_{c+}t_{+c}} \tag{12}$$

where N denote the number of testing samples and n_c is the number of classes. t_{cc} denotes the number of samples correctly classified in class c, t_{c+} is the number of testing samples labeled as class c, and t_{+c} is the number of samples predicted as belonging to class c.

The obtained values of accuracy and reliability versus the iteration of algorithm in classification of test data are illustrated in Fig. 4. We can see from Fig. 4 that accuracy after 13 iterations and reliability after 12 iterations are converged to their final values. Confusion matrices acquired from test data for GML classifier and proposed classifier are shown in table I and table II respectively. Without any adding pseudo-training samples and just using original training set, 0.76 accuracy and 0.76 reliability were obtained from GML classifier. With using proposed classifier, we can obtain 0.89 accuracy and 0.88 reliability in classification of test data.

Tables III and IV show the confusion matrices for classification of F210 data using GML and proposed classifiers respectively. In this experiment, 0.75 accuracy and 0.59 reliability are acquired from GML classifier. Also 0.83 accuracy and 0.73 reliability are obtained using proposed method. Fig. 5 shows the classification maps using conventional GML classifier and the proposed adaptive classifier for F210 dataset. In both of used dataset (test data and F210 data), the performance of proposed classifier is better than GML classifier considerably.

We repeat the represented experiment for Indian dataset. Because the number of necessary training samples for training the GML classifier is equal to the number of features plus one, we have to reduce the number of features to 15 for using 16 training samples in our experiment. The conventional feature extraction method, maximum noise fraction (MNF), is used for feature reduction. The obtained average accuracy and average reliability versus the number of iterations in proposed adaptive procedure is shown in Fig. 6. As seen from Fig. 6, the accuracy and reliability of classification attain a maximum value after 3 iterations and after that, the performance of classifier is decreased and converges to a less value of efficiency. Doing a feature extraction process before classification of hyperspectral image may cause this behavior for Indian data set that is partially different from multispectral data (test and F210 data sets). Further, if the unlabeled samples are not properly selected, added semi-labeled samples may confuse the classifier. Thus, they may reduce the classification performance. Beside conventional GML classifier, we also compare our proposed method with semi-labeled-sample-driven bagging technique which proposed in [17]. The classification maps using conventional GML classifier, bagging technique and adaptive proposed method are illustrated in Fig. 7.

A summary of classification results which contain average accuracy, average reliability and kappa coefficient for all datasets and three methods (conventional GML classifier, bagging technique [17] and proposed method) are represented in table V. one sees from this table that proposed method has the better performance than conventional GML classifier and the bagging method. The selection of reliable semi-labeled samples with high confidence, which obtained by using a proper threshold, makes our proposed method as an interesting technique for small sample size situation.

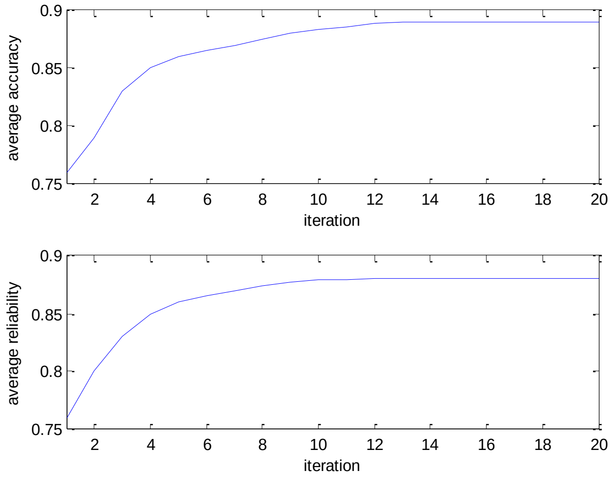


Fig. 4: The average accuracy and average reliability versus the iteration of algorithm (for test data).

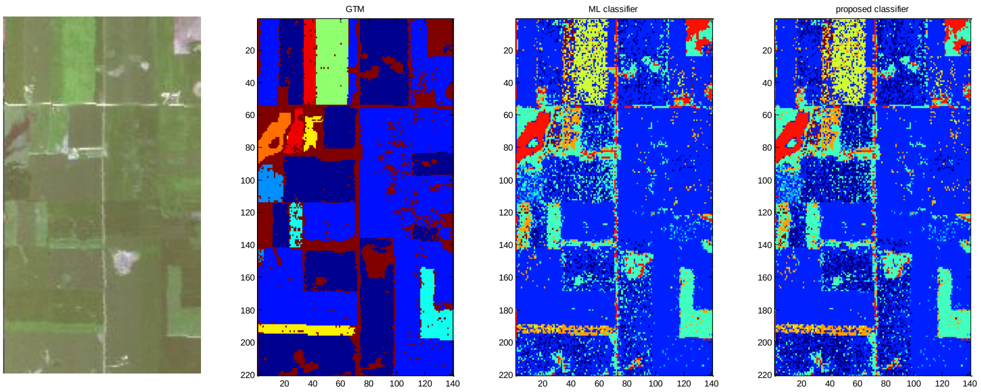


Fig. 5: The result class maps for F210 data set: (Left to right) false -color image, Ground Truth Map (GTM), the result of conventional GML classifier, the result of proposed classifier.

Table I: Confusion Matrix of Conventional GML classifier (test data)

ML classifier	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	accuracy
Class 1	2133	4	5	69	0	8	1	180	0.89
Class 2	8	1940	305	1	7	58	0	81	0.81
Class 3	2	62	555	33	0	10	1	137	0.69
Class 4	7	0	1	633	0	75	0	84	0.79
Class 5	0	0	16	0	736	0	0	48	0.92
Class 6	0	0	1	15	4	538	70	172	0.67
Class 7	0	0	0	11	0	407	290	92	0.36
Class 8	4	0	24	16	2	21	8	725	0.91
reliability	0.99	0.97	0.61	0.81	0.98	0.48	0.78	0. 48	**Average accuracy** **0.76**
									Average reliability **0.76**

Table II: Confusion Matrix of proposed classifier (test data)

Proposed classifier	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	accuracy
Class 1	2291	23	1	45	0	3	0	37	0.95
Class 2	0	2342	50	0	0	3	1	4	0.98
Class 3	7	45	666	48	0	1	0	33	0.83
Class 4	5	2	1	784	0	8	0	0	0.98
Class 5	0	0	1	1	787	6	2	3	0.98
Class 6	0	9	3	32	0	593	159	4	0.74
Class 7	0	1	0	21	0	215	552	11	0.69
Class 8	0	3	6	27	0	10	5	749	0.94
reliability	0.99	0.97	0.91	0.82	1	0.71	0.77	0. 89	**Average accuracy** **0.89**
									Average reliability **0.88**

Table III: Confusion Matrix of Conventional GML classifier (F210 data)

ML classifier	**Corn**	**Soybeans**	**Woods**	**Wheat**	**Sudex**	**Oats**	**Pasture**	**Hay**	accuracy
Corn	7110	648	798	248	0	9	0	362	0.77
Soybeans	61	8509	1538	322	358	816	0	233	0.72
Woods	27	39	270	12	0	3	1	3	0.76
Wheat	25	25	47	594	0	78	10	45	0.72
Sudex	0	105	3	2	971	102	0	11	0.81
Oats	1	70	11	38	35	334	0	54	0.62
Pasture	1	6	6	4	0	4	316	3	0.93
Hay	14	52	12	36	15	85	0	439	0.67
reliability	0.98	0.90	0.10	0.47	0.70	0.23	0.97	0. 38	**Average accuracy** **0.75**
									Average reliability **0.59**

Table IV: Confusion Matrix of proposed classifier (F210 data)

Proposed classifier	**Corn**	**Soybeans**	**Woods**	**Wheat**	**Sudex**	**Oats**	**Pasture**	**Hay**	accuracy
Corn	7856	917	135	126	0	1	0	139	0.86
Soybeans	17	10534	810	14	219	213	0	30	0.89
Woods	20	109	210	10	1	1	4	1	0.59
Wheat	2	29	6	714	0	57	15	1	0.87
Sudex	1	12	0	2	1144	32	0	3	0.96
Oats	4	26	0	21	51	417	0	23	0.77
Pasture	0	0	0	1	0	0	338	0	1
Hay	45	37	0	34	38	36	1	462	0.71
reliability	0.99	0.90	0.18	0.77	0.79	0.55	0.94	0. 70	**Average accuracy** **0.89**
									Average reliability **0.88**

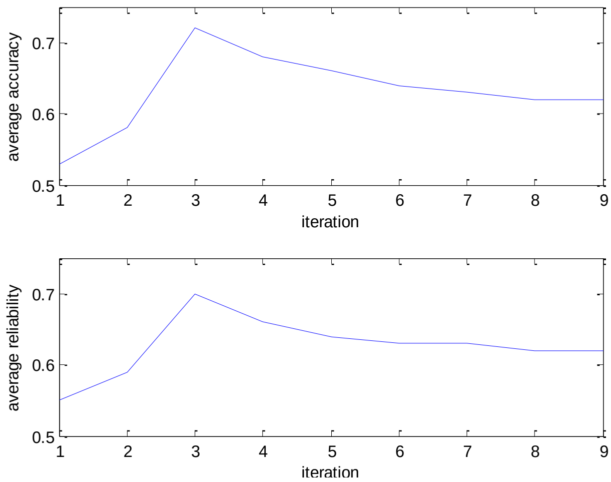


Fig. 6: The average accuracy and average reliability versus the iteration of algorithm for Indian data set

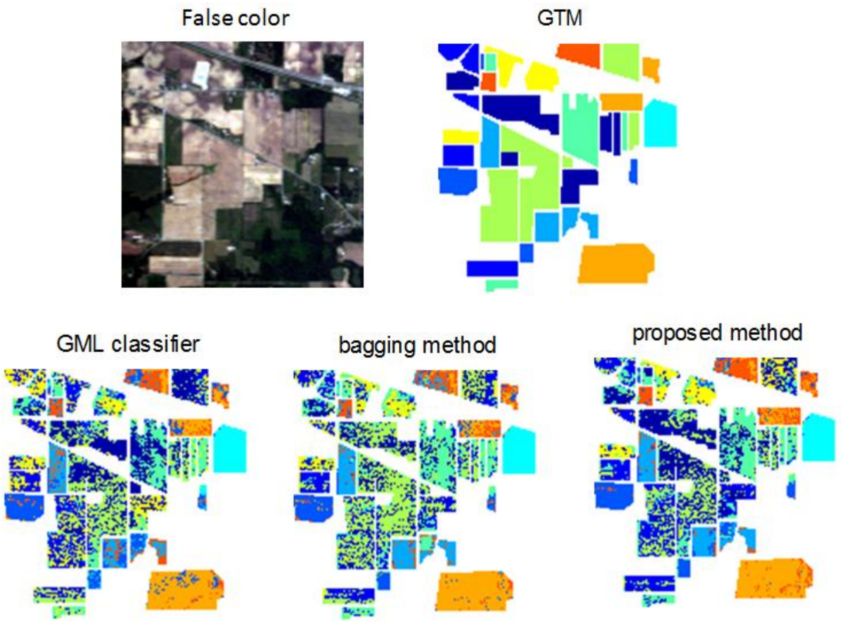


Fig. 7: The result class maps for Indian data set: (Left to right and up to down) false color image, GTM, classification map using GML classifier, classification map using bagging technique, classification map using proposed method.

Table V: Summary of classification results

dataset	Method	Average accuracy	Average reliability	Kappa coefficient
test data	Conventional GML	0.76	0.76	0.65
	Bagging technique [17]	0.82	0.81	0.73
	Proposed method	0.89	0.88	0.78
F210	Conventional GML	0.75	0.59	0.48
	Bagging technique [17]	0.79	0.68	0.60
	Proposed method	0.83	0.73	0.69
Indian	Conventional GML	0.62	0.60	0.54
	Bagging technique [17]	0.67	0.66	0.63
	Proposed method	0.72	0.70	0.67

5. Conclusions

For a limited number of available training samples, the classification performance is decreased as the number of features (spectral bands) is increased. This is an important challenge especially in hyperspectral data sets where the ratio of available training samples to dimension of data is small. In this paper, we proposed an adaptive method for classification of hyperspectral images to solve the limitation of available training samples. We select high-confidence labeled samples after primary classification and consider them as semi-labeled (pseudo-training) samples. The selected pseudo-training samples are added to the original training samples and the new extended training set is used to re-estimate the statistics of classifier. This process is continued sequentially until the accuracy and reliability of classifier converge to the final values for multispectral data or gain the maximum accuracy and reliability (the best possible performance) for hyperspectral images. Our experiment results show that proposed method has better performance than conventional GML classifier and semi-labeled-sample-driven bagging technique. The proposed method can be an effective solution to cope with the small sample size situation.

References

[1] Q. Z. Jackson and D. Landgrebe, “Design of an Adaptive Classification Procedure for The Analysis of High-dimensional Data With Limited Training Samples,” TR-ECE 01-5, Purdue University, West Lafayette, Indiana, Dec. 2001.

[2] G. F. Hughes, “On the mean accuracy of statistical pattern recognition,” IEEE Trans. Inf. Theory, Vol. IT-14, No. 1, Jan. 1968, pp. 55–63.

[3] T. Achalakul and S. Taylor, “Real-time multispectral image fusion,” Concurr. Computat.: Pract. Exper., Vol. 13, No. 12, Sep. 2001, pp. 1063–1081.

[4] A. Keshavarz, H. Ghassemian1, and H. Dehghani, “Hierarchical classification of hyperspectral images by using SVMs and "same class neighborhood property",” IEEE Symposium on Geoscience and Remote Sensing, Vol. 5, July 2005, pp. 3219 – 3222.

[5] A.C. Jensen and A.S. Solberg, “Fast hyperspectral feature reduction using piecewise constant function approximations,” IEEE Trans. Geosci. Remote Sens. Lett., Vol 4, No. 4, Oct. 2007, pp. 547–551.

[6] L. O. Jimenez and D. A. Landgrebe, “Hyperspectral data analysis and feature reduction via projection pursuit,” IEEE Trans. Geosci. Remote Sensing, Vol. 37, Nov. 1999, pp. 2653–2667.

[7] C. Cariou, K. Chehdi, and S. Le Moan, “BandClust: An Unsupervised Band Reduction Method for Hyperspectral Remote Sensing,” IEEE Trans. Geosci. Remote Sens. Lett., Vol. 8, No. 3, MAY 2011, pp. 565-569.

[8] C.-H. Li, B.-C. Kuo, C.-T. Lin, and C.-S. Huang, “A Spatial–Contextual Support Vector Machine for Remotely Sensed Image Classification,” IEEE Trans. Geosci. Remote Sensing, Vol. 50, No. 3, March 2012.

[9] F. Bovolo, L. Bruzzone, and Lorenzo Carlin, “A Novel Technique for Subpixel Image Classification Based on Support Vector Machine,” IEEE Trans. Image Proces. Vol. 19, No. 11, Nov. 2010, pp. 2983 - 2999.

[10] X. Zhu, “Semi-supervised learning literature survey,” Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, Tech. Rep. 1530, 2005.

[11] H. Dehghani and H. Ghassemian, “Adaptive Classification of Hyperspectral Image,” 12th Iranian Conference on Electrical Engineering, 11-13 May, 2004.

[12] A. Keshavarz and H. Ghassemian, “An Improving in Estimate of Classifier Parameters for Hyperspectral Images Using Semi-labeled Samples and Deletion of Outlier Samples”, 16th Iranian Conference on Electrical Engineering, 13-15 May, 2008, pp. 306-311, (Printed in Persian).

[13] M. Chi, Q. Kun, J.-A. Benediktsson, and Rui Feng, “Ensemble Classification Algorithm for Hyperspectral Remote Sensing Data,” IEEE Trans. Geosci. Remote Sens. Lett, Vol. 6, No. 4, oct. 2009, pp. 762-766.

[14] M., G. Camps-Valls, and L. Bruzzone, “A Composite Semisupervised SVM for Classification of Hyperspectral Images,” IEEE Trans. Geosci. Remote Sens. Lett, Vol. 6, No. 2, April 2009, pp. 234-238.

[15] G. Camps-Valls, T. Bandos Marsheva, and D. Zhou, “Semi-Supervised Graph-Based Hyperspectral Image Classification,” IEEE Trans. Geosci. Remote Sensing, Vol. 45, No. 10, Oct. 2007, pp. 3044 - 3054.

[16] I. Dopido, J. Li, P.R. Marpu, A. Plaza, J.M. Bioucas Dias, and J.A. Benediktsson, “Semisupervised Self-Learning for Hyperspectral Image Classification,” IEEE Trans. Geosci. Remote Sensing, Vol. 51, No. 7, July 2013, pp. 4032-4044.

[17] M. Chi and L. Bruzzone,“A semilabeled-sample-driven bagging technique for ill-posed classification problems,” IEEE Geosci. Remote Sensing lett., Vol. 2, No. 1, Jan. 2005, pp. 69-73.

[18] J. Munoz- Mari, D. Tuia, and G. Camps- Valls, “Semisupervised Classification of Remote Sensing Images With Active Queries,” IEEE Trans. Geosci. Remote Sensing, Vol. 50, No. 10, Oct. 2012, pp. 3751 - 3763.

[19] http://speclab.cr.usgs.gov/spectral-lib.html.

[20] D. A. Landgrebe, Signal Theory Methods in Multispectral Remote Sensing, Hoboken, NJ: Wiley, 2003.

[21] J. Cohen,“A coefficient of agreement from nominal scales,” Edu. Psychol. Meas., Vol. 20, No. 1, 1960, pp. 37–46.

Maryam Imani received the B.Sc. and M.Sc. Degrees in Electrical Engineering from Shahed University, Tehran, Iran in 2009 and 2011 respectively. She is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering at the Tarbiat Modares University, Tehran, Iran. Her research interests include the hyperspectral image analysis, feature extraction and pattern recognition in remote sensing applications.

Hasan Ghasemian received the B.S.E.E. degree from Tehran College of Telecommunication in 1980 and the M.S.E.E. and Ph.D. degree from Purdue University, West Lafayette, USA in 1984 and 1988 respectively. Since 1988, he has been with Tarbiat Modares University in Tehran, Iran, where he is a Professor of Computer and Electrical Engineering. His research interests focus on Multi-Source Signal/Image Processing and Information Analysis and Remote Sensing.

Internet Banking, Cloud Computing: Opportunities, Threats

Monireh Hosseini*
Department of Industrial Engineering, K.N.Toosi University of Technology, Tehran, Iran
hosseini@kntu.ac.ir
Elias Fathi Kiadehi
Department of Industrial Engineering, K.N.Toosi University of Technology, Tehran, Iran
eliasfathi@gmail.com

Received: 25/Sep/2013 Accepted: 12/Apr/2014

Abstract
With the extension of Internet and its applications, internet banking is introduced as an efficient and cost effective way to provide services to customers. Towards the end of previous decade, cloud computing has been offered as a revolution in Internet application as a service which effect on the way that service is provided. Regarding the service improvement based on customer's needs, cloud computing is a quick move in informational services. This study tried to consider each aspect of internet banking and cloud computing strengths, weaknesses, opportunities and threats and provide SWOT analysis for Internet banking using cloud computing. In the following, the study tried to provide a practical solution for financial agencies and banks to provide better Internet banking services using cloud computing technology. Finally a SWOT analysis of internet banking using cloud computing technology is discussed and approved with expert opinions using fuzzy Delphi method.

Keywords: Cloud Computing, Internet Banking, Strengths, Weaknesses, Opportunities, Threats.

1. Introduction

The advent of internet and its applications caused a revolution in service provision in financial sector. This revolution in financial services caused changes in banking service provision leading to internet banking [1]. Using cost effective and helpful solution, internet banking caused a decrease in the time required for financial activities [2]. This technology resulted in bank customers' cooperation in banking activities such as payments, statements, account information review, money transfer and etc. without having any physical locations [3]. Internet banking lead to a reduction in costs related to other banking situations and provides complete and useful customer information [4]. Internet banking is widely adopted by most countries; the degree of progress in this technology in pioneer countries is more than fifty percent's [5,6]. Business strategies, plans and policies should be reviewed in order to increase the performance and decrease operational costs [7,8,9, and 10]. The power that leads to move toward using internet banking causes a break in adoption barriers, creates new products and services and provides opportunities for internet banking [11].

Based on NIST (National Institute of Standard and Technology) definition, cloud computing is a model to access, to share and to configure computing resources such as networks, servers, storage area, and software and

Services that form the internet [12]. Cloud computing is a new computing method in Information Technology provision [13]. Provided services by cloud computing are focused on two factors of quality and low costs. Customers can increase or decrease their needed services. Customers should pay based on their usage and they are able to increase or decrease the amount of usage rate and shared resources [14]. By workload division on different centers, cloud computing is able to optimize IT infrastructure usage. Also, users are able to access to services from anywhere and anytime [15].

Aim of this paper is to survey and extract cloud computing and internet banking strengths ,weaknesses, opportunities and threats (SWOT) and use these SWOT to provide a SWOT analysis for internet banking using cloud computing technology. Aim of this SWOT analysis is to show which opportunities are created by cloud computing strengths, which strengths will decrease the threats, which opportunities are lost by cloud computing weaknesses and which threat is created by cloud computing weaknesses. In section one, an introduction to cloud computing and internet banking and in section two, paper methodology in order to collect required data is provided. In section three cloud computing features is discussed and cloud computing and internet banking literature is provided. In section four and five a SWOT analysis of internet banking using cloud computing technology is discussed and approved with expert opinions using fuzzy Delphi method.

2. Methodology

In this article, cloud computing subject is considered and is searched for the related articles in Science Direct indexing database. The objective of this search was on

* Corresponding Author

cloud computing strengths, weaknesses, opportunities and threats (SWOT).

Twenty three articles with cloud computing subject were selected in this database and cloud computing SWOT is extracted with their comparison.

In internet banking issue, twenty one articles were found in Science Direct indexing database. With assessment of these articles, internet banking SWOTs was extracted.

In Internet banking using cloud computing technology issue, two articles are found in Science Direct indexing database. Regarding these articles, internet banking using cloud computing technology SWOTs is discussed and resulted. In addition, some recommendations are proposed, accordingly.

In order to assess results' correctness and to obtain experts opinions, fuzzy Delphi method has been used. Fuzzy Delphi method was developed in the 1980s. The application of this approach is to make decision and consensus on issues that are not explicitly specified goals and parameters, can lead to very significant results. Feature of this method, providing a flexible framework that many of the barriers related to lack of precision and clarity are covered.

3. Literature Review

NIST defined five qualities for cloud computing including "Widely network access", "Resource pooling", "Rapid elasticity" and "Measured services". Also, NIST defined four deployment models for cloud computing including public cloud, private cloud, community cloud and hybrid cloud. In Fig. 1 these deployment models are shown and will be described in table 1.

Based on NIST definition, cloud services can be categorized in three levels which are described in table 2.

From 1990 to now, many works are done about internet banking. Sathye [18] described the main problems of internet banking adoption as security issues, lack of knowledge about internet banking and unreasonable price.

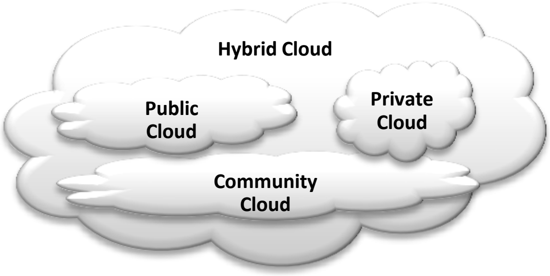


Fig. 1. Cloud computing deployment model

Table 1. Cloud computing deployment models

Public cloud	Public cloud is shared infrastructures among different users and these services are provided for anyone [14].
Private cloud	In private cloud, IT infrastructure is provided and supported for special organizations. In this model, there is no sharing in hardware or software among different users. This service can be inside an organization [16] and customers are responsible for this service management.
Community cloud	In community cloud, Cloud infrastructure is shared among a specific community which has special mission or goals, such as military organizations, banks.
Hybrid cloud	Hybrid cloud is a combination of other models for special purpose.

Table 2. Cloud Computing provided services

Infrastructure as a service	This service includes virtual servers which are run on virtual environments [17]. Customers have full controls on operating systems, storage areas, software and some configurations. Also, customers can provide their needed services based on their required computing power, storage area and required resources. Amazon EC2 service is an example of such service.
Platform as a service	In this service, users can deploy developed software by other users or create their own software using provided programming tools. Microsoft Azure is an example of such service.
Software as a service	This service includes provided software by cloud service providers in order to be used by customers [14]. Google Doc is an example of such service.

Howcraft [19], Liao [20], Akinci [21] described different factors of internet banking which have impacts on internet banking adoption. Some of these factors are, for instance, twenty four hour access, time performance, good quality services, current media support, security issues, and ease of use, transaction speed and user convenient.

Gerrard [22] used questionnaires to analyze customers' opinion about internet banking and described eight factors which prevent customers to adopt internet banking. Risks, lack of comprehensible requirements, service knowledge, and lack of access, face to face interaction and price related issues are some of these factors.

Sayer [23] described internet banking from customers' view and compared internet banking in Turkey and the UK. Focusing on cultural difference between Turkey and the UK in his article, he has done some researches on private financial information methods.

Laakkanen [24] assessed customers' attitude toward internet banking in Finland. Based on this article, resisting customers to internet banking form functional and psychological perspectives were evaluated. Resisting customers to internet banking were only measured from psychological perspectives. Non-resisting customers were more unpleased about information provided for internet banking.

Mirza [25] assessed internet banking adoption by private and public sectors in Iran. Based on this article, private banks were more successful in internet banking adoption by customers.

Subsorn [7] provided a comparative analysis for internet banking security from customers' perspective in Thailand and Sarakolaei [5] described four barriers to internet banking in Iran.

Lee [11] assessed internet banking and private financial information security in South Korea. In this article, he tried to focus on security issues and protection methods for private financial information.

Riffali [26], focuses on the acceptance factor and internet banking usage in Oman and Normalini [27] assessed the biometric technology impacts on the reduction of security problem. Based on this article, researches on biometrics lead to secure the logs process, to eliminate vulnerability problems and to reduce service desk call for password resets.

From 2009, cloud computing is noticed by many researchers because of its different applications. Owing to a wide variety of applications of this technology in all IT related issues, it can be used to accelerate the IT service uses.

Misra [27], worked on companies suitability in cloud adoption and Modeling its return of investments (ROI). This article tried to help companies for cloud adoption based on their specifications. In this article, some factors such as IT resources size, the amount of servers, the amount of user bases, IT annual revenue, the amount of covered countries, the percent of usage, data criticality, sensibility of work done by the company were used and by pointing to these factors, the researchers tried to provide some factors in order to create a suitability index. Finally, it tried to provide some solutions in ROI calculation which include cloud computing annual costs, saved costs, traditional costs, profit and etc.

Mastron [29], in his article tried to assess cloud computing from a business perspective. In his article, the strength, weakness, opportunities and threats of cloud computing for industries were identified. Also, different issues which affect cloud stakeholders were assessed. Some suggestions were provided for cloud service provides and managers.

Trang [30] focused on the role of cloud computing in competitive advantages improvement, and developed a research model for cloud computing form managerial perspective with a focus on small businesses. Also, the impact of cloud related resources in small businesses was assessed in this article.

Paquette [31], assessed security risks related to governmental uses of cloud computing. This article provided these risks as known or tangible risks including access, availability, infrastructure and integration and unknown or intangible risks including reliability, security, privacy and confidentiality, data location and etc.

Zissis [32] focused on cloud related security issues. Trust, security threats identification, confidentiality and privacy, integration, availability were some security factors which were focused in this article. Also, the writer tried to propose some security solutions for cloud computing challenges.

Subashini [33] in his lecture tried to assess security issues for each cloud provided services (SaaS, PaaS, IaaS) and current security solutions for these challenges.

Xunxu [34] worked on manufacturing based cloud computing. This article focused on cloud computing roles in manufacturing industries and its impacts on traditional manufacturing business models changes.

Zissis [35] in his article worked on electronic government and electronic voting security using cloud computing structures. In this article, the increase of complexity and corporation in electronic government services using cloud technology are assessed. In addition, they tried to identify cloud vulnerabilities using structural assessment. Finally, a high level solution for electronic government and electronic voting was presented using cloud computing solutions.

Aposta [36] worked on cloud computing modeling in banking system. In this article, he tried to analyze and assess cloud computing implementation in internet banking. In addition, the key requirements and tools of Cloud implementation are presented. Finally, he tried to identify business challenges in using cloud computing through a case study.

Bose [37] in his article, compared cloud computing and Internet banking form security and confidence perspective. He proposed that customers should have equal confidence about information storage in cloud computing and save money in internet banking. Some recommendations in technological, behavioral and regulatory aspects are presented which include “Critical security thinking”, “Access control and availability”, “confidentiality and privacy” and “long term viability and regulation”. These studies is summarized in table 3.

Table 3. Summary of studies related to Internet Banking and Cloud Computing

Author	Date	Research Fields	Results
Howcraft [19]	2002	Internet banking and customer attitude	• Different factors of internet which have impact on internet banking adoption
Liao [20]			
Akinci [21]	2004		
Gerrard [22]	2006	Internet banking and customer attitude	• Customer opinion about internet banking • Eight factors that prevent customers to adopt internet banking
Sayer [23]	2007	Internet banking adoption and customer attitude	• Internet banking from customer view • Comparison of cloud banking in UK and Turkey • Research on private financial information method
Laakkanen [24]	2009	Internet banking and customer attitude	• Customer attitude toward internet banking in Finland • Evaluation of resisting customers to internet banking from functional and psychological perspective.
Mirza [25]	2009	Internet banking adoption	• Internet banking adoption by private and public sectors in Iran
Subsorn [7]	2012	Internet banking security	• Comparative analysis for internet banking security from customers’ perspective in Thailand

Author	Date	Research Fields	Results
Sarakolaei [5]	2012	Internet banking adoption	• anking in Iran
Lee [11]	2011	Internet banking security	• Security issues for private financial information • Protection method for private financial information
Riffali [26]	2011	Internet banking adoption and customer attitude	• Acceptance factor and internet banking usage in Oman
Normalini [27]	2012	Internet banking security	• Biometric technology impacts on the reduction of security problems
Misra [27]	2010	Cloud computing Adoption	• Suitability index in cloud adoption and modeling its return of investment(ROI)
Mastron [29]	2010	Cloud computing SWOT and adoption	• Cloud computing SWOT • Some suggestion for cloud service providers and managers
Trang [30]	2010	Cloud computing SWOT and adoption	• Research model for cloud computing from managerial perspective • Impact of cloud related resources in small businesses
Paquette [31]	2010	Cloud computing security	• Security risks related to governmental use of cloud computing
Zissis [32] Subashini [33]	2010	Cloud computing security	• Cloud related security issues • Security solution for cloud computing challenges
Xunxu [34]	201	Cloud computing SWOT and adoption	• Cloud computing roles in manufacturing industries
Zissis [35]	2011	Cloud computing security	• High level solution for electronic government and electronic voting using cloud computing technology
Aposta [36]	2012	Internet Banking and Cloud Computing	• Key requirements and tools of Cloud implementation • Identify business challenges in using cloud computing
Bose [37]	2013	Internet Banking and Cloud Computing	• recommendations in technological, behavioral and regulatory aspects

4. The SWOT for Internet Banking and Cloud Computing

Cloud computing technology is introduced as a new way to provide services using internet. User convenient, performance, cost reduction and efficiency are the key factors which increased cloud applications and expansion in different services. After four years of cloud service introduction and assessment of every aspect of this technology, there is a good confidence from researchers, industries and customers about the provided services by this technology. Internet banking as a service can use cloud computing strengths and opportunities in order to improve their provided services for customers.

Internet banking can be provided using both public and private cloud. Each of these platforms has some opportunities and threats.

ING direct is one of famous internet banking systems which uses cloud Computing in order to expand its performance and decrease its costs.

Cloud computing has different strengths, weaknesses, opportunities and threats which are summarized in table 4.

Table 4. Cloud Computing SWOT

Strength	Weakness [39]	Opportunities	Threats
• Scalability [38][39] • Cost efficiency [38][40] • Efficiency [39][42] • Agility [40] • Availability [14] • Innovation [14]	• Security • Privacy • Structure • Performance • Financial • Legal • Learning	• Low Costs in IT infrastructure [29] • Cloud based e-education and e-learning [41] • Cloud based e-voting and e-government [35] • Cloud CRM and Cloud ERP [16] • Cloud based Tele-working • Green Cloud [42] • Service innovation	• Threat of Substitution [14] • Change in IT Culture [14] • The Loss of Physical control [14] • Critical Mission applications on Cloud environment [40]

Cloud computing strengths

For either high or low scale of computing demands, cloud computing shows its potential benefits. Traditional IT systems may fail against unpredicted demands. In contrast, cloud computing services can answer these demands quickly [38]. Unlimited capacity makes cloud computing flexible and responsive against changes. Quick responses against demands cause an improvement in IT service for both customers and organizations [39].

Primary benefits of cloud computing are cost saving as the main purpose of cloud computing to decrease purchase, maintenance and update costs of software, tools, and development area and transition of these costs to cloud providers. Because of the shared platform among different users, customers use a pay-as-you-go model that decreases the capital expenditure for each customer. In addition, by using cloud computing, high costs of IT infrastructures and operational costs of maintenance will be decreased [38]. These costs include energy consumption, IT systems maintenance, and support and transition management from old system to newer one [39]. Also, the cost related to servers including hardware and software purchase, annual licenses, technology updates, maintenance management costs and etc. can be saved using cloud computing technology.

Cloud computing efficiency includes an increase of IT infrastructure usage (more than sixty percent) [40], an increase of research and development in software innovation in the way of business and production growth [39], a creation of new ways which are not technically and economically possible without using cloud computing [39], prototyping and surveying on market acceptance for new approach quickly [39].

Cloud computing is agile and responsive for emergency needs because of its capacity to increase or decrease the ability to buy as a service from valid cloud providers [40].

Cloud computing availability can be assessed from some perspectives. From the first perspective, because cloud based software development is based on network performance; high levels of accessibility from this kind of software are excepted [14]. From the second perspective, high flexibility and accessibility of shared information cause an access to these services from anywhere through using internet.

Cloud computing is a motivation in the way of innovation and Creativity [14]. Using cloud computing, new technologies (such as cell phone and tablets) can be used to provide services for customers. For cloud computing implementation and use, vision of managers should change from ownerships of equipment view to service management view. Innovation can be concluded from this view. Also private sectors and organizations can benefit from cloud computing technology in innovation, organizational culture encouragement and relation with new technologies [40].

Cloud computing weaknesses

Security is the most important topic in cloud computing challenges. In cloud computing providers' selection, the care about provided level of security is critical. In this technology, users can have a basic description about their security and can determine security details. Cloud Providers should provide these security parameters. Users are depended for data access to Internet. Any lack of access to internet is a barrier to provide cloud services. So the existence of confident and stable communication platform to internet is critical. Different matters such as sanctions and storage location are important. Cloud's nature is in a way that storage locations can be different. Even data may be stored in different countries or continents. In this case, if service customer in a country is faced with sanctions, data accessibility will be ambiguous. In public platforms, because of neighboring data with other users, different security issues occur. Some security issues like side channels and covert channels are some example of this threat [39].

Privacy may be compromised because of the possibilities of access to stored critical and confidential information. Damage caused by security problems and privacy or inaccessibility to service can also damage reputation of both the customer and the organization [39]. Because of the related risks to cloud computing technology, consideration to priority of provided cloud services is too important.

Cloud computing models (especially SaaS model), for their provided software and type of service, decreased the possibilities of customization [39]. Especially when the organizational process is predefined and organization works based on specific process, implementation and compatibility of cloud services based on current organizational process imposes some complexities. When such compatibility with processes is essential, priority of Business Process Reengineering (BPR) is high. Before moving to this new technology, organizations must check on the effect of this new technology on their business processes and solve any problems or technical obstacles.

Providers must choose their level of service. This level of service should be guaranteed and be mentioned in Service Level Agreement (SLA). Service level should be monitored frequently (by both the providers and users) to ensure that this SLA is based on agreements. Every change in service level causes a disability of full and correct service delivery. So in cloud service provider's selection reputation, background, and service sustainability are critical factors. Because of daily increase in cloud computing service providers, coordination between them is too complex. For example, when a provider wants to finish its services, an organization must have the ability of information, application and processes transition to another compatible cloud provider [39].

Based on cloud pay-as-you-go model, customers will pay based on provided services for them. This model will change organizational IT budget. IT budget for traditional services is in form of capital costs, on the other hand, using pay-as-you-go model, IT budget should change to operational costs. This change may challenge the organization in budget estimation processes [39].

Information may be stored in different locations such as third country, this information is dependent on destination countries and organizations should be aware of destination countries laws to avoid trouble. Problems such as sanctions which may cause obstacle to accessibility of information should be considered by organizations [39]. From educational perspective, organizations should notice that transition to cloud computing technology needs more business analysis, change management and contract managers [39].

Cloud computing opportunities

Sharing IT infrastructures by cloud computing causes a decrease in IT capital costs especially in developing countries which have low capabilities of investment in high costs of IT infrastructures. These countries can use cloud computing to develop their IT services. Software provided by cloud technology creates a decrease in software and infrastructures costs which improve the software usage both for the organizations and users [29].

Software, as a service in electronic learning and electronic education, leads to improve in service quality in these two scopes. Using cloud computing, users can access the educational contents from educational centers or homes cooperatively and geographically which increases both users' and students' cooperation from different locations and educational levels [41].

Cloud computing will provide a secure platform for e-government and e-voting leading to people's cooperation in this subject. With the introduction of new services and provision of monotonic solutions, real time cooperation among agencies, location of free services, new communication and operational channels and elimination of cooperation barriers lead to provide high quality services in e-government and e-voting [35].

Small and medium size industries (SME's) have many problems in ERP and CRM systems implementation.

These problems include high financial power for implementation, managerial risks, business process reengineering (BPR) and etc. using cloud computing technology, SAP and Oracle companies; provide ERP systems under SaaS and IaaS platforms. SME's can use cloud ERP and cloud CRM in order to make benefit from ERP and CRM systems opportunities.

In today's world, tele-working is an important issue. Organizations move toward decreasing the employees' physical presence and increase the working time using tele-working. With data, software, infrastructure availability, cloud computing plays an important role in this way.

Another opportunity of cloud computing is green cloud. Using some mechanisms to save consumption power, cloud computing expands green IT. Some of these mechanisms are energy performance and systems scheduling [42]. Many researches are done in green cloud field in order to decrease energy consumption and increase effective services.

Cloud computing threats

Substitution by other providers is one of the threats which have impacts on cloud computing. Other providers can compete with a provider with higher quality and lower price. Because of the changes in IT culture in organizations, cloud computing may become threatened. These changes include changes in IT budget structure from capital to operational, changes in infrastructure managements and changes in the use of IT in organizations goals. With the elimination and transition of IT infrastructures to cloud computing platform, availability problems can endanger organizations safety. This unavailability may be occurred as a consequence of sanctions, end of provider services or etc. Organizations should be aware of the threat of losing the control on their physical infrastructures and services. Lack of suitable standards to implement cloud Computing shows itself as a threat in cloud computing implementations. In this topic, cloud computing implementation strategies are introduced by the U.S federal government and others [14,38,39,40]. Lack of suitable standards creates some problems in service transition from one provider to another returning from cloud services to traditional IT services.

In table 5 internet banking strength, weakness, opportunity and threat (SWOT) are summarized.

Table 5. Internet banking SWOT

Strength	Weakness	Opportunity	Threat
• Better customer service[43] • Wider customer base Availability [1] • Cost reduction [1] • Convenience [43] • Accuracy [43] • Transaction speed [43]	• Security and Privacy [7][11] • Comp ability • Accessibility • Performance • Maintenance • Legal regulation	• Market expand[43] • Building trust on brand name[43] • New product and service[43] • Customer purchase behavior[43] • Profitability [43] • User experience and User involvement [43]	• Loose of market [14] • Information leakage [14] • Continuity of service[44]

Internet banking strength

Because of twenty four hours services from anywhere and at any time, internet banking provides high level of availability for users. Users can easily go to internet web sites and benefit from internet banking services. This level of availability provides better services to customers and increase customers' convenience in using internet banking. Internet provides a wide platform for banking. Customers can access their requested service from any countries and there is no need for new branches in different countries for service expansion. Because of these, all users which have access to internet can have access to internet banking service from a specific bank. In addition, lack of physical presence of customers in physical branches will decrease paper work, official process, staff costs and etc. electronic structure of internet banking and service delivery methods such as payment, money transfer, purchase and etc. created more accuracy and transaction speed vs. traditional banking systems which increase service performance. Internet banking opportunities open new ways for banking agencies in providing innovative and new products and services which improve the service level, profit and efficiency [1,48].

Internet banking weakness

Security is one of the main and most important subjects in internet banking. Internet vulnerabilities such as the rejection of services, information leakage, losing customers' confidential information, virus attacks, credit card frauds, stealing account information and etc. are some of these vulnerabilities. In privacy, internet banking has many security challenges. Privacy assurance is one of the challenging issues in this section which endanger expansion of this service.

Customers should be able to transfer their money form an account to another account on internet banks or traditional banks. So these services should be compatible with other banking systems. Trust on internet banking services is hard due to lack of face to face interactions which cause lack of total confidence to internet banking from customers.

Traditional users need extra education and learning activities to learn how to work and trust with internet banking systems and websites. This may cause a decrease in users' attitude to internet banking. Sites and infrastructure maintenance for service provision should be considered. Users and internet banking systems providers should consider regulations and banking and money laws in destination and providers' countries in order to have no problems for both users and business.

Internet banking opportunities

Using internet as a service provision platform by internet banking leads to expand target market and improves bank customers' base. Considering many people's access to internet and provider of internet banking through internet and web sites, a trust on commercial brands is created. Internet banking platform is a service that can be accessed from anywhere and at any time and changes the social behavior of the people. Service innovations in internet banking will be increased and new and innovative software will be provided by internet banking industries.

All transactions and user visits from bank websites will be stored. Using this stored information, users' behaviors can be underhanded and forecasted easily. Using this understanding, new or changed product or services will be presented in order to improve service quality and convenience. Provided services by internet banking will cause new customer's experience and more customers' cooperation in internet banking services.

Internet banking threats

All of the threats that internet and cloud computing have can be occurred in internet banking service. Information leakage, sanctions, security problems will endanger service accessibility and users' trust which cause the loss of market or even end of an internet banking business.

5. Result and Discussion

Both internet banking and cloud computing are using internet platform to provide services for customers. Cloud computing shares software, platforms and infrastructures through the internet. Using web sites, Internet banking will provide different services for customers. So all SWOT of internet is shared between these two technologies.

In table 6 the strength, weakness, opportunity and threat of internet banking using cloud computing platform is summarized and compared together. In addition, we tried to answer these following questions:

1. Which opportunities are created by cloud computing strengths?
2. Which strengths will decrease the threats?
3. Which opportunities are lost by cloud computing weaknesses?
4. Which threat is created by cloud computing weaknesses?

Table 6. Internet banking using cloud computing technology SWOT

	Opportunities	Threats
Strengths	• More profit for both bank and customers. • Decrease in IT infrastructure costs and result in decrease of organization’s IT costs. • Innovation in services and new products. • Service access from anywhere and at any time. • Target market expand and more customer acquisition. • Increase customer satisfaction. • Staff education and training will be more effective and efficient. • Service expansion using mobile cloud. • Increase in small and medium size organization to implement internet banking. • Service management and control form anywhere and at any time. • Service integration using current standards. • Convenience in transactions using current standards.	• Prevention of market loose using innovative and cost effective services. • Security threats prevention using centralized services by provider. • Quick service recovery with low costs. • Better response to change in demand. • High capability for high requests responses.
Weaknesses	• Loose of market and investment because of sanctions. • Loose of customers because of security and access problems. • Loose of physical control on internet banking operations. • Lack of service integration with other internet banking services because of lack of standards.	• Lack of access to data because of sanctions. • Information leakage because of security threats. • Loose of reputation because current weaknesses. • Low service quality and substitution by rivalry.

As a consequence of a decrease in IT costs, cloud computing led to more profit for both banking system and customers. This profit is produced by a decrease in IT infrastructure costs, convenience in service, a decrease in paperwork costs and etc. Innovative specification of cloud computing and application of this technology in internet banking will release innovative and new services and products.

With the expansion of Cloud Computing, every day, new and innovative product and services are provided which can be used in internet banking section. Scalable, cost efficient and available services which are provided by cloud computing technology lead to expand the target market and attract more customers in wider scopes. Low costs services, wider service in geographical scope and availability from anywhere and at any time cause an increase in provided service quality and finally leads to increase in customers' satisfaction which expand target market. The application of cloud computing in internet banking improves mobile banking and staff education activities. On the other hand, some threats will endanger internet banking system based on cloud computing technology. Sanction is the most important threat which endangers these services. Sanctions cause services unavailability, loss of market and money. Information leakage is one of the most important threats to internet banking. This threat may cause a complete destruction of bank. These information leakages can be caused by some vulnerability such as internet threats, information accessibility to other providers which cause the loss of customers, target market, reputation and whole business.

Because of the provided infrastructure and software by cloud service providers; there is a threat of losing physical controls on internet banking operations from service providers. So organizations should consider providers' reputation, background, commitments and current customers. If provided services have low quality compared with other competitors and customers are not encountered by fast, convenient and accurate services, threat of substitution by competitors is considerable. Due to lack of standards for cloud computing and necessity of internet banking services integration with other internet banking services for transaction and other banking activities, data transaction between different providers may face some problems. Many agencies such as ISO are working on cloud computing standards.

Because of the high priority of security issues for internet banking and investment power in cloud computing technology by these agencies, medium and small size internet banking services can be chosen to migrate on cloud computing technology and take advantage of its unique specifications. In Fig. 2, a method to help internet banking in the way of cloud computing adoption is presented. Fig. 2 shows that whatever banking agencies want to invest more on internet banking based on cloud computing technology and have high security considerations, they can benefit from internet banking on private cloud IaaS.

On the other hand, if they want to invest less on this technology and have lower security considerations, they can choose internet banking on public cloud IaaS.

Cloud Computing provides more profit than traditional internet banking, so each of these two technologies will improve internet bank’s efficiency, profit, performance and etc.

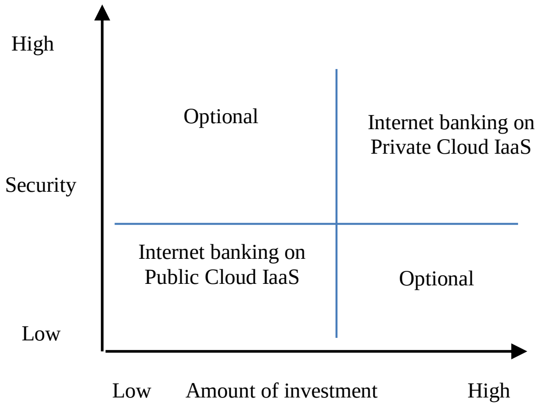

High
Security
Low
Optional
Internet banking on Private Cloud IaaS
Internet banking on Public Cloud IaaS
Optional
Low
Amount of investment
High

Fig. 2. Cloud computing adoption by internet banking

The Fuzzy Delphi Method is a method based on the Delphi Method and the Fuzzy Theory. Using this model, extracted factors can be evaluated based on expert opinion. Due to the accurate collection and analysis of expert opinions, this model has been used. In order to use fuzzy Delphi method [45], Experts offer their opinion in the form of minimum value and most likely value, then average of expert opinion (number provided) and the degree of disagreement of any expert from the average is calculated.

Fuzzy Delphi steps are as follows:

1. Experts selection
2. Preparation of the questionnaire
3. Getting expert opinion and analysis
4. If there is good consensus fuzzy Delphi process is finished otherwise back to step two

In the first step, ten experts in the field of information technology and banking are selected then a questionnaire based on a literature research in four parts is prepared. These four parts include which Internet banking opportunities can be pursued utilizing the strengths of the cloud computing? (SO), which threats to the Internet banking can be reduced or removed through cloud computing strengths (ST), which opportunities for Internet banking are lost by cloud computing weaknesses (WO) and to which threats Internet banking are exposed due to cloud computing weaknesses (WT).

The result of questionnaire is summarized in table 7.

After experts opinion collection, their opinion is transformed to qualitative variables as a trapezoid-shaped fuzzy numbers low (0,0,2,4), medium (3,4,6,7) High (6,8,10,10) using equation (1). Then the average (mean) A_m of all $A^{(i)}$ is computed using equation (2) [45].

$$A^{(i)} = \left(a_1^i, a_2^i, a_3^i, a_4^i\right), \quad i = 1,2,3,\ldots,n \qquad (1)$$

$$A_m = \left(a_{m1}, a_{m2}, a_{m3}, a_{m4}\right) = \left(\tfrac{1}{n}\sum a_1^{(i)}, \tfrac{1}{n}\sum a_2^{(i)}, \tfrac{1}{n}\sum a_3^{(i)}, \tfrac{1}{n}\sum a_4^{(i)}\right) \qquad (2)$$

where : *n* denotes the number of experts or opinions, m denotes the mean or average

The result of these equations is summarized in table 8.

Using equation (3) [45], for each expert, the difference between expert’s opinion and A_m is calculated and sent back to the expert for reexamination. Then each expert sends their revised opinion as a trapezoidal fuzzy number based on this difference. In table 9, the revised expert opinion and the average of expert opinions obtained from the questionnaires are shown.

$$e = \left(a_{m1} - a_1^{(i)}, a_{m2} - a_2^{(i)}, a_{m3} - a_3^{(i)}, a_{m4} - a_4^{(i)}\right) = \left(\tfrac{1}{n}\sum a_1^{(i)} - a_1^i, \tfrac{1}{n}\sum a_2^{(i)} - a_2^i, \tfrac{1}{n}\sum a_3^{(i)} - a_3^i, \tfrac{1}{n}\sum a_4^{(i)} - a_4^i\right) \qquad (3)$$

At next step, using equation (4) [45] the distance of two average fuzzy numbers from step one and two is calculated. If calculated difference be less than 2.0, the Delphi process achieved consensus and will be stopped [45]. The process could be repeated again and again until a consensus emerges.

$$S\left(A_{m2}, A_{m1}\right) = \left|\frac{1}{4}\left[\begin{array}{l}\left(a_{m21} + a_{m22} + a_{m23} + a_{m24}\right) - \\ \left(a_{m11} + a_{m12} + a_{m13} + a_{m14}\right)\end{array}\right]\right| \qquad (4)$$

According to smaller mean difference of 2.0, there is expert consensus on these factors after four rounds (the Results of 3rd and 4th questionnaires was the same). The result of this step is shown in tables 10-12.

Table 7. Result of first questionnaire

	SWOT Statements	**The opinion among 10 experts**		
		The degree of agreement		
		High	Medium	Low
SO	More profit for both bank and customers.	8	2	0
	Decrease in IT infrastructure costs and result in decrease of organization's IT costs.	9	1	0
	Innovation in services and new products.	7	2	1
	Service access from anywhere and at any time.	10	0	0
	Target market expand and more customer acquisition.	7	3	0
	Increase customer satisfaction.	6	2	2
	Staff education and training will be more effective and efficient.	5	4	1
	Service expansion using mobile cloud.	7	3	0
	Increase in the implementation of Internet banking by small and medium size organizations.	8	2	0
	Service management and control form anywhere and at any time.	8	1	1
	Service integration using current standards.	6	2	2
	Convenience in transactions using current standards.	9	0	1
ST	Prevention of market loose using innovative and cost effective services.	5	3	2
	Security threats prevention using centralized services by provider.	5	4	1
	Quick service recovery with low costs.	8	2	0
	Better response to change in demand.	7	2	1
	High capability for high requests responses.	7	3	0
WO	Loose of market and investment because of sanctions.	9	0	1
	Loose of customers because of security and access problems.	4	5	1
	Loose of physical control on internet banking operations.	3	4	3
	Lack of service integration with other internet banking services because of lack of standards.	5	5	0
WT	Lack of access to data because of sanctions.	4	3	3
	Information leakage because of security threats.	7	2	1
	Loose of reputation because current weaknesses.	3	5	2
	Low service quality and substitution by rivalry.	6	1	3
S: Strengths, W:weaknesses, O: opportunities, T: threats				

Table 8. Average of opinions obtained from the first

	SWOT Statements	**Average of experts' opinions (trapezoidal fuzzy number)**
SO	More profit for both bank and customers.	[5.4,7.2,9.2,9.4]
	Decrease in IT infrastructure costs and result in decrease of organization's IT costs.	[5.7,7.6,9.6,9.7]
	Innovation in services and new products.	[4.8,6.4,8.4,8.8]
	Service access from anywhere and at any time.	[6,8,10,10]
	Target market expand and more customer acquisition.	[5.1,6.8,8.8,9.1]
	Increase customer satisfaction.	[4.2,5.6,7.6,8.2]
	Staff education and training will be more effective and efficient.	[4.2,5.6,7.6,8.2]
	Service expansion using mobile cloud.	[5.1,6.8,8.8,9.1]
	Increase in the implementation of Internet banking by small and medium size organizations.	[5.4,7.2,9.2,9.4]
	Service management and control form anywhere and at any time.	[5.1,6.8,8.8,9.1]
	Service integration using current standards.	[4.2,5.6,7.6,8.2]
	Convenience in transactions using current standards.	[5.4,7.2,9.2,9.4]
ST	Prevention of market loose using innovative and cost effective services.	[3.9,5.2,7.2,7.9]
	Security threats prevention using centralized services by provider.	[4.2,5.6,7.6,8.2]
	Quick service recovery with low costs.	[5.4,7.2,9.2,9.4]
	Better response to change in demand.	[4.8,6.4,8.4,8.8]
	High capability for high requests responses.	[5.1,6.8,8.8,9.1]
WO	Loose of market and investment because of sanctions.	[5.4,7.2,9.2,9.4]
	Loose of customers because of security and access problems.	[3.9,5.2,7.2,7.9]
	Loose of physical control on internet banking operations.	[3,4,6,7]
	Lack of service integration with other internet banking services because of lack of standards.	[4.5,6,8,8.5]
WT	Lack of access to data because of sanctions.	[3.3,4.4,6.4,7.3]
	Information leakage because of security threats.	[4.8,6.4,8.4,8.8]
	Loose of reputation because current weaknesses.	[3.3,4.4,6.4,7.3]
	Low service quality and substitution by rivalry.	[3.9,5.2,7.2,7.9]
S: Strengths, W:weaknesses, O: opportunities, T: threats		

Table 9. Result of the second questionnaire

	SWOT Statements	**The opinion among 10 experts**	
		The degree of agreement	
		High, Medium, Low	**Average of experts' opinions (trapezoidal fuzzy number)**
SO	More profit for both bank and customers.	8,2,0	[5.4,7.2,9.2,9.4]
	Decrease in IT infrastructure costs and result in decrease of organization's IT costs.	9,1,0	[5.7,7.6,9.6,9.7]
	Innovation in services and new products.	8,1,1	[5.1,6.8,8.8,9.1]
	Service access from anywhere and at any time.	10,0,0	[6,8,10,10]
	Target market expand and more customer acquisition.	8,2,0	[5.4,7.2,9.2,9.4]
	Increase customer satisfaction.	7,2,1	[4.8,6.4,8.4,8.8]
	Staff education and training will be more effective and efficient.	6,4,0	[4.8,6.4,8.4,8.8]
	Service expansion using mobile cloud.	8,2,0	[5.4,7.2,9.2,9.4]
	Increase in the implementation of Internet banking by small and medium size organizations.	9,1,0	[5.7,7.6,9.6,9.7]
	Service management and control form anywhere and at any time.	9,0,1	[5.4,7.2,9.2,9.4]
	Service integration using current standards.	7,2,1	[4.8,6.4,8.4,8.8]
	Convenience in transactions using current standards.	10,0,0	[6,8,10,10]
ST	Prevention of market loose using innovative and cost effective services.	6,3,1	[4.5,6,8,8.5]
	Security threats prevention using centralized services by provider.	6,4,0	[4.8,6.4,8.4,8.8]
	Quick service recovery with low costs.	9,1,0	[5.7,7.6,9.6,9.7]
	Better response to change in demand.	8,1,1	[5.1,6.8,8.8,9.1]
	High capability for high requests responses.	9,1,0	[5.7,7.6,9.6,9.7]
WO	Loose of market and investment because of sanctions.	9,0,1	[5.4,7.2,9.2,9.4]
	Loose of customers because of security and access problems.	3,7,0	[3.9,5.2,7.2,7.9]
	Loose of physical control on internet banking operations.	2,6,6	[3,4,6,7]
	Lack of service integration with other internet banking services because of lack of standards.	6,4,0	[4.8,6.4,8.4,8.8]
WT	Lack of access to data because of sanctions.	4,4,2	[3.6,4.8,6.8,7.6]
	Information leakage because of security threats.	7,2,1	[4.8,6.4,8.4,8.8]
	Loose of reputation because current weaknesses.	2,7,1	[3.3,4.4,6.4,7.3]
	Low service quality and substitution by rivalry.	7,0,3	[4.2,5.6,7.6,8.2]
S: Strengths, W:weaknesses, O: opportunities, T: threats			

Table 10. Mean difference of the first and second questionnaires

	SWOT Statements	**Difference of experts opinions**
SO	More profit for both bank and customers.	0
	Decrease in IT infrastructure costs and result in decrease of organization's IT costs.	0
	Innovation in services and new products.	0.35 > 0.2
	Service access from anywhere and at any time.	0
	Target market expand and more customer acquisition.	0.35 > 0.2
	Increase customer satisfaction.	0.7 > 0.2
	Staff education and training will be more effective and efficient.	0.7 > 0.2
	Service expansion using mobile cloud.	0.35 > 0.2
	Increase in the implementation of Internet banking by small and medium size organizations.	0.35 > 0.2
	Service management and control form anywhere and at any time.	0.35 > 0.2
	Service integration using current standards.	0.7 > 0.2
	Convenience in transactions using current standards.	0.7 > 0.2
ST	Prevention of market loose using innovative and cost effective services.	0.7 > 0.2
	Security threats prevention using centralized services by provider.	0.7 > 0.2
	Quick service recovery with low costs.	0.35 > 0.2
	Better response to change in demand.	0.35 > 0.2
	High capability for high requests responses.	0.7 > 0.2
WO	Loose of market and investment because of sanctions.	0
	Loose of customers because of security and access problems.	0
	Loose of physical control on internet banking operations.	0
	Lack of service integration with other internet banking services because of lack of standards.	0.35 > 0.2
WT	Lack of access to data because of sanctions.	0.35 > 0.2
	Information leakage because of security threats.	0
	Loose of reputation because current weaknesses.	0
	Low service quality and substitution by rivalry.	0.35 > 0.2
S: Strengths, W:weaknesses, O: opportunities, T: threats		

Table 11. Result of the third questionnaire

	SWOT Statements	**The opinion among 10 experts**		**Difference of experts opinions (2nd and 3rd round)**
		High, Medium, Low	**Average of experts' opinions**	
SO	More profit for both bank and customers.	8,2,0	[5.4,7.2,9.2,9.4]	0
	Decrease in IT infrastructure costs and result in decrease of organization's IT costs.	9,1,0	[5.7,7.6,9.6,9.7]	0
	Innovation in services and new products.	8,1,1	[5.1,6.8,8.8,9.1]	0
	Service access from anywhere and at any time.	10,0,0	[6,8,10,10]	0
	Target market expand and more customer acquisition.	8,2,0	[5.4,7.2,9.2,9.4]	0
	Increase customer satisfaction.	7,2,1	[4.8,6.4,8.4,8.8]	0
	Staff education and training will be more effective and efficient.	7,3,0	[5.1,6.8,8.8,9.1]	0.35 > 0.2
	Service expansion using mobile cloud.	8,2,0	[5.4,7.2,9.2,9.4]	0
	Increase in the implementation of Internet banking by small and medium size organizations.	9,1,0	[5.7,7.6,9.6,9.7]	0
	Service management and control form anywhere and at any time.	9,0,1	[5.4,7.2,9.2,9.4]	0
	Service integration using current standards.	8,1,1	[5.1,6.8,8.8,9.1]	0.35 > 0.2
	Convenience in transactions using current standards.	10,0,0	[6,8,10,10]	0
ST	Prevention of market loose using innovative and cost effective services.	7,2,1	[4.8,6.4,8.4,8.8]	0.35 > 0.2
	Security threats prevention using centralized services by provider.	8,1,1	[5.1,6.8,8.8,9.1]	0.35 > 0.2
	Quick service recovery with low costs.	9,1,0	[5.7,7.6,9.6,9.7]	0
	Better response to change in demand.	8,1,1	[5.1,6.8,8.8,9.1]	0
	High capability for high requests responses.	9,1,0	[5.7,7.6,9.6,9.7]	0
WO	Loose of market and investment because of sanctions.	9,0,1	[5.4,7.2,9.2,9.4]	0
	Loose of customers because of security and access problems.	2,8,0	[3.6,4.8,6.8,7.6]	0.35 > 0.2
	Loose of physical control on internet banking operations.	1,8,1	[3,4,6,7]	0
	Lack of service integration with other internet banking services because of lack of standards.	7,3,0	[5.1,6.8,8.8,9.1]	0.35 > 0.2
WT	Lack of access to data because of sanctions.	3,6,1	[3.6,4.8,6.8,7.6]	0
	Information leakage because of security threats.	8,1,1	[5.1,6.8,8.8,9.1]	0.35 > 0.2
	Loose of reputation because current weaknesses.	1,9,0	[3.3,4.4,6.4,7.3]	0
	Low service quality and substitution by rivalry.	7,0,3	[4.2,5.6,7.6,8.2]	0
S: Strengths, W:weaknesses, O: opportunities, T: threats				

Table 12. Result of the fourth questionnaire

	SWOT Statements	**The opinion among 10 experts**	**Average of experts' opinions (trapezoidal fuzzy number)**	**Difference of experts opinions (3nd and 4th round)**
		High, Medium, Low		
SO	More profit for both bank and customers.	8,2,0	[5.4,7.2,9.2,9.4]	0
	Decrease in IT infrastructure costs and result in decrease of organization's IT costs.	9,1,0	[5.7,7.6,9.6,9.7]	0
	Innovation in services and new products.	8,1,1	[5.1,6.8,8.8,9.1]	0
	Service access from anywhere and at any time.	10,0,0	[6,8,10,10]	0
	Target market expand and more customer acquisition.	8,2,0	[5.4,7.2,9.2,9.4]	0
	Increase customer satisfaction.	7,2,1	[4.8,6.4,8.4,8.8]	0
	Staff education and training will be more effective and efficient.	7,3,0	[5.1,6.8,8.8,9.1]	0
	Service expansion using mobile cloud.	8,2,0	[5.4,7.2,9.2,9.4]	0
	Increase in the implementation of Internet banking by small and medium size organizations.	9,1,0	[5.7,7.6,9.6,9.7]	0
	Service management and control form anywhere and at any time.	9,0,1	[5.4,7.2,9.2,9.4]	0
	Service integration using current standards.	8,1,1	[5.1,6.8,8.8,9.1]	0
	Convenience in transactions using current standards.	10,0,0	[6,8,10,10]	0
ST	Prevention of market loose using innovative and cost effective services.	7,2,1	[4.8,6.4,8.4,8.8]	0
	Security threats prevention using centralized services by provider.	8,1,1	[5.1,6.8,8.8,9.1]	0
	Quick service recovery with low costs.	9,1,0	[5.7,7.6,9.6,9.7]	0
	Better response to change in demand.	8,1,1	[5.1,6.8,8.8,9.1]	0
	High capability for high requests responses.	9,1,0	[5.7,7.6,9.6,9.7]	0
WO	Loose of market and investment because of sanctions.	9,0,1	[5.4,7.2,9.2,9.4]	0
	Loose of customers because of security and access problems.	2,8,0	[3.6,4.8,6.8,7.6]	0
	Loose of physical control on internet banking operations.	1,8,1	[3,4,6,7]	0
	Lack of service integration with other internet banking services because of lack of standards.	7,3,0	[5.1,6.8,8.8,9.1]	0
WT	Lack of access to data because of sanctions.	3,6,1	[3.6,4.8,6.8,7.6]	0
	Information leakage because of security threats.	8,1,1	[5.1,6.8,8.8,9.1]	0
	Loose of reputation because current weaknesses.	1,9,0	[3.3,4.4,6.4,7.3]	0
	Low service quality and substitution by rivalry.	7,0,3	[4.2,5.6,7.6,8.2]	0
S: Strengths, W:weaknesses, O: opportunities, T: threats				

6. Conclusions

Cloud computing is a new technology in IT service provision. Using cloud computing as a platform in internet banking service provision creates opportunities and threats. Because of the cost efficiency, availability, scalability, accuracy, innovation and efficiency features of cloud computing, it is proposed for internet banking section. Using this technology leads to more profit for both customers and banking, new products and services, market expansion and etc. nevertheless, some threats such as sanctions, information leakage, change in IT culture and substitution affect this technology.

Banks can use cloud computing as a base for improving their services. Internet banking sectors have low level of security considerations and amount of investment on internet banking on public cloud IaaS. Other internet banking sectors which have high level of security needs and high amount of investment power can choose internet banking on private cloud IaaS. Others can choose between these two kinds of implementation.

This paper has two main limitations. First limitation is the access to experts who have the knowledge and the experience to support researchers. Only limited access to experts in the research areas of cloud computing and internet banking was available in this study. Second limitation of this paper is the restricted number of literature reviews in terms of cloud computing and internet banking which can be expanded using more research on different aspect of these technologies.

In future, we will try to assess internet banking states, opportunities and threats in Iran and opportunities and threat of cloud computing for Iranians' internet banking systems. Finally, we will try to provide some recommendation for Iranians internet banking in order to adopt cloud computing technology.

References

[1] P. Hanafizadeh, B. W. Keating and H. R. Khedmatgozar, "A systematic review of internet banking adoption", Telematics and informatics, 2013.

[2] M. Markis, V. Koumaras, H. Koumaras, A. Konstantopoulou, S. Konidis and S. Kostakis, "Qualifying factors influencing the adoption of internet banking service in Greece", International journal of e-adoption, Vol. 1, 2009, pp. 20-32.

[3] M. Tan and T. S. H. Teo, "Factors influencing the adoption of Internet banking", Journal of associate information systems, Vol. 5, 2000, pp. 1–42.

[4] K. Rouibah, T. Ramayah and O. S. May, "User acceptance of internet banking in Malaysia: test of three competing models", International journal of e-adoption, Vol. 1, 2009, pp. 1–19.

[5] M. A. Sarokolai, A. Rahimipoor, S. Nadimi and M. Taheri, "The investigating of barriers of development of e-banking in Iran", Social and behavioral science, 2012, pp. 110-1106.

[6] P. Tero, P. Kari, K. Heikki and P. Seppo, "Consumer acceptance of online banking: An extension of the technology acceptance model", Internet research, Vol. 14, 2004, pp. 224-235.

[7] P. Subsorn and S. Limwiriyakul, "A comparative analysis of internet banking security in Thailand: A customer perspective", Procedia engineering, Vol. 32, 2012, 260-272.

[8] Karim Z, Rezaul K. M. and Hossain A., "Towards secure information systems in online banking", in International conference for internet technology and secured transactions (ICITST), London, 2009.

[9] P. Subsorn and S. Limwiriyakul, "A comparative analysis of the security of internet banking in Australia: A customer perspective", in 2nd international cyber resilience conference (ICR), Perth, Western Australia, 2011.

[10] C. Gurau, "Online banking in transition economies: The implementation and development of online banking systems", International journal of bank marketing, Vol. 20, No. 6, 2012, pp. 285-296.

[11] J. H. Lee, W. G. Lim and J. I. Lim, "A study of the security of internet banking and financial private information in South Korea", Journal of mathematical and computer modeling, Vol. 58, No.1, 2011, pp. 117-131.

[12] NIST, "Cloud computing definition", U.S. Department of Commerce, 2011.

[13] N. A. Sultan, "Reaching for the cloud: how SME’s can manage", International Journal of Information Management Vol. 31, No. 3, 2011, pp. 272-278.

[14] V. Kundra, "Federal cloud computing strategy", The White House, Washington, 2011.

[15] "Unleashing the potential of cloud computing in Europe", European Commission, 2011.

[16] S. Ramgovind, M. M. Eloff and E. Smith, "The management of security in cloud computing", in Information Security for South Africa (ISSA) conference, Johannesburg, South Africa, 2010.

[17] M. Gregg, "10 security concern of cloud computing", Global knowledge training, 2010.

[18] M. Sathye, "Adoption of internet banking by Australian consumers: an empirical investigation", International journal of bank marketing, Vol. 17, No. 7, 1999, pp. 324–334.

[19] B. Howcraft, R. Hamilton and P. Hewer, "Consumer attitude and the usage and adoption of home-based banking in the United Kingdom", International journal of bank marketing Vol. 20, No. 2, 2002, pp. 111–121.

[20] Z. Liao and M. T. Cheung, "Internet-based e-banking and consumer attitudes: an empirical study", Journal of information management, Vol. 39, No. 4, 2002, pp. 283–295.

[21] S. Akinci, S. Aksoy and E. Atilgan, "Adoption of internet banking among sophisticated consumer segments in an advanced developing country", International journal of bank marketing Vol. 22, No. 3, 2004, pp. 212–232.

[22] P. Gerrard, J. B. Cunningham and J. F. Devlin, "Why consumers are not using internet banking: a qualitative study. Journal of service marketing", Vol. 20, No. 3, 2006, pp. 160–168.

[23] C. Sayar and S. Wolfe, "Internet banking market performance: Turkey versus the UK", International journal of bank marketing Vol. 25, No. 3, 2007, pp. 122–141.

[24] T. Laukkanen, S. Sinkkonen and P. Laukkanen, "Communication strategies to overcome functional and psychological resistance to Internet banking", International journal of information management Vol. 29, No. 2, 2009, pp.111–118.

[25] A. P. Mirza, A. Wallstorm, M. T. Hamidi Beheshti and O. P. Mirza, "Internet banking service adoption: private bank versus governmental bank", Journal of applied science Vol. 9, No. 24, 2009, pp. 4206-4214.

[26] M. M. M. A. Riffai, K. Grant and D. Edgar, "Big TAM in Oman: Exploring the promise of on-line banking, its adoption by customers and challenges of banking in Oman", International journal of information management, Vol. 32, No. 3, 2012, pp. 239-250.

[27] M. K. Normalini and T. Ramayah, "Biometrics technologies implementation in internet banking reduces security issues, Procedia- social and behavioral sciences, Vol. 65, 2012, pp. 364-369.

[28] S. Marston, Z. Li, S. bandyopadhyay, J. Zhang and A. Ghalsasi, "Cloud computing- the business perspective", Journal of decision support systems, Vol. 51, No. 1, 2011, pp. 176-189.

[29] D. Truong, "How cloud computing enhances competitive advantages, A research model for small business", the business review Cambridge, Vol. 15, 2010.

[30] S. Paquette, P. T. Jaeger and S. C. Wilson, "Identifying the security risks associated with governmental use of cloud computing”. Journal of government information quarterly, Vol. 27, No. 3, 2010, pp. 245-253.

[31] D. Zissis and D. Lekkas, "Addressing cloud computing security issues”. Journal of future generation computer systems, Vol. 28, No. 3, 2012, pp. 583-592.

[32] S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud computing", Journal of network and computer applications, Vol. 34, No. 1, 2011, pp. 1-11.

[33] X. Xu, "From cloud computing to cloud manufacturing", Journal of robotics and computer-integrated manufacturing Vol. 28, No. 1, 2012, pp. 75-86.

[34] D. Zissis and D. Lekkas, "Securing e-government and e-voting with an open cloud computing architecture", Journal of government information quarterly, Vol. 28, No. 2, 2011, pp. 239-251.

[35] A. Apostu, E. Rednic and F. Puican, "Modeling cloud architecture in banking systems", Procedia economics and finance, Vol. 3, 2012, pp. 543-548.

[36] R. Bose, X. Luo and Y. Liu, "The roles of security and trust: Comparing cloud computing and banking", Procedia-social and behavioral sciences, Vol. 73, 2013, pp. 30-34.

[37] "FAA Cloud Computing strategy", Federal Aviation Administration, 2012.

[38] "Cloud computing strategic direction paper, opportunities and applicability for use by the Australian government", Australian government department of finance and deregulation, 2011.

[39] "Cloud computing strategy", U.S department of defense chief information officer, 2012.

[40] H. M. Fardoun, S. R. Lopez, D. M. Alghazzawi and J. R. Castillo, "Education system in the cloud to improve student communication in the institutes of: C-learnXML++", Procedia – social and behavioral sciences, Vol. 47, 2012, pp. 1762-1759.

[41] C. Lin, "A novel green cloud computing framework for improving system efficiency", in Proceeding of International conference on applied physics and industrial engineering, Vol. 24, 2012, pp. 2326-2333.

[42] M. H. Shah and F. A. Siddiqui, "Organizational critical success factors in adoption of e-banking at the Woolwich bank", International journal of information management, Vol. 26, No. 6, 2006, pp. 442-456.

[43] J. W. Gikandi and C. Bloor, "Adoption and effectiveness of electronic banking in Kenya", Journal of electronic commerce research and applications, Vol. 9, No. 4, 2010, pp. 277-282.

[44] C. H. Cheng, Y. Lin, “Evaluating the Best Main Battle Tank Using Fuzzy Decision Theory with Linguistic Criteria Evaluation”, European Journal of Operational Research Vol. 142, 2002, pp. 174–186.

Monireh Hosseini holds a PhD from Tarbiat Modares University. She is currently an assistant professor at the Department of Industrial Engineering at K.N.Toosi University of technology. Her work deals with customer relationships management, internet marketing, e-commerce strategies and managements' information systems.

Elias Fathi Kiadehi holds a master degree in the field of Information Technology at K.N. Toosi University of technology. His work deals with cloud computing, information systems and data bases.

Low Complexity Median Filter Hardware for Image Impulsive Noise Reduction

Hossein Zamani HosseinAbadi*
Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran
h.zamanihosseinabadi@ec.iut.ac.ir
Shadrokh Samavi
Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran
samavi96@cc.iut.ac.ir
Nader Karimi
Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran
nader.karimi@cc.iut.ac.ir

Received: 11/Aug/2013 Accepted: 12/Apr/2014

Abstract

Median filters are commonly used for removal of the impulse noise from images. De-noising is a preliminary step in online processing of images, thus hardware implementation of median filters is of great interest. Hence, many methods, mostly based on sorting the pixels, have been developed to implement median filters. Utilizing vast amount of hardware resources and not being fast are the two main disadvantages of these methods. In this paper a method for filtering images is proposed to reduce the needed hardware elements. A modular pipelined median filter unit is first modeled and then the designed module is used in a parallel structure. Since the image is applied in rows and in a parallel manner, the amount of necessary hardware elements is reduced in comparison with other hardware implementation methods. Also, image filtering speed has increased. Implementation results show that the proposed method has advantageous speed and efficiency.

Keywords: Image Processing, Noise Reduction, Median Filter, Hardware Implementation, FPGA.

1. Introduction

Impulse noise or salt and pepper noise usually corrupts images during image capture or transmission. This noise may be found in situations where quick transients, such as faulty switching, take place during imaging. Impulse noise appears as black and white dots in some random pixels of an image. As a result, the effectiveness and accuracy of image's subsequent processes (such as edge detection, segmentation, feature extraction and pattern recognition) will be negatively affected [1]. Median filter is a nonlinear filter, which is common and effective tool for removing impulse noise from images. By sweeping a noisy image and outputting median value of the median window, median filter can significantly reduce the amount of impulse noise. De-noising is an important operation in image pre-processing. This process is usually performed online and inside camera's hardware. The de-noised image can then be further processed to obtain necessary information. Therefore, hardware implementation of median filter is of great importance for the purpose of online image de-noising and preparing the image for principal processing.

A variety of studies have been done on hardware implementation of median filters [2]-[6]. For implementing the median filter a method, called standard method, is the sorting of pixels and extracting the middle value of the sorted pixels as the filter's output [2]. Karaman et al. [3] propose a change to the standard method by dealing with samples in a bitwise manner, needing only single bit sorters. The strength of regular array architectures is that they can be pipelined down to a single compare-swap stage. This means high throughput and frequency. Instead of standard sorting of all pixels, multilayer sorting, as explained in [5], could also be used. Implementation of the median filter by means of trace transform is yet another important study. In that method, instead of sorting of the pixels, numbers of all pixel values are recorded. The pixel that sum of the recorded numbers of its previous pixels is half of the sum of all pixel value numbers, will be the output. This method is effective for large windows, but in small windows (such as 3×3) vast hardware resources are required and this method is not suitable [5]. Use of a threshold value for computing the output of the median filter is presented in [6]. Some other studies are about implementing other filters with equivalent performance with median filters for image de-noising applications [7]-[12]. There are also studies about implementation of median and equivalent filters at VLSI level for costume IC design in order to boost clock pulse frequency and decrease usage area of the chip [13,14].

In this paper, a pipelined structure is proposed for implementing median 3×3 filter. The structure reduces necessary registers for filter implementation up to 50%.

* Corresponding Auhor

Afterwards, based on this structure, a pipelined method is introduced for reading images in a row by row manner and de-noising them. By using parallel modules in the proposed method, speed of the filter is significantly increased in comparison with sweeping the image by a 3×3 window.

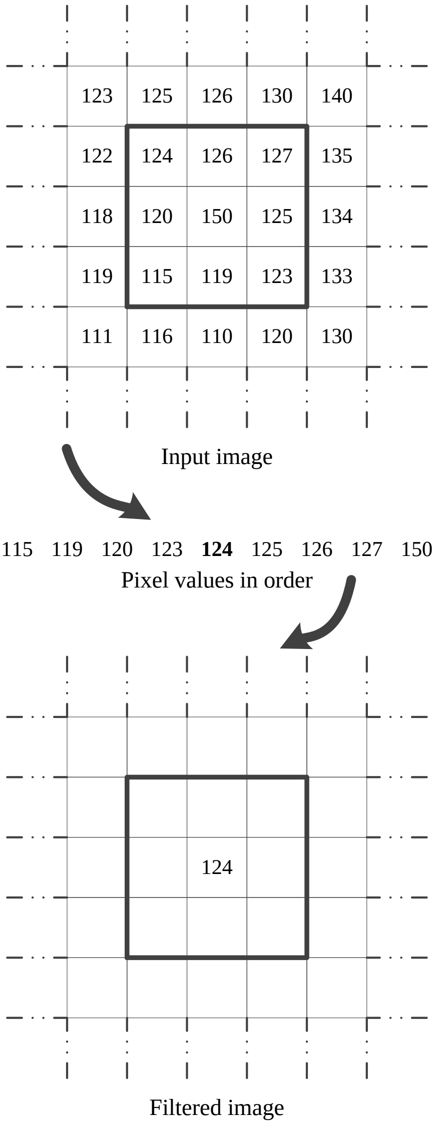
In Section 2, various median filter implementation methods including standard and multilayer implementations are described. Afterwards, in Section 3, the proposed method for implementing median filter, reading images and applying the filter on them, is introduced. In Section4, simulation results are presented. Eventually, concluding remarks are presented in Section 5.

2. Standard and Multilayer Implementations of Median Filter

Median filter is a spatial filtering operation, so it uses a 2-dimensional mask that is applied to each pixel in the input image. To apply the mask means to center it in a pixel, evaluating the covered pixel values and determining which brightness value is the median value. The median value is determined by placing the pixel values in ascending order and selecting the center value [1]. The obtained median value will be the value for that pixel in the output image. Fig. 1 shows how a median window is applied to a part of an input image.

Since bigger windows may eliminate small edges of the input image, usually a 3×3 median window is used for image filtering. Therefore, we will focus on hardware implementation of the 3×3 median filter.

Sorting windowed pixels for extracting the median value is done usually with two methods: standard and multilayer. In the standard method, all 9 pixels of a 3×3 median window are sorted together. For sorting the pixels, 8 bit comparator elements (or pixel comparators) are used [2]. The standard sorting algorithm structure for 9 pixels of the window is depicted in Fig. 2. In this figure, each of the displayed elements, are 8 bit comparators. As shown in the figure, for sorting 9 pixels, 9×(9-1)/2=36 comparators are needed. However, dark colored comparators are used for sorting all pixels and are not required for extracting the median value. Thus, 30 comparator elements will be adequate to implement the standard method [5]. Interior structure of comparator elements is shown in Fig. 3. These elements compare two pixels, and send higher value pixel into the H output and lower pixel into the L output.



123	125	126	130	140
122	124	126	127	135
118	120	150	125	134
119	115	119	123	133
111	116	110	120	130

Fig. 1. Filtering of an image by median filter.

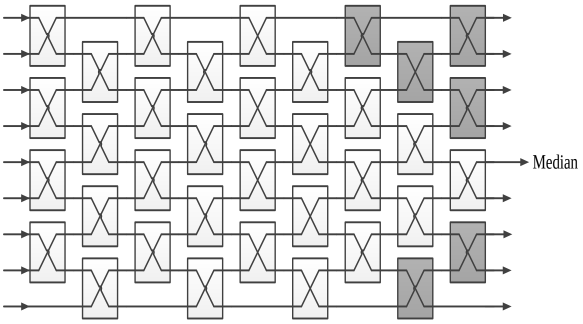


Fig. 2. Standard sorting algorithm [5].

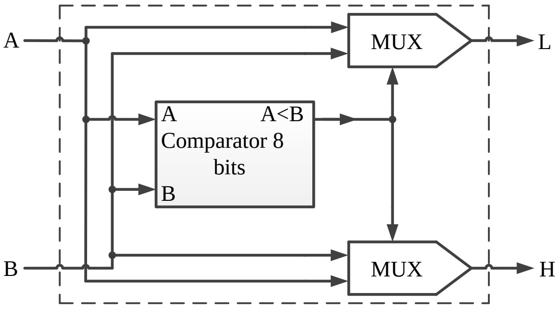


Fig. 3. Internal structure of comparators [2].

Some changes could be made in the standard sorting algorithm to reduce required hardware elements for implementing the median filter. For this purpose, instead of sorting all 9 pixels together, each column of the 3×3 median window is sorted independently in 3 sorting blocks. These sorting blocks sort the 3 input pixels in ascending order. Afterwards, outputs of the sorting blocks are grouped in 3 new sorting blocks and are sorted again in layer 2. By continuing this procedure, the median value could be extracted in the output of layer 3. This method is a multilayer implementation of the median filter. The multilayer sorting algorithm structure for 9 pixels of a 3×3 window is depicted in Fig. 4 [4]. Each of the sorting blocks in Fig. 4 is constructed from 3 comparators, as shown in Fig. 5.

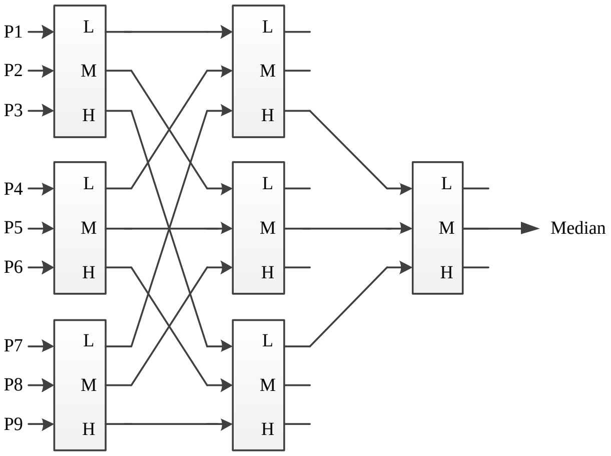


Fig. 4. Multilayer sorting of pixels.

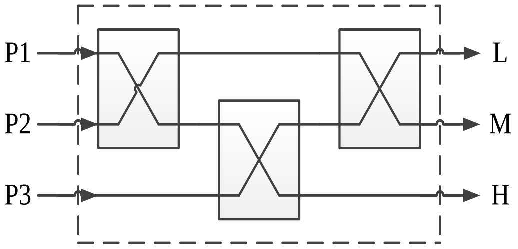


Fig. 5. Structure of a 3 pixels sorting block.

Multilayer implementation method has been proposed based on deleting non-median pixels. In fact, any pixel that is higher or lower than 5 of the other pixels in median window could not be the median value. By utilizing this logic, it could be proved that the output of the filter shown in Fig. 4 is the median pixel. This method uses only 7×3=21 comparator elements [15]. Pipeline structure for this method is introduced in [15]; this structure is displayed in Fig. 6. In the displayed structure, number of comparator elements is further reduced to only 15. In this structure, instead of sorting columns of the median window in different 3 pixels sorting blocks, by pipelining the hardware, one block is used to sort all columns in the layer one of filter.

3. Proposed Method

Median filter implementations that are introduced in previous section, must sweep the image to filter it; thus filtering of the whole image is slow and time consuming. To improve the filtering speed, at first, we propose a pipelined median multilayer structure, called 3-level pipelined filter. This structure is depicted in Fig. 7. In the proposed structure, there are 3 levels of filter and all levels are pipelined. This structure accepts 9 pixels as inputs and returns the median pixel as its output.

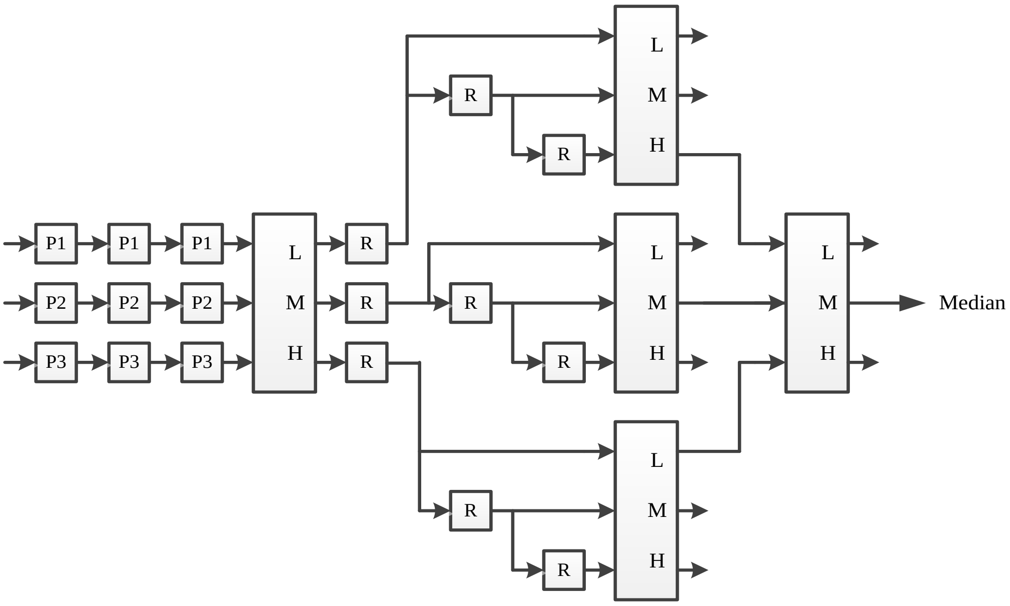


Fig. 6. Pipelined multilayer sorting structure [15]

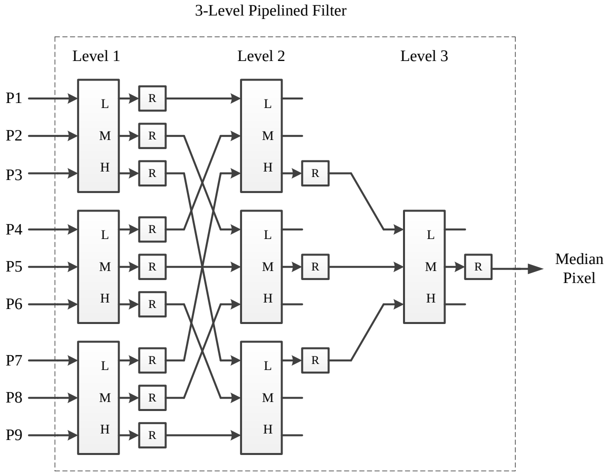


Fig. 7. Proposed pipeline architecture for multilayer median filter (3-level pipelined filter).

Although, in the 3-level pipelined filter, required hardware elements are more than the structure shown in Fig. 6; but, by introducing a useful method for reading images, these hardware elements could be decreased. By using some of the proposed 3-level pipelined filters in parallel form, hardware elements will be reduced. For the purpose of using parallel filters, image is transmitted to a pipelined hardware in a row by row structure. Thus, number of required filters is as same as number of pixels in a row of the image. Fig. 8 illustrates the manner of transmitting image pixels into the hardware. To increase clarity of the figure, only a few of the wires are illustrated.

Each median block in Fig. 8 has the proposed 3-level pipelined filter structure demonstrated in Fig. 7. According to Fig. 7, there is no need to feed all 9 pixels into a 3-level pipelined filter block of Fig. 8. In fact, by using parallel blocks, only 3 pixels could be enough to be fed into one block. To modify the 3-level pipelined filter structure, in level 1 of the structure, the upper and lower 3 pixel elements could be removed and instead of them, the 3 pixel elements from the adjacent 3-level pipelined filters could be used. Thus only 3 pixels are fed into each 3-level pipelined filter block of Fig. 8.

For reading image pixels and transmitting them into the hardware, primary pixels of each 3 rows in Fig. 8 will be transmitted into block 1; secondary pixels of each 3 rows in Fig. 8 will be transmitted into block 2, and so on, until the end of every 3 rows.

In level 2 of the 3-level pipelined filter, outputs of level 1 are used. In each 3-level pipelined filter block in Fig. 8, outputs of levels 1 of next and previous blocks can be used as well as outputs of level 1 of the block itself as the inputs of level 2.

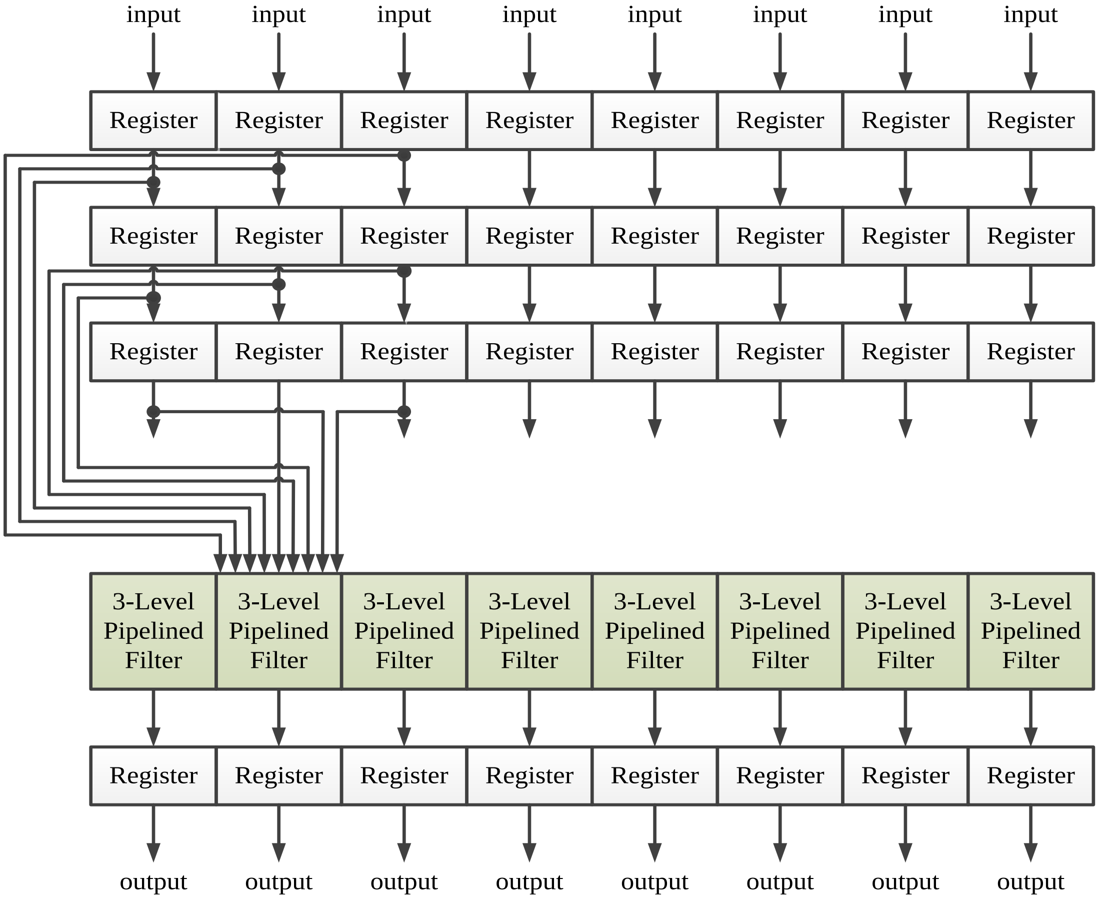


Fig. 8. Feeding of image into pipelined architecture.

According to the proposed image feeding technique, the architecture per each output pixel to de-noise a noisy image is shown in Fig. 9. As indicated in the figure, this structure accepts three pixels of a column of median window as inputs, and computes the output of median window. Since median window should be repeated in a whole row of a noisy image to de-noise that row, the structure demonstrated in Fig. 9 should be repeated for each pixel of the row. In this architecture, for each pixel, 5 comparators of the 3-inputs type are needed. Suppose that there are N pixels in each row of an image; for de-noising this image with our proposed method N demonstrated structures are required; therefore, 5×3×N=15N comparators are needed.

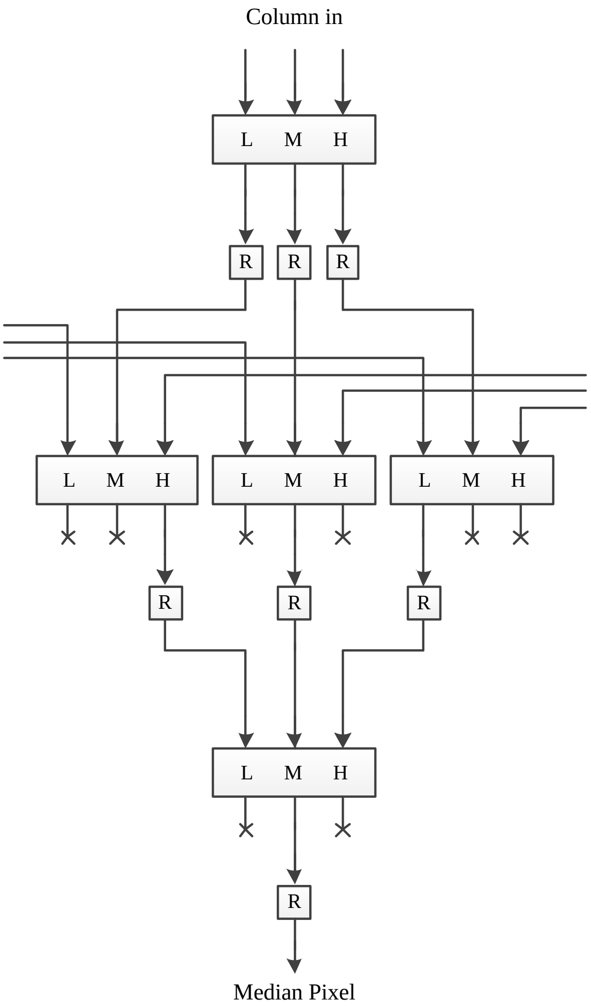


Fig. 9. Proposed pipelined multilayer median filter architecture for each pixel of image.

For de-noising the whole row of an image, and subsequently de-noising the image, the proposed pipeline multilayer median filter architecture, indicated in Fig. 9, should be repeated for all pixels of the row. Furthermore, primary registers those are shown in Fig. 8, should be included. Accordingly, by replacing each of the 3-level pipelined filter blocks and below registers in Fig. 8, by the proposed pipelined multilayer median filter shown in Fig. 9, our proposed structure for low complexity hardware image de-noising, can be constructed. Architecture of the proposed method for de-noising images is depicted in Fig. 10. This architecture will be repeated for the whole row of an image.

4. Simulation Results

The performance accuracy of the proposed architecture for image de-noising was proved by simulating it using Active-HDL 8.1 software. Several images with different row lengths used as the input noisy images. Impulsive noises by different noise factors were added to the images to corrupt them. Afterwards, the noisy images were de-noised by the proposed structure. Fig. 11 displays simulation signals including: clock pulse, input row, and output row for a de-noising process. The output signal activates after 6 clock pulses; hence, 6 clock pulses are required for the first row to be de-noised. Afterwards, subsequent rows need only 1 clock pulse to be de-noised. Therefore, the latency of the proposed structure is 6. This was expected since there are 6 registers in the path of each output pixel in the proposed structure as shown in Fig. 10. If the number of rows in an image is supposed to be M, the number of clock pulses those are needed to make the image de-noised, is equal to M+6. Moreover, as stated before in Section 3, if there are N pixels in each row of image; 5×3×N=15N pixel comparator elements are needed for de-noising this image with our proposed method.

Fig. 12 shows the results of image de-noising by means of the proposed median filter. In this figure, the results of image de-noising by the proposed structure are compared by the outputs of image de-noising by the 3×3 median filter in MATLAB software for three pictures. The pictures are corrupted by impulsive noise with different noise factors and the lengths of their rows are different. The length of image rows is 500 pixels for the left image and is 512 for the other ones. Accordingly, the proposed structure is established twice, both for 500 pixels and for 512 pixels.

In Fig. 12, original images, noisy images (10% of pixels are corrupted by salt and pepper noise for the two left images and 20% of pixels are corrupted for the two right ones), and de-noised images, both with the proposed median filter and with median filter in MATLAB, are displayed for both images. Also, the images which are de-noised by means of the 3×3 mean filter in MATLAB are displayed in the figure. As displayed in the figure, the proposed median filter could de-noise the noisy images with desired efficiency. Moreover, the priority of the proposed filter over the mean filter in image de-noising can be concluded.

In order to further illustrate the performance of the proposed filter, in Fig. 13 we have compared the produced histograms of different filters. Figure 13 (a) illustrates the original image and its histogram. The image has a relatively distributed histogram with all levels of gray-scale values. Figure 13(b) shows the noisy image due to 10 percent salt and pepper noise. The histogram of noisy image shows high number of induced zero points (peppers) and large number of white pixels (salts). The effect of mean filter is shown in Fig. 13 (c) where the quality of the produced image may be better than the noisy image but still is far from desirable. While the histogram of the original image has three distinct peaks, the histogram of the mean filter is very much flat and barely contains the three mentioned peaks. On the other hand, Fig. 13 (d) shows the output of our median filter where the image and its histogram are very close to the original image and its histogram.

The proposed architecture was further simulated in Xilinx ISE 11.1 software to determine the required hardware elements and maximum clock frequency. A Virtex5 family FPGA, XC5VTX240T, was used as the target device. Different implementation methods of the median filter were synthesized separately; and necessary hardware elements for implementing the filter using each method were computed based on the device's resources. The proposed median window architecture was compared with the standard median filter [2], multilayer median filter [4] and the pipelined multilayer median filter introduced in [15]. Required hardware resources and minimum filter's delay for all methods were computed and are shown in Table 1.

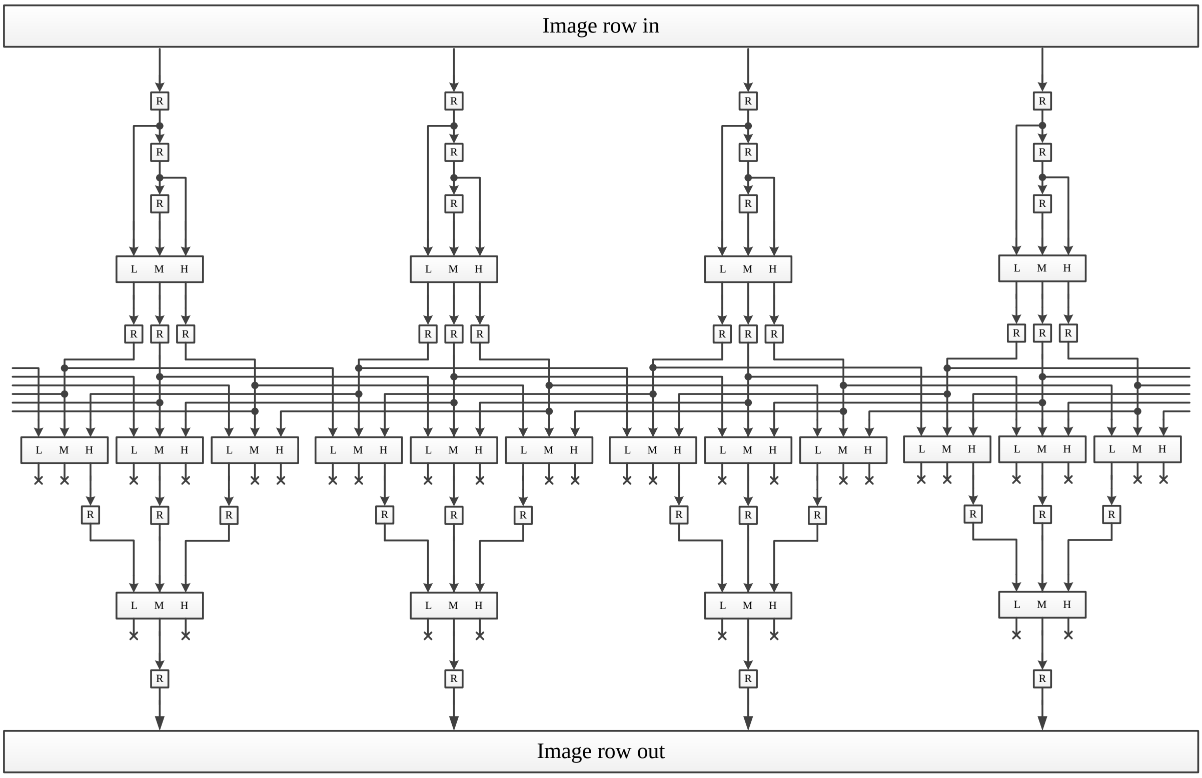


Fig. 10. Architecture of proposed method for image de-noising (structure for only 4 pixels is shown).

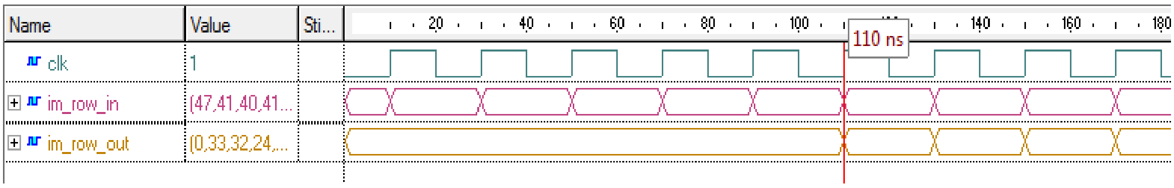
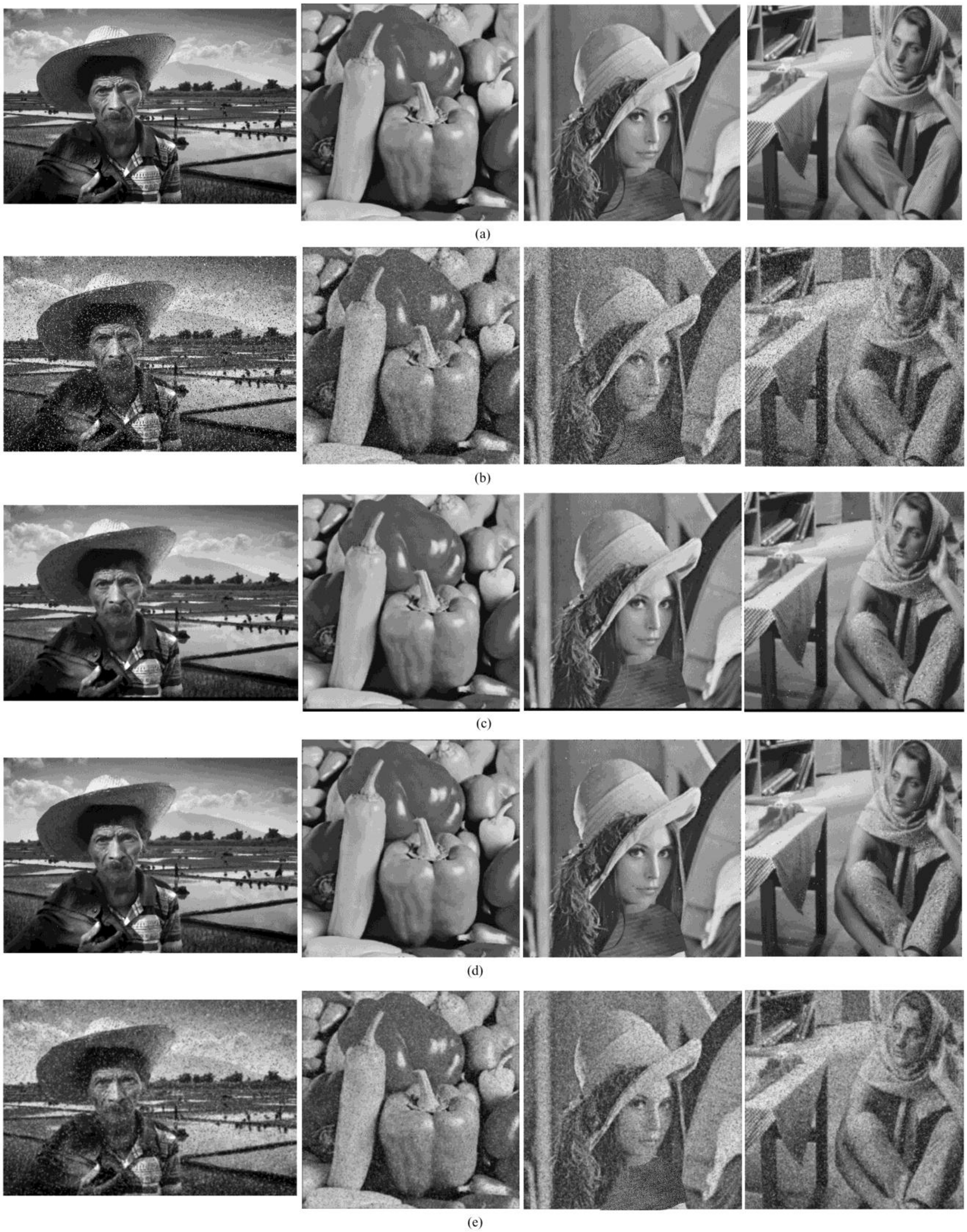


Fig. 11. Three principle simulation signal



(a)

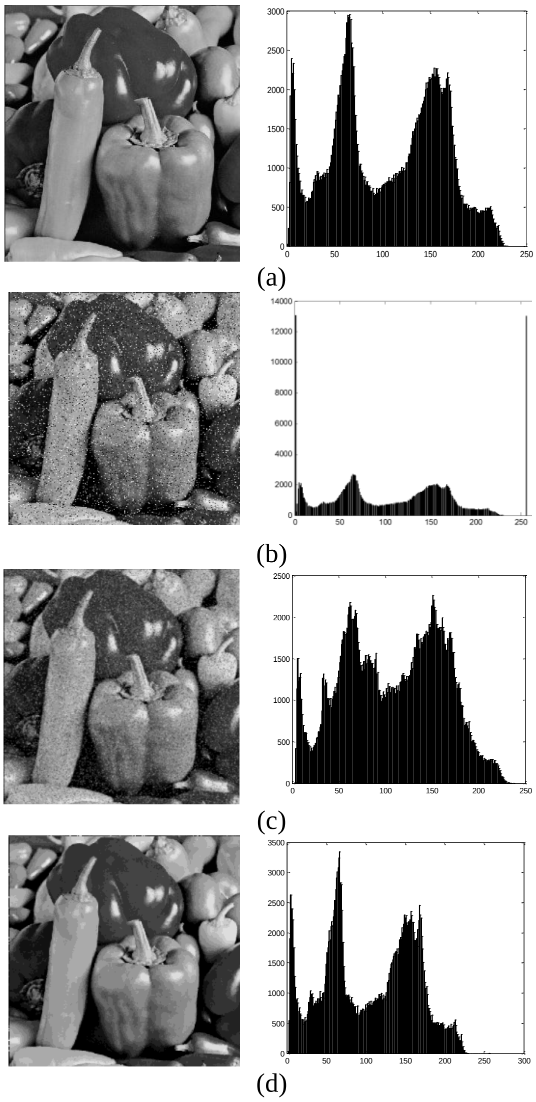
(b)

(c)

(d)

(e)

Fig. 12. (a) Original images; (b) noisy images by 10% and 20% salt and pepper noise for the left and right images respectively; (c) images denoised by the proposed method; (d) images denoised by median filter in MATLAB software; (e) image de-noising by mean filter in MATLAB software.



(a)

(b)

(c)

(d)

Fig. 13. Comparison between output of our filter and that of a 3x3 mean filter using histograms, (a) original image, (b) noisy image due to 10% salt and pepper noise, (c) output of mean filter, (d) output of our median filter

Table 1. Synthesis results for different median filter implementation methods.

Method	*# of LUTs*	*# of DFFs*	*Delay (ns)*
Standard [2]	552	0	34.614
Multilayer [4]	316	0	33.787
Pipelined multilayer [15]	224	768	2.006
Proposed method	224	384	1.947

According to the results demonstrated in Table 1, in comparison with pipelined multilayer median filter [15], the proposed architecture has up to 50% reduction in the required registers; also it has increased speed of the hardware up to 3%. In each of these methods (the proposed and the pipelined multilayer[15]), needed LUTs for implementing the filter is 29% lower than the multilayer median and 57% lower than the standard median filter. Also, using the proposed structure can lead to filter speed increase between 1.025 to 17.78 times in comparison with the other pipelined and not pipelined median filter implementation methods.

5. Conclusions

In this paper, a method for implementing the median filter was proposed to reduce required hardware elements and to increase the processing speed. In the proposed method, images are applied to the filter in rows and with a pipelined method. Filter elements are pipelined, too. Accuracy of the proposed architecture in removing salt and pepper noise was demonstrated by de-noising of sample noisy images. The synthesis results revealed that the proposed method could increase filter speed up to 3% and decrease the required hardware elements up to 50% in comparison with the existing pipelined multilayer median filter structure.

References

[1] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 2nd ed. New Jersey: Prentice Hall, 2008, pp. 225-227.

[2] D. Richards, "VLSI median filters", IEEE Transaction on Acoustics, Speech and Signal Processing, Vol. 38, No. 1, 1990, pp. 145-153.

[3] M. Karaman, L. Onural and A. Atalar, "Design and implementation of a general-purpose median filter unit in CMOS VLSI," IEEE Journal of Solid-State Circuits, Vol. 25, No. 2, 1990, pp. 505–13.

[4] J. L. Smith, "Implementing Median Filters in XC4000E FPGAs", XCell, Vol. 23, No. 4, 1996, p. 16. [Online]. Available: http://users.utcluj.ro/~baruch/resources/Image/xl23_16.pdf

[5] S. A. Fahmy, P. Y. K. Cheung and W. Luk, "Novel FPGA-based implementation of median and weighted median filters for image processing", in Proc. 2005 International Conference on Field Programmable Logic and Applications FPL, pp. 142-147.

[6] A. Burian and J. Takala, "VLSI-efficient implementation of full adder-based median filter ", in Proc. 2004 IEEE International Symposium on Circuit and Systems, Vol. 2, pp. 817-820.

[7] H. S. Yu, J. Y. Lee and J. D. Cho, "A fast VLSI implementation of sorting algorithm for standard median filters", in Proc. 1999 IEEE International ASIC\SOC Conf., pp. 387-390.

[8] C. T. Chen, L. G. Chen and J. H. Hsiao, "VLSI implementation of a selective median filter", IEEE Transaction of Consumer Electronics, Vol. 42, No. 1, 1996, pp. 33-42.

[9] L. Breveglieri and V. Piuri, "Digital median filters", Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology, Vol. 31, No. 3, 2002, pp. 191-206.

[10] K. S. Srinivasan and D. Ebenezer, "A New Fast and Efficient Decision- Based Algorithm for Removal of High-Density Impulse Noises", IEEE Signal Processing Letter, Vol.14, No.3, 2007, pp. 189-192.

[11] T. Matsubara, V.G. Moshnyaga and K. Hashimoto, "A FPGA implementation of low-complexity noise removal.", 17th IEEE International Conference on Electronics, Circuits, and Systems (ICECS), IEEE, 2010.

[12] D. Prokin and M. Prokin, "Low hardware complexity pipelined rank filter", IEEE Transactions on Circuits and Systems II: Express Briefs, Vol.57, No.6, 2010, pp. 446-450.

[13] C.Y. Lien, C.C. Huang, P.Y. Chen and Y.F. Lin, "An efficient denoising architecture for removal of impulse noise in images", IEEE Transactions on Computers, Vol.62, No. 4, 2013, pp. 631-643.

[14] P. Chen, C. Lien and H. Chuang, "A low-cost VLSI implementation for efficient removal of impulse noise", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol.18, No.3, 2010, pp. 473-481.

[15] K. Vasanth, S. Nirmal raj, S. Karthik and P. Preetha mol, "FPGA implementation of optimized sorting network algorithm for median filters", in Proc. 2010 International Conference on Emerging Trends in Robotics and Communication Technologies (INTERACT), pp. 224-229.

Hossein Zamani Hosseinabdi was born in 1988 in Isfahan, Iran. He received his B.Sc. and M.Sc. degrees in Electrical Engineering from the Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran in 2010 and 2013 respectively. His research interests include analog/mixed signal integrated circuits, implementable and wearable electronics, hardware implementation of signal processing algorithms and digital signal processing.

Shadrokh Samavi is a Professor of Computer Engineering at Isfahan University of Technology, Iran. He is also an Adjunct Professor at the ECE department of McMaster University where he is a member of the Multimedia Signal Processing Lab. Professor Samavi completed a B.S. degree in Industrial Technology and received a B.S. degree in Electrical Engineering at California State University, a M.S. degree in Computer Engineering at the University of Memphis and a Ph.D. degree in Electrical Engineering at Mississippi State University, U.S.A. Dr. Samavi is a Registered Professional Engineer (PE), USA. He is also a member of IEEE and a member of Eta Kappa Nu and Tau Beta Pi honor societies. Shadrokh Samavi's research interests are in the areas of image processing and hardware implementation and optimization of image processing algorithms. He is also interested in compression and processing of biomedical images, as well as, VLSI design and computer arithmetic.

Nader Karimi received the B.S. degree (summa cum laude) in computer engineering from Azad University, Arak Branch, Iran, in 2002 and the M.Sc. and Ph.D. degrees (honor) in computer engineering and electrical engineering from Isfahan University of Technology (IUT), Iran, in 2004 and 2012, respectively. He is currently an Assistant Professor at the Department of Electrical and Computer Engineering, Isfahan University of Technology. His research interests are image compression, hardware implementation and optimization of image processing algorithms, and watermarking.

A Stochastic Lyapunov Theorem with Application to Stability Analysis of Networked Control Systems

Babak Tavassoli*
Department of Electrical and Computer Engineering, K.N Toosi University of Technology., Tehran, Iran
tavassoli@kntu.ac.ir
Parviz Jabehdar-Maralani
Department of Electrical and Computer Engineering, University of Tehran, Tehran, Iran
pjabedar@ut.ac.ir

Received: 02/Jul/2013 Accepted: 12/Apr/2014

Abstract

The source of randomness in stochastic systems is an input with stochastic behavior as treated in the existing literature. Special types of stochastic processes such as the Wiener process or the Brownian motion have served as an adequate model of such an input for years. The body of stochastic systems theory is elegantly shaped around such input models. An example is the Itô's formula. With development of new applications, we are faced with various phenomena that are more demanding from a stochastic modeling approach.

To cope with this problem we restate the stochastic Lyapunov theorem such that it can be applied to a wider class of stochastic systems. In this paper stochastic systems are considered without imposing assumptions on the nature of the stochastic input and the way it affects the sample trajectories. Lyapunov stability theorem is represented for this type of systems in terms of a stability notion that generalizes the notion of stability in moments. As a result, the new theorem finds a larger domain of applications while it can be reduced to some known versions of the stochastic Lyapunov theorem. As an application, an existing deterministic result for nonlinear networked control systems is extended to a more practical probabilistic setting which extends the available analysis tools for checking the stability of continuous-time nonlinear networked control systems in the stochastic setting. The results are applied to a two-channel magnetic levitation system which is controlled over a local communication network to obtain a bound on the rate of transmission failures due to the presence of noise in the industrial environment.

Keywords: Networked Control Systems, Nonlinear Systems, Stochastic Systems, Lyapunov Theorem, Delayed Systems.

1. Introduction

In many applications, stability of a control system should be studied in presence of some random behavior. The theory of stochastic differential equations (SDEs) is used for this purpose [1,2,3,4]. An SDE can be regarded as a differential equation which depends on a stochastic process. The theory of SDEs is mainly developed for Itô SDE [4] and its applications to stochastic control problem are usually based on extensions of Lyapunov theorem [5,6,7]. There are problems that cannot be modeled using the Itô SDE like the random switched system in [8]. Another problem is the networked control system (NCS) analysis problem considered in this work. The difficulty is to model the stochastic phenomena as a Brownian motion to act as an input of the Itô SDE which is not always possible.

Several approaches have been used for handling NCS problems [9,10]. Two important issues are handling the stochastic effects such as communication delay and loss in the NCS [11]. LQG problem for linear NCSs with stochastic delays and packet losses is studies for example in [12,13]. An in-depth investigation of an NCS problem may lead to more detailed modeling and analysis, such as the relationship between stability and noise characteristics in [14], or network scheduling and topology related issues in [15,16]. Some basic works regarding nonlinear NCS are [17,18,19] where the network induced error is defined and modeled as a perturbation.

In this work, the Lyapunov stability method is presented in an abstract setting with respect to the Itô SDE stability analysis. In the Itô SDE, the usual source of randomness is an stochastic input. This input has certain properties that result in the Itô formula which is the basis of the related Lyapunov based stochastic stability analysis methods. However, our results are not based on the Itô formula which enables us to apply our results to an NCS problem (and possibly other new applications). For this purpose, a stability notion which is more suitable is used. Additional efforts may be required for applying the results to a problem. But, in return the results may be used for a wider class of applications. The main motivation of this work is its application in extending the NCS analysis performed in [17] in which the effect of shared communication on a nonlinear NCS is studied in a deterministic setting. The results of this paper are applied to obtain a practical probabilistic NCS analysis. Due to possibility of significant delays in an NCS, the Lyapunov results are presented for delayed systems to facilitate extension of the NCS analysis to delayed case in future. This paper is an enhanced version of [20] where the

* Corresponding Author

formulation of NCS analysis is improved and a new section is added to present an application of the results to a practical NCS problem.

In section two, the considered SDE is described. Also, the stability notions will be presented and their relations will be clarified. Section three contains the Lyapunov stability results. Extension of the NCS problem in [17] is studied in section four followed by an example. A practical case-study is studied in section five and conclusions are made at the end.

2. Preliminaries

Notation: Throughout the paper, the set of real numbers $(-\infty,\infty)$ is indicated by $\mathbb{R}$ and the set of non-negative real numbers $[0,\infty)$ is indicated by $\mathbb{R}^{-}+$ where ∞ is the positive infinity. Euclidian norm of a vector x is denoted by $\|x\|$. For an arbitrary set A, the set of mappings from A to $\mathbb{R}^n$ is denoted by $C^n(\mathsf{A})$ and the set of stochastic processes with sample paths in $C^n(\mathsf{A})$ is denoted by $\mathcal{B}^n(\mathsf{A})$. For $\varphi \in C^n(\mathsf{A})$, ∞-norm is defined as $\|\varphi\|\infty = \sup_{\alpha\in\mathsf{A}} \|\varphi(\alpha)\|$. For $\varphi \in C^n(\mathbb{R})$ and $t\in\mathbb{R}$, history of φ at t denoted by $\varphi_t\in C^n(\bar{\mathbb{R}}^+)$ is defined as $\varphi_t(\alpha) = \varphi(t-\alpha)$ for any $\alpha\in\bar{\mathbb{R}}^+$. This convention is used to indicate an argument of a functional [21,22]. Accordingly, if $\psi\in\mathcal{B}^n(\mathbb{R})$ and $t\in\mathbb{R}$ then the history $\psi_t\in\mathcal{B}^n(\bar{\mathbb{R}}^+)$ can be defined as $\psi_t(\alpha) = \psi(t-\alpha)$. Probability of an event A is denoted by $\mathsf{P}(A)$.

2.1 The class of systems to be considered

The mathematical description of the class of systems considered in this paper is given by the stochastic functional differential equation (1) where $t\in\mathbb{R}$ is a time instant, $x(t)\in\mathbb{R}^n$ is the state vector, $\theta\in\mathcal{B}^m(\mathbb{R})$ and f: $C^n(\bar{\mathbb{R}}^+)\times\mathbb{R}\times\mathbb{R}^m \to \mathbb{R}^n$ is a functional.

$$\dot{x}(t) = f\big(x_t, t, \theta(t)\big) \tag{1}$$

Because of the randomness caused by θ, the state x is also a stochastic process. The initial time is denoted by t_0 up to which the state information is available.

The functional f is assumed to satisfy (2) for every $\varphi\in C^n$, $t\in\mathbb{R}$, $\theta\in\mathbb{R}^m$ which indicates that x=0 (or the origin) is an equilibrium solution of (1).

$$\|\varphi\|_\infty = 0 \quad \Rightarrow \quad f(\varphi, t, \theta) = 0 \tag{2}$$

Remark 2.1: The theory of stochastic differential equations (SDEs) is mainly concerned with the Itô SDE [4]. In an Itô SDE, $\theta(t)$ is basically a white noise process and f is affine with respect to θ. There are existence and uniqueness results for solution of Itô SDEs with delays [21].

Assumption 1: In this paper it will be assumed that (1) has a unique solution $x\in\mathcal{B}^n(\mathbb{R})$ for every initial conditions.

We are interested in determining the stability of (1) where the concept of stability is presented in the next part.

2.2 Definition of stability

Three important stability notions in the literature are stability in probability, stability in p-th moment and almost sure stability. In this paper we will work with a generalized version of stability in p-th moment, which is also related to stability in probability (definition 2.2 in the following).

Definition 2.1: A continuous function u: $\bar{\mathbb{R}}^+ \to \bar{\mathbb{R}}^+$, $u(0)=0$, is said to belong to class K_d if it is non-decreasing and $u(\alpha)>0$ for $\alpha>0$. For u_1, $u_2 \in K_d$ it is said that u_1 covers u_2 if there exist $c>0$ such that $u_1(\alpha) \geq c\, u_2(\alpha)$ for every $\alpha\in\bar{\mathbb{R}}^+$.

Definition 2.2: For a class K_d function h, the equilibrium x=0 of system (1) is *h-mean stable* if for any $\varepsilon>0$ there exist $\delta(\varepsilon, t_0) > 0$ such that for any $t \geq t_0$ (3-1) holds. Moreover, x=0 is *asymptotically h-mean stable* if it is h-mean stable and there exists $\delta(t_0) > 0$ such that (3-2) holds.

$$\|x_{t_0}\| < \delta \Rightarrow E\{h(\|x(t)\|)\} < \varepsilon \tag{3-1}$$

$$\|x_{t_0}\| < \delta \Rightarrow \lim_{t\to\infty} E\{h(\|x(t)\|)\} = 0 \tag{3-2}$$

Remark 2.2: If we select h as $h(\alpha) = \alpha^p$ for some $p>0$, definition of stability in p-th moment in [3] is retrieved. Stability in second moment or mean square stability, is a very practical stability concept specially for linear systems.

Definition of stability in probability from [3] with a few modifications to express it for (1) is as below.

Definition 2.3: The equilibrium x=0 of system (1) is *stable in probability* if for every pair ε_1, $\varepsilon_2 > 0$, there exists $\delta(\varepsilon_1,\varepsilon_2,t_0) > 0$ such that (4-1) holds for any $t \geq t_0$. Also, x=0 is *asymptotically stable in probability* if it is stable in probability and for every $\varepsilon > 0$, there exists $\delta(\varepsilon, t_0) > 0$ such that (4-2) holds.

$$\|x_{t_0}\| < \delta \quad \Longrightarrow \quad P\{\|x(t)\| \geq \varepsilon_1\} \leq \varepsilon_2 \tag{4-1}$$

$$\|x_{t_0}\| < \tilde{\delta} \quad \Longrightarrow \quad lim_{t\to\infty} P\{\|x(t)\| \geq \tilde{\varepsilon}\} = 0 \tag{4-2}$$

Relationship between h-mean stability and stability in probability is stated as proposition 2.1 (proof is omitted).

Proposition 2.1: The system (1) is stable in probability if and only if there exists a K_d function h such that (1) is h-mean stable. Moreover, (1) is asymptotically stable in probability if there exists a K_d function h such that (1) is asymptotically h-mean stable.

Remark 2.3: Any of the above stability properties is said to be global when the value of related functions δ or δ can be made arbitrarily large by adjusting their first argument. Moreover, a stability property is said to be uniform if the related functions δ or δ are independent from t_0 (similar to the delay-free deterministic case in [23]).

Remark 2.4: The stability property of a stochastic system can have different qualities. A system may be stable in first moment but not mean square stable. Due to proposition 2.1, quality of stability for a system that is stable in probability can be studied by finding h. For example, a faster growth of h can imply a better convergence. Therefore, the h-mean stability is a strong and exact stability notion.

3. Stability Theorem

In this section the main results of the paper is presented. First a Lyapunov theorem is proposed for the delayed system (1). Then, the theorem is rewritten for the special case of a delay free system.

3.1 Delayed systems

The Lyapunov stability theorem for (1) is as below.

Theorem 3.1: The system (1), is w_1-mean stable if there is a differentiable functional V: $C^n(\bar{\mathbb{R}}^+)\times\mathbb{R}\rightarrow \bar{\mathbb{R}}^+$ satisfying (5) for some $w_1,w_2\in K_d$ and there exist $r\in\bar{\mathbb{R}}^+$ such that $E\{^d/_{dt}V(x_t,t)\}$ is well-defined and non-positive for every $\|x_{t_0}\|_\infty< r$, $t \geq t_0$. Also, (1) is asymptotically h-mean stable for $h\in K_d$ if there exist $u\in K_d$ such that $\|x_{t_0}\|_\infty< r$ implies $E\{^d/_{dt}V(x_t,t)\} \leq -E\{u(\|x(t)\|)\}$ for every $t \geq t_0$ and h is covered by both u and w_1.

$$w_1(\|x(t)\|) \leq V(x_t,t) \leq w_2(\|x\|_\infty) \tag{5}$$

Proof: According to Equation (1), every sample path of x is differentiable with respect to t. Hence, due to differentiability of V, the time derivative of V exists and it can be easily shown that the time derivation operator *d/dt* commutes with the expectation operator E as below.

$$E\left\{\frac{d}{dt}V(x_t,t)\right\} = E\left\{\lim_{s\to t}\frac{V(x_s,s)-V(x_t,t)}{s-t}\right\} =$$
$$\lim_{s\to t}\frac{E\{V(x_s,s)\}-E\{V(x_t,t)\}}{s-t} = \frac{d}{dt}E\{V(x_t,t)\} \qquad \Rightarrow$$
$$\frac{d}{dt}E\{V(x_t,t)\} = E\left\{\frac{d}{dt}V(x_t,t)\right\} \tag{6}$$

Also, according to continuity of w_2, for every $\varepsilon > 0$ we can select $0 < \delta < r$ such that (7) is satisfied.

$$w_2(\delta) < \varepsilon \tag{7}$$

Theorem 3.1 has two parts, proved in the following.

Part 1: According to the selected $\delta < r$ and conditions of the theorem, if we select $\|x_{t_0}\|_\infty< \delta$ then $E\{^d/_{dt}V\} \leq 0$ and consequently $^d/_{dt}E\{V\} \leq 0$ due to (6). This implies non-increasing behavior of $E\{V\}$ with time. Using this fact, (5) and (7) one can write (8) which proves the first part according to definition 2.2.

$$E\{w_1(\|x(t)\|)\} \leq E\{V(x_t,t)\} \leq V(x_{t0},t_0) \leq w_2(\delta) < \varepsilon \tag{8}$$

Part 2: For ε and δ in (7) and $\|x_{t_0}\|_\infty< \delta$ we have (9) which implies decreasing behavior of $E\{V\}$ and (8) as in previous part of proof.

$$E\{^d/_{dt}V(x_t,t)\} < -E\{u(\|x(t)\|)\} < 0 \tag{9}$$

Since $0\leq E\{V\} < \varepsilon$, its decreasing behavior implies that $m = \lim_{t\to\infty} E\{V\}$ is a constant between 0 and ε. Now we define $\hat{V}(x_t,t) = m + \int_t^\infty E\{u(\|x(s)\|)\}ds$. According to (9), it follows that $m \leq \hat{V}(x_t,t) \leq E\{V(x_t,t)\}$ and consequently $\lim_{t\to\infty} E\{\hat{V}\}=$ m (using the squeeze lemma). By definition of $\hat{V}$ we have $^d/_{dt}\hat{V} = -E\{u(\|x(t)\|)\}$ and we can apply the Barbalat's lemma to conclude (10) as below. Because $^d/_{dt}\hat{V}$ is continuous with respect to time according to continuity of u and differentiability of x with respect to t in (1).

$$\lim_{t\to\infty} E\{u(\|x(t)\|)\} = 0 \tag{10}$$

Since u covers h, there exist a constant $c_1\in\bar{\mathbb{R}}^+$ such that $0<c_1h(\|x\|)<u(\|x\|)$. Taking expectation from this inequality and tending t to infinity we obtain $\lim_{t\to\infty} E_t\{h(\|x(t)\|)\} = 0$ according to (10). This fact together with Lemma 3.1 in the following proves the second part. □

Lemma 3.1: if (1) is w-mean stable for some $w\in K_d$ then it is h-mean stable for every $h\in K_d$ that is covered by w.

Proof: The w-mean stability of (1) can be written as (11) according to definition 2.2.

$$\forall\bar{\varepsilon} > 0, \exists\delta > 0 \mid \|x_{t_0}\| < \delta \Rightarrow E_t\{w(\|x(t)\|)\}< \bar{\varepsilon} \tag{11}$$

There exist $c_1\in\bar{\mathbb{R}}^+$ such that $w(\|x\|) > c_1h(\|x\|)$. Combining this inequality with consequent part of (11) and setting $\bar{\varepsilon}$ to $c_1\varepsilon$ for an arbitrary $\varepsilon > 0$, one can write the following and obtain (12) which is equivalent to h-mean stability of (1) according to definition 2.2.

$$\forall\varepsilon > 0, \exists\delta > 0 \mid \|x_{t_0}\| < \delta \Rightarrow$$
$$c_1 E_t\{h(\|x(t)\|)\} < E_t\{w(\|x(t)\|)\} < \bar{\varepsilon} = c_1\varepsilon$$
$$\forall\varepsilon > 0, \exists\delta > 0 \mid \|x_{t_0}\| < \delta \ \Rightarrow\ E_t\{h(\|x(t)\|)\}< \varepsilon \tag{12}$$

The h-mean stability notion has a natural relationship with theorem 3.1, which results in shortening the proof of theorem. This fact and remark 2.4, are main reasons of using h-mean stability notion in this work.

Remark 3.1: For every set $U = \{u_i\in K_d : 1\leq i\leq n\}$ of K_d functions, there always exist functions that are covered by all elements of U. An example is $h_1(\alpha) = \inf_i \{b_i u_i(\alpha)\}$ where b_i ($1\leq i\leq n$) are arbitrary positive real numbers (since $u_i \geq b_i^{-1} h_1$). This ensures that in second part of theorem 3.1 there always exists a function h that is covered by u and w_1.

Remark 3.2: The stochastic control theory ([5,7]) is mainly concerned with systems modeled by Itô SDEs (remark 2.1). As a result, second derivatives of V appear in calculation of $E_t\{^d/_{dt}V\}$. In this work, no assumption is made about θ and $E_t\{^d/_{dt}V\}$ is not calculated. As a result, theorem 3.1 is applicable to problems that are different from the problems commonly modeled by the Itô SDE. An application is the case of next section. Other applications may include randomly switched systems [8].

3.2 Delay free systems

State vector $x(t)$ is a part of history x_t. Hence, the delay free system (13) is a special case of delayed system (1).

$$\dot{x}(t) = f\big(x(t),t,\theta(t)\big) \tag{13}$$

Stability definitions 2.2 and 2.3 are written for (13) by replacing $\|x_{t_0}\|$ in antecedents of (3) and (4) with $\|x(t_0)\|$.

Accordingly, theorem 3.1 in previous part is simplified to theorem 3.2 for the delay free system (13).

Theorem 3.2: System (13), is w_1-mean stable if there exist a function V: $\mathbb{R}^n\times\mathbb{R}\rightarrow\mathbb{R}^+$+ and $w_1,w_2\in K_d$ such that (14) is satisfied and there exist $r\in\mathbb{R}^+$ such that $\|x(t_0)\| < r$ implies $E\{{}^d/_{dt}V(x(t),t)\} \leq 0$ for every $t \geq t_0$. Additionally, for $h\in K_d$ (1) is asymptotically h-mean stable if there exist $u\in K_d$ such that $\|x(t_0)\| < r$ implies $E\{{}^d/_{dt}V(x(t),t)\} \leq -E\{u(\|x(t)\|)\}$ for every $t \geq t_0$ and h is covered by both u and w_1.

$$w_1(\|x\|) \leq V(x,t) \leq w_2(\|x\|) \tag{14}$$

4. Application to NCS Problem

In this section, a nonlinear NCS problem will be studied which has been originally proposed in [17]. The configuration of this NCS is depicted in Fig.1. In [17] it is assumed that there is no communication delay and the goal is to obtain a bound on maximum allowed time interval between data transmissions that can guarantee the stability of NCS. This bound is known as MATI (maximum allowable transfer interval).

Using Theorem 3.1, the problem can be extended from two different aspects. First, instead of finding a bound, we will be able to check stability when some probabilistic data about the transfer intervals is available as a PDF. Second, it will be possible to handle an NCS with communication delays. However, due to complexities of the extension to delayed case and the limited space, we will only focus on the extension to the probabilistic case in this work.

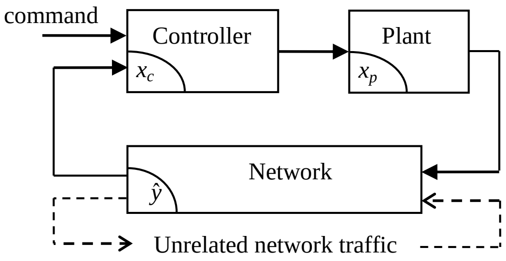


Fig 1. The networked control system (NCS).

The plant and controller can be modeled as (15) and (16) respectively in which x_p is the state of plant, x_c is the controller state, u_p is the plant input, y is the plant output and $\hat{y} = y(\hat{t})$ is the latest sample of y available at controller which is obtained at sampling instant $\hat{t}$ ([17]).

$$\dot{x}_p = f_p(x_p, u_p, t),\ \ y = g_p(x_p, t) \tag{15}$$

$$\dot{x}_c = f_c(x_c, \hat{y}, t),\ \ \ u_p = g_c(x_c, \hat{y}, t) \tag{16}$$

The above equations can be combined during time interval that $\hat{y}$ is constant (no updated data is received). The result is (17) in which x^T=[x_p^T x_c^T], $\hat{x} = x(\hat{t})$ and the network induced error is defined as $e = x - \hat{x}$.

$$\dot{e} = f(e, \hat{x}, t) \tag{17}$$

For simplicity, we will consider the situation where all feedback data is transmitted at once. This is the case for example when the plant is single-input single-output. However, the results can be extended to the case of multiple transmitters. It is assumed that a continuously differentiable and positive definite Lyapunov function $V(x,t)$ exists that satisfies (18) to (20) globally with positive real numbers c_1, c_2, c_3, c_4.

$$c_1\|x\|^2 \leq V(x,t) \leq c_2\|x\|^2 \tag{18}$$

$$\frac{\partial}{\partial t}V + \frac{\partial}{\partial x}V\, f(0,x,t) \leq -c_3\|x\|^2 \tag{19}$$

$$\left\|\frac{\partial}{\partial x}V\right\| \ \leq c_4\|x\| \tag{20}$$

The functions f and g are also assumed to be globally Lipschitz such that one can write (21).

$$\|f(e,\hat{x},t)\| \leq k_1\|e\| + k_2\|\hat{x}\| \tag{21}$$

$$\|f(e, x-e, t) - f(0,x,t)\| \leq k_p\|e\| \tag{22}$$

Time derivative of V can be calculated as below using (19), (20) and (21-1).

$$\dot{V} = \frac{\partial V}{\partial t} + \frac{\partial V}{\partial x}\,\dot{x} = \frac{\partial V}{\partial t} + \frac{\partial V}{\partial x}\,f(e,\hat{x},t) =$$

$$\frac{\partial V}{\partial t} + \frac{\partial V}{\partial x}\,f(0,x,t) + \frac{\partial V}{\partial x}\,[f(e,x-e,t) - f(0,x,t)]$$

$$\dot{V} \leq -c_3\|x\|^2 + c_4\|x\|k_p\|e\|$$

Using triangle inequality we have

$$\dot{V} \leq -c_3[\|\hat{x}\| - \|e\|]^2 + c_4(\|\hat{x}\| + \|e\|)k_p\|e\|$$

$$\dot{V} \leq -c_3\|\hat{x}\|^2 + (c_4k_p + 2c_3)\|\hat{x}\|\|e\| + (c_4k_p - c_3)\|e\|^2 \tag{23}$$

Using (17) and (21), we can write

$$\frac{d}{dt}\|e\| \leq \|\dot{e}\| = \|f(e,\hat{x},t)\| \leq k_1\|e\| + k_2\|\hat{x}\|$$

Using the comparison lemma ([23]) we obtain an upper bound on $\|e\|$ as below

$$\|e\| \leq \frac{[e^{k_1\Delta t} - 1]k_2}{k_1}\|\hat{x}\| \tag{24}$$

$$\Delta t = t - \hat{t}$$

In the same way, we can obtain a differential inequality and solve in the reverse direction of time from t to $\hat{t}$ to obtain the following bounding

$$\|\hat{x}\| \leq \frac{[e^{(k_1+k_2)|\Delta t|} - 1]k_1}{k_1 + k_2}\|x\| \tag{25}$$

Relations (23) and (24) give an upper bound on $\dot{V}$

$$\dot{V} \leq \phi_a(\Delta t)\|\hat{x}\|^2 \tag{26}$$

$$\phi_a(\Delta t) = -c_3 + (c_4k_p + 2c_3)\frac{[e^{k_1\Delta t} - 1]k_2}{k_1} + |c_4k_p - c_3|\left[\frac{[e^{k_1\Delta t} - 1]k_2}{k_1}\right]^2 \tag{27}$$

For every i, the ith transfer interval is denoted by T_i. In deterministic case, $V \leq 0$ guarantees stability due to the common Lyapunov theorem. According to (26) $V \leq 0$ is resulted from $\phi_a(\Delta t) \leq 0$.

The function, $\phi_a(s)$ is increasing with $\phi_a(0) < 0$. Therefore, $\phi_a(s)=0$ has a unique positive solution τ_a which is a lower bound for MATI because $T_i \leq \tau_a$ implies $\phi_a(\Delta t) \leq 0$. This is similar to the results in [17]. But how would be the NCS stability if T_i can exceed τ_a. To answer such a question we can use theorem 3.2 in previous section. Taking expectation from (25) we have (28).

$$E\{\dot{V}\} \leq E\{\phi_a(\Delta t)\}\|\hat{x}\|^2 \qquad (28)$$

Also we can obtain (29) from (26) by a few manipulations and taking expectation.

$$E\{\phi_b(\Delta t)\}\|\hat{x}\|^2 \leq E\{\|x\|^2\} \qquad (29)$$

$$\phi_b(\Delta t) = \left[\frac{\left[e^{(k_1+k_2)|\Delta t|}-1\right]k_1}{k_1+k_2}\right]^{-2}$$

Now we can apply Theorem 3.2 as explained in the following. Due to (18), Lyapunov function V satisfies the condition (14) of Theorem 3.2 with $w_1(\alpha) = c_1\alpha^2$ and $w_2(\alpha) = c_2\alpha^2$. Therefore, the NCS is asymptotically mean square stable if $E\{\dot{V}\} \leq -c_5E\{\|x\|^2\}$ for some $c_5 > 0$. But, if $E\{\phi_a(\Delta t)\}<0$ then we can combine (28) and (29) to obtain a positive constant $c_5 = -E\{\phi_a(\Delta t)\} / E\{\phi_b(\Delta t)\}$. However, for every i, $E\{\phi_a(\Delta t)\}<0$ is satisfied if $E_{T_i}\{\phi_a(T_i)\} < 0$ because ϕ_a is increasing and $\Delta t \leq T_i$. The result can be summarized as following corollary.

Corollary 4.1: NCS (15), (16) is asymptotically mean square stable (asymptotically stable in second moment) if the transfer intervals T_i from y to $\hat{y}$ have a common PDF denoted by p_T and there exists a Lyapunov function V for the closed loop system with $\hat{y} = y$ that satisfies (18), (19), (20), and the following condition is satisfied with ϕ_a defined in (26).

$$E\{\phi_a(T_i)\} = \int_0^\infty p_T(s)\,\phi_a(s)ds < 0 \qquad (30)$$

Remark 4.1: If the random variations of transfer intervals T_i are due to data packet losses (packet transmission errors) with probability $\mathrm{p_e}$ and the sampling period is equal to h, then we have (31) in which δ is the Dirac's delta function.

$$p_T(s) = \sum_{i=1}^{\infty}(1-p_e)p_e^{i-1}\delta(s-ih) \qquad (31)$$

Replacing (27) and (31) in the left hand side of (30), eliminating the Delta function and integration we obtain

$$E\{\phi_a(T_i)\} = \left(c_4k_p + 2c_3\right)\frac{k_2}{k_1}\sum_{i=1}^{\infty}(1-p_e)p_e^{i-1}\left[e^{k_1ih}-1\right]$$
$$+\left|c_4k_p - c_3\right|\left[\frac{k_2}{k_1}\right]^2\sum_{i=1}^{\infty}(1-p_e)p_e^{i-1}\left[e^{k_1ih}-1\right]^2 - c_3$$

The right hand side of the above equation contains two geometric series with common ratios $p_e\exp(k_1h)$ and $p_e\exp(2k_1h)$ that must be smaller than one to ensure the convergence. Since the later one is always greater, it suffices to have $p_e < \exp(-2k_1h)$. Simplifying the result, the condition (30) can be represented as below

$$p_e < e^{-2k_1h} \qquad (32\text{-}1)$$

$$E\{\phi_a(T_i)\} =$$
$$\beta_0 + \beta_1\frac{e^{k_1h}(1-p_e)}{1-e^{k_1h}p_e} + \beta_2\frac{e^{2k_1h}(1-p_e)}{1-e^{2k_1h}p_e} < 0 \qquad (32\text{-}2)$$
$$\beta_0 = -c_3 - \left(c_4k_p + 2c_3\right)\frac{k_2}{k_1} + \left|c_4k_p - c_3\right|\left[\frac{k_2}{k_1}\right]^2$$
$$\beta_1 = \left(c_4k_p + 2c_3\right)\frac{k_2}{k_1} - 2\left|c_4k_p - c_3\right|\left[\frac{k_2}{k_1}\right]^2$$
$$\beta_2 = \left|c_4k_p - c_3\right|\left[\frac{k_2}{k_1}\right]^2$$

which can be calculated to rewrite (30) as below provided that $p_e < \exp(-2k_1h)$.

Example 4.1: The following NCS is considered.

$$\dot{x}_1 = x_2 - x_1$$
$$\dot{x}_2 = (x_2 + x_1)\,sin\,x_1 + u$$
$$y = x_2$$
$$u = -3\hat{y}$$

A Lyapunov function for the closed loop with $\hat{y} = y$ can be $V=½(x_1^2 + x_2^2)$ with $c_1 = c_2 = ½$, $c_3 = 0.38$, $c_4 = 1$ which guarantees deterministic stability globally. Equations (17) can be written as below.

$$\dot{e}_1 = \hat{x}_2 - \hat{x}_1 + e_2 - e_1$$
$$\dot{e}_2 = (\hat{x}_2 + \hat{x}_1 + e_2 + e_1)\,sin(\hat{x}_1 + e_1) - 3\hat{x}_2$$

Using the above equations, we can obtain k_1=k_2=3.34 and k_p=3. This completely determines the function ϕ_a in (26). The obtained ϕ_a gives τ_a = 0.0215 which guaranties the stability for $T_i \leq 0.0215$.

Many random communication effects can be studied using Corollary 4.1. It is assumed that we have transmissions with possibility of error as explained in Remark 4.1. In Figure 2 the maximum value of p_e that satisfies (32) is plotted as a function of h for the obtained ϕ_a. This plot gives a lower bound for the stability margin of the packet loss probability p_e.

5. A Practical NCS Application

In this section we study a practical NCS that consists of a dual axis magnetic levitation system controlled over a communication link to a computer. In the following, in the first part we describe the control system. In the second part we describe the communication system and its limitations. In the third and last part we use the results of this paper to select the communication data rate such that the stochastic stability of the control system is preserved.

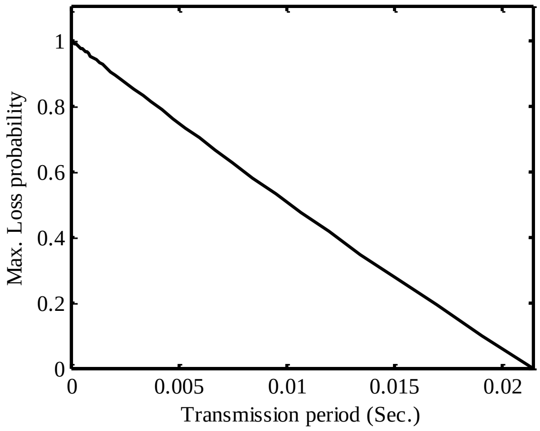
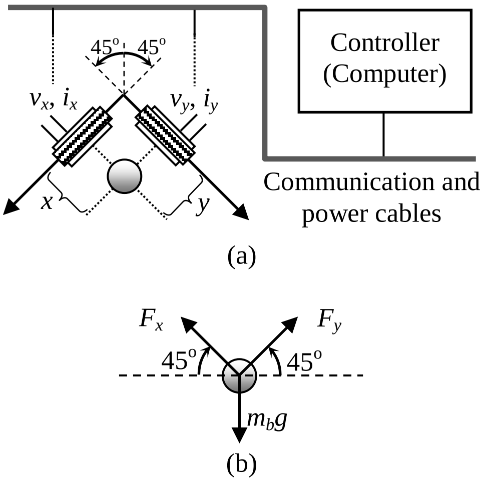


Fig. 2. Transmission period h versus loss probability p_L

5.1 Control system

The magnetic levitation system is composed of a steel ball with mass m_b=40 g affected by two magnetic forces F_x, F_y generated by two identical solenoids with voltages v_x, v_y and currents i_x, i_y respectively as shown in Figure 3 (a). The forces acting on the ball are Fx, F_y and the weight of ball $m_b g$ (g = 9.8 m/sec.^{2} is the acceleration of gravity) as depicted in Figure 3 (b).

Assuming that the axes of coils (x and y axes) are perpendicular, the magnetic forces F_x, F_y can be calculated from the following equations in which K_f=0.0047 Kg m^3/C^2 is magnetic force constant.



(a)

(b)

Fig. 3. (a) The magnetic levitation control system. (b) The forces acting on the ball.

$$F_x = K_f \left({}^{i_x}/_{x}\right)^2, \quad F_y = K_f \left({}^{i_y}/_{y}\right)^2 \tag{33}$$

The resistance and inductance of the coils are R=0.62 Ω and L=0.32 H such that we can write

$$v_x = R\, i_x + L\, \dot{i}_x\,, \quad v_y = R\, i_y + L\, \dot{i}_y \tag{34}$$

The equations of motion of the ball are also as below.

$$m_b\ddot{x} = m_b g\sqrt{2}/2 - F_x\,,$$
$$m_b\ddot{y} = m_b g\sqrt{2}/2 - F_y \tag{35}$$

The control algorithm is implemented in a computer (Figure 1) that receives the measurement feedbacks i_x, i_y, x, y, $\dot{x}, \dot{y}$ from the node which is connected to the coils and sends back the control commands v_x and v_y to the coils.

The control commands to the input voltages v_x and v_y are calculated using the feedback linearization method as follows. First we differentiate the equations in (35) and summarize the results as in the following two equations (detailed representation of functions f_a and g_a are omitted for brievity).

$$\dddot{x} = f_a(x,\dot{x},i_x) + g_a(x,\dot{x},i_x)v_x\,,$$
$$\dddot{y} = f_a(y,\dot{y},i_y) + g_a(y,\dot{y},i_y)v_y \tag{36}$$

Based on the above equations we design the control laws in (37) to achieve the closed loop transfer functions in (38) where x_d and y_d are the desired values for x and y.

$$v_x = \frac{-\alpha_1\ddot{x} - \alpha_2\dot{x} - \alpha_3(x - x_d) - f_a(x,\dot{x},i_x)}{g_a(x,\dot{x},i_x)}$$
$$v_y = \frac{-\alpha_1\ddot{y} - \alpha_2\dot{y} - \alpha_3(y - y_d) - f_a(y,\dot{y},i_y)}{g_a(y,\dot{y},i_y)} \tag{37}$$

$$\frac{X(s)}{X_d(s)} = \frac{Y(s)}{Y_d(s)} = \frac{1}{s^3 + \alpha_1 s^2 + \alpha_2 s + \alpha_3} \tag{38}$$

The values of $\ddot{x}$ and $\ddot{y}$ in (37) are obtained from equations (33) through (35) in terms of the measured variables.

The controller parameters are selected as α_1 = 4.47, α_2 = 8.64, α_3 = 6.1 and the control objective is to maintain the ball at position x_d=y_d = 0.65 m.

5.2 Communication

The solenoids require electric power which is supplied through power cables. To reduce wiring we would like to transmit the control data through the power cables. The transmission bit-rate is denoted by f_{tx}. The communication through the power cables suffers from the noise in an industrial environment (we assume that there is no bandwidth limitation). We denote the noise at the bit detection stage by $w(t)$ and assume that it is Gaussian with $E\{w(t)\}$ = 0, $E\{w^2(t)\}$ = σ_w^2 = 1 and power spectral density $S_w(f)$ in (39) where f_b=10^{8}Hz is the noise bandwidth.

$$S_w(f) = \begin{cases} \sigma_w^2/2f_b & |f| \le f_b \\ 0 & |f| > f_b \end{cases} \tag{39}$$

The signal level at receiver is assumed to be v_{bit}=1 v. In general, the reliability of data transmission increases if we reduce f_{tx}. For example with a smaller f_{tx}, we can decrease the bandwidth of the low-pass filter at the baseband processing stage in the receiver to reduce the effect of noise on the bit detection. We use a simple low-pass filter with transfer function $H_f(s)$ = 1/[1+s/(4πf_{tx})]. Then the power spectral density of the filtered noise $w_f(t)$ is

$$S_{w_f}(f) = \frac{1}{1 + (f/f_{tx})^2} S_w(f) \tag{40}$$

The variance (power) of $w_f(t)$ can be calculated as

$$\sigma_{w_f}^2 = f_{tx} \tan^{-1}(f_b/f_{tx}) \quad (41)$$

Assuming that the sampling for bit detection is performed at the end of the bit hold time $1/f_{tx}$ and neglecting the effect of filter on the signal amplitude at the sampling instant, the probability of a bit transmission error p_{bit} is

$$p_{bit} = 1 - \frac{1}{2} erf\left(\frac{v_{bit}}{\sqrt{2}\,\sigma_{w_f}}\right) \quad (42)$$

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-s^2} ds$$

Consequently probability of a frame transmission error p_e is obtained as below where n_f is the length of frame in bits

$$p_e = 1 - (1 - p_{bit})^{n_f} \quad (43)$$

5.3 Analysis

In this part we study the interaction of the control system and communication described in the previous parts of this section to select bit-rate of communication f_{tx} such that the control loop remains stable.

Each control cycle begins with a sampling at sensors on the solenoids side, transmission of the measurement data through the coils-controller link to the control computer, execution of the control algorithm which is assumed to take τ_c = 50 μs and sending back the voltage commands through the controller-coils link to the solenoids. The sensor measurements include 6 values and control commands include 2 values as described previously. Assuming that each value is encoded in 10 bits and that the framing adds 10 extra bits as header, the transmission time of the measured values and command values are τ_s=70/f_{tx} and τ_a=30/f_{tx} respectively. We assume that the control loop is allowed to use 33 percents of the communication capacity (time division). Hence, the length of control cycle becomes

$$h = \tau_c + 3[\tau_s + \tau_a] = 300/f_{tx} + 50\times10^{-6} \quad (44)$$

Since the controller in (37) is static (it does not have states), we can concatenate the communication delays and assume that there is a single delay of length h during a control cycle.

Now, for a given value of transmission bit-rate f_{tx} we can obtain p_e and h from (43) and (44) and use condition (32) in remark 4.1 to calculate $E\{\phi_a(T_i)\}$ in Corollary 4.1 and check the stability. This is performed for a range of values for f_{tx} in Figure 4.

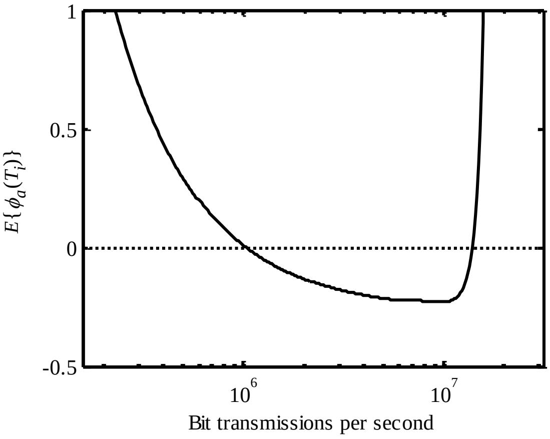


Fig. 4. (a) The magnetic levitation control system.

According to Figure 4, the stability of control loop is preserved between f_{tx}=1 MBPS (Mega bit per sec.) and f_{tx}=14 MBPS. A good selection is f_{tx}=5 MBPS which is sufficiently away from the stability margins and gives a sampling frequency of 9.1 KHz.

6. Conclusions

In this paper, we observed that the focus of the theory of stochastic systems has been centered on a special kind of approach to modeling the stochastic phenomena. Even if this approach has been sufficient for the past applications, with the growing complexity of the new systems, it is expectable that we will need to expand the capabilities of the existing analysis frameworks.

Based on this observation, a stochastic Lyapunov theorem was presented. This theorem benefits from a higher level of generality. This was shown by applying the theorem to a practical NCS problem. The result is a new stability analysis criterion for a stochastic nonlinear NCS that cannot be obtained using the traditional versions of the stochastic Lyapunov theorem. However, we had to carry out some additional calculations in section four in order to be able to apply the theorem from section three to the NCS problem. This seems to be the cost that we have to pay for the generality that we have obtained. There are more potential applications to the various NCS problems that will be investigated in future works.

References

[1] X. Mao, Stochastic Differential Equations and Applications, 2nd Ed. Horwood Pub., 2007.

[2] R. Z. Khasminskii, Stochastic Stability of Differential Equations, 2nd Ed., Springer, 2012.

[3] Y. K. Lin and G. Q. Cai, Probabilistic Structural Dynamics, McGraw-Hill. 2004.

[4] B. Øksendal, Stochastic Differential Equations, 6th edition, Springer-Verlag, New York, 2003.

[5] H. J. Kushner, Stochastic Stability and Control. Academic Press, 1967.

[6] W. Chen, L.C. Jiao, "Finite-time stability theorem of stochastic nonlinear systems", Automatica, Vol. 45, pp. 2105-2108, 2010.

[7] E. Samiei, S. Torkamani, E.A. Butcher, "On Lyapunov stability of scalar stochastic time-delayed systems", Int. Journal of Dynamics and Control, Vol. 1, No. 1, pp. 64-80, 2013.

[8] J. Xiong, J. Lam, Z. Shu, X. Mao, "Stability Analysis of Continuous-Time Switched Systems With a Random Switching Signal", IEEE Trans. on Automatic Control, Vol. 59, No. 1, pp. 180-186, 2014.

[9] L. Zhang, H. Gao, O. Kaynak, "Network-induced constraints in networked control systems—A survey", IEEE Transactions on Industrial Informatics, Vol. 9, No. 1, pp. 403-416, 2013.

[10] W. Zhang, M. S. Branicky and S. M. Phillips, “Stability of networked control systems”, IEEE Control Systems Magazine. Vol. 21, No. 2, pp.84 99, 2001.

[11] S. Yuksel, T. Basar, "Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints", Birkhauser-Springer, 2013.

[12] J. Nilsson, Real-time control systems with delays, Ph.D. dissertation, Dept. Automatic Control, Lund Institute of Technology, Lund, Sweden, 1998.

[13] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, S.S. Sastry, “Foundations of Control and Estimation Over Lossy Networks”, Proceedings of the IEEE Vol.95, No.1, pp.163-187, Jan. 2007.

[14] S. Yuksel, "Characterization of information channels for asymptotic mean stationarity and stochastic stability of nonstationary/unstable linear systems", IEEE Transactions on Information Theory, Vol. 58, No. 10, pp. 6332-6354, 2012.

[15] B. Safarinejadian, A. Rahimi, "Optimal Sensor Scheduling Algorithms for Distributed Sensor Networks", Journal of Information Systems and Telecommunication, Vol. 1, No. 3, pp. 175-181, 2013.

[16] M. Pajic, R. Mangharam, G.J. Pappas, "Topological conditions for in-network stabilization of dynamical systems", IEEE Journal on Selected Areas in Communications, Vol. 31, No. 4, pp. 794-807, 2013.

[17] G.C. Walsh, O. Beldiman, L.G. Bushnell, “Asymptotic behavior of nonlinear networked control systems”, IEEE Trans. on Automatic Control, Vol.46, No.7, pp.1093-1097, 2001.

[18] W. P. M. H. Heemels, A. R. Teel, N. van de Wouw, D. Nesic, “Networked Control Systems With Communication Constraints: Tradeoffs Between Transmission Intervals, Delays and Performance”, IEEE Transactions on Automatic Control, Vol. 55, No. 8, pp. 1781-96, Aug. 2010.

[19] B. Tavassoli, “Stability of Nonlinear Networked Control Systems over Multiple Communication Links with Asynchronous Sampling”, IEEE Transactions on Automatic Control, Vol. 59, No. 2, pp. 511-515, 2014.

[20] B. Tavassoli, P. Jabehdar-Maralani, "A Probabilistic Approach to Lyapunov Method with Application to Networked Control Systems", 20th Iranian Conference on Electrical Engineering, Tehran, Iran, 2012.

[21] Mohammed, S.E. (1984) Stochastic functional differential equations. Pitman advanced publishing program, research notes in mathematics, 99.

[22] K. Gu, V. L. Kharitonov and J. Chen, Stability of Time Delay Systems, Birkhauser, Boston, 2003.

[23] H. K. Khalil, Nonlinear Systems. 2nd Edition, Prentice Hall, NJ. 1996.

Babak Tavassoli received the B.S. in electronics engineering in 1998, M.S. and Ph.D. degrees in control engineering in 2001 and 2009 from the University of Tehran, Tehran, Iran. During 2009-2010 he has been involved in process monitoring and control projects at the Research Institute of Petroleum Industry, Tehran, Iran. He has been involved in developing industrial automation systems based on Foundation Fieldbus™ standards at Farineh Fannavar Co., Tehran, Iran between 2003 and 2008. Since 2010, he is an assistant professor at K.N. Toosi University of Technology, Tehran, Iran where five M.S. theses have been completed under his supervision by the end of 2013. His research and teaching is mainly focused on Networked Control Systems, Hybrid Systems, Industrial Automation and process control.

Parviz Jabehdar-Maralani was born 1941 in Tabriz, Iran. He received the B.Sc. degree in electrical (1st. rank) from University of Tehran in 1963 and MS and Ph.D. degrees both in electrical engineering from The University of California, Berkeley in 1966 and 1969 respectively. From 1969 to 1970 he was with the AT&T Bell labs in Holmdel, NJ. He joined the electrical engineering department of the University of Tehran in 1970 and worked as a professor of the School of Electrical and Computer Engineering for 41 years. He retired in 2011 upon his own request. He was selected as a distinct professor of the University of Tehran in 1994. He served as the head of the school of electrical and computer engineering, the head of electro-technique institute, chairman of the department, director of graduate studies and the director of the Informatics Center at the University of Tehran. His current research interests are: Circuits and Systems, Control Systems, Computer Aided Analysis and Design, and Electrical Engineering Curriculum Development and Planning. He is a fellow of the Academy of Science of Iran. He has written more than 150 technical papers and authored 16 textbooks.

GoF-Based Spectrum Sensing of OFDM Signals over Fading Channels

Seyed Sadra Kashef
Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran
s.kashef@modares.ac.ir
Paeiz Azmi*
Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran
pazmi@modares.ac.ir
Hamed Sadeghi
Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran
hamed.sadeghi@modares.ac.ir

Received: 09/Dec/2013 Accepted: 12/Apr/2014

Abstract

Goodness-of-Fit (GoF) based spectrum sensing of orthogonal frequency-division multiplexing (OFDM) signals is investigated in this paper. To this end, some novel local sensing methods based on Shapiro-Wilk (SW), Shapiro-Francia (SF), and Jarque-Bera (JB) tests are first studied. In essence, a new threshold selection technique is proposed for SF and SW tests. Then, three studied methods are applied to spectrum sensing for the first time and their performance are analyzed. Furthermore, the computational complexity of the above methods is computed and compared to each other. Simulation results demonstrate that the SF detector outperforms other existing GoF-based methods over AWGN channels. Furthermore simulation results demonstrate the superiority of the proposed SF method in additive colored Gaussian noise channels and over fading channel in comparison with the conventional energy detector.

Keywords: Cognitive Radio, Spectrum Sensing, Goodness-of-Fit (GoF), Orthogonal Frequency Division Multiplexing (OFDM).

1. Introduction

The motivation for presentation of Cognitive Radio (CR) is the increasing need for higher bandwidth in wireless communications despite limited or licensed spectrum resources. Licensed spectrum is allocated over long time periods and is intended to be used only by licensed users. Different measurements of spectrum utilization have shown significant unused resources in three dimensions of frequency, time, and space [1]. Discovering these underutilized spectrum sources is the main idea behind CR by reusing spectrum holes in an opportunistic way [2]. In a CR network, spectrum sensing (SS) is the main duty of each CR user to find the unused spectrum, or equivalently, the primary users (PUs). One of the most challenging problems in this area is to find a solution to detect the existence and absence of PUs in the wireless communication [3].

Reliable PU detection problem is the main end of many SS algorithm proposals. For example, in the presence of PUs, when PU's signal is known, the best sensing method is matched filtering. However, when the primary signal is not perfectly known, energy detection (ED) method can be used instead of matched filtering [4]. In situations that SNR is low, distinguishing between PUs and noise is not simple. ED method is often considered for SS because of simplicity and admirable performance over SNR situations. However, uncertainty of the noise power quickly destroys the performance of ED. In practice, noise is an summation of various sources which can be changed significantly; therefore, usually uncertainty of noise variance exists and ranges about 1 to 2 dB [5]. By knowing characteristics of the incoming signals, different algorithms were recommended to increase the performance of ED consisting of waveform-based sensing and cyclostationarity-based sensing (See [6]). According to mathematical statistics, these methods are part of parametric hypothesis testing. It means that incorrect assumption about the received signals' parameters will degrade the performance. Accordingly, it is not easy to perform detecting the signal without key information. The appropriate signal features should be reported in the feature detecting methods (e.g. cyclostationarity). On the other hand, having information about the PUs is impossible practically in a CR receiver [7]. For instance, the matched filtering method, which provides the maximum SNR at the output of the detector, requires the exact knowledge of PU waveform. In addition, in the cyclostationary feature detection method, the cycle frequency of the primary signal should be known completely. Some PU detection algorithms based on statistical properties of eigenvalues of the covariance matrix of the received signals were devised in [8] in an attempt to compensate for the weaknesses of the above methods. However, the order of computational

* Corresponding Author

complexity of these algorithms is generally huge, which limits their practicality in CR devices [9].

To compensate weakness of the above methods, some PU detection methods have been proposed in the literatures. Recently, a higher-order-statistics (HOS) techniques was applied in [9] in SS problem for a reliable detection of PUs in the low SNR situations. In addition, a powerful sensing algorithm based on JB test [10] has been devised in [9], which is inherently a GoF testing problem. It has been shown that this method provides a high detection performance in very low SNRs [9]. Several GoF-based sensing methods have been recently proposed in the literatures [11],[12],[13],[14], where they provide superior performance in challenging opportunistic applications. Note that GoF-based sensing methods do not require any prior knowledge about the transmitted PU signals.

In this paper, we review three GoF techniques: Shapiro-Wilk (SW), Shapiro-Francia (SF), and Jarque-Bera (JB) tests. Then, we propose two SS methods based on SW and SF methods. Since these tests are kinds of Gaussianity tests, we assume that the distribution of channel noise is Gaussian. We compare these algorithms with each other, and also with a conventional GoF-based sensing approach, i.e., the Anderson-Darling (AD) test. In essence, we show that the computational complexities of the proposed methods are lower than the AD method. Also, we show that SF is faster than JB and SW since its computational complexity is lower. Furthermore, we will show through simulation results that SF outperforms the other candidates in different SNR values, signal sample sizes, and channel characteristics. Thus, the SF detector can effectively contribute to future CR networks.

It is straightforward to show that the first-order distribution of OFDM signals will converge to a Gaussian variable [15]. Suppose that an OFDM-based primary user signal is already present in the spectrum. In this case, the Gaussianity-based SS techniques would fail in PU detection; they will wrongly decide the hypothesis H_0 instead of H_1. Thus, we propose using FFT block as a preprocessing method to fix this problem when using GoF-based Gaussianity tests for sensing of OFDM signals. The idea is the fact that OFDM signals do not show Gaussian behavior in the frequency domain [16].

The organization of this paper is as follows. Section II introduces the GoF-based sensing and colored noise concept. Section III presents the statistical GoF tests. In Section IV, we propose the GoF-based spectrum sensing method. Section 5 presents the simulation results and finally, Section 6 concludes the paper.

2. GoF-based Spectrum Sensing Methods

Spectrum sensing algorithms must detect the presence/absence of PUs as quick as possible. If CR decides that the considered channel is empty, then it uses that frequency band for opportunistic communication. However, if CR misses the PU detection, it will cause a harmful interference to PU. Thus, the detection performance of the SS algorithm is an important factor in CR networks.

It is well-known that SS is a binary hypothesis testing problem as follows,

H_0: Presence of noise only

H_1: Presence of Primary User + noise

Let $y = \{y_i\}_{i=1}^{N}$ denotes N local time-domain observation samples collected at each CR. Without loss of generality, $y_i,\ i = 1,\cdot\cdot\cdot, N,$ are assumed to be real-valued. In situations that there is no primary transmission, $y_1,\cdot\cdot\cdot, y_N$ are only the noise samples. In this situation, they could be considered as an independent and identically distributed (i.i.d.) sequence with cumulative distribution function $F_0(y)$. However, based on the kind of modulation and communication link characteristics, the gathered samples $y_1,\cdot\cdot\cdot, y_N$ may not have distribution $F_0(y)$ when the PU's signal is present. In other words, this situation occurs when the received samples are not coming from the distribution function $F_0(y)$. Thus, the null hypothesis can be described as:

H_0 : $\boldsymbol{y}$ is an i.i.d. sequence obtained from distribution $F_0(y)$.

Furthermore, the alternative hypothesis (H_1) is the situation in which the received samples $\boldsymbol{y}$ do not form an i.i.d. sequence coming from distribution $F_0(y)$. There is no need for the knowledge of any information about the PU's signal in the above-mentioned GoF-based hypothesis testing problem and the type of noise distribution is the only assumption for detection.

Due to the presence of a colored channel interferer or some other reasons, the conventional white Gaussian noise may become colored. Therefore, in presence of colored noise the performance of SS methods is degraded. Independence of received samples is negligible in GoF based SS problem. The model of colored noise will be introduced in the next section.

2.1 Colored noise

It is clear that knowing the exact covariance matrix of the noise is necessary for the conventional energy detection method in colored noise, which is impractical in AWGN channel. The uncertainty of covariance matrix will lead to performance degradation because of inaccurate estimate of the noise parameters. According to the work in [17], here some new definitions on noise uncertainty in colored noise is introduced.

Based on [18], the colored Gaussian noise can be seen as the output of a single pole recursive filter stimulated by a white Gaussian noise (WGN). In mathematics, it can be expressed as $w(t) = -\eta w(t-1) + u(t)$, where $w(t)$ is the colored noise, $u(t)$ is the WGN with variance σ_u^2 and η ($|\eta| < 1$) is the correlation strength of the noise $w(t)$.

To have the exact covariance matrix of the colored noise, the exact σ_u^2 and η should be known. Practically, the two parameters should be estimated and thus there are uncertainties. According to what is mentioned in [17] we assume the estimated parameters $\hat{\sigma}_u^2, \hat{\eta}$ are the multiples of the actual values $\sigma_u^2, \eta,$ i.e. $\hat{\eta} = \beta\eta$ و $\hat{\sigma}_u^2 = \alpha\sigma_u^2$. So, we say there are noise uncertainties for signal detection in colored Gaussian noise [19].

2.2 Anderson-darling test

The GoF tests quantify a distance between the distribution functions of two sample sets. The transmitted signal assumption is not needed in these tests at all. Anderson-Darling (AD) test is a popular GoF test in statistics that has been applied to SS in [11] and [12]. It will be discussed in the following.

AD test is an extension of the Cramer-von Mises (CM) test, so we will have a short description about the CM statistic.

The CM statistic W^2 is defined by [11],

$$W^2 \triangleq N \int_{-\infty}^{+\infty} \left(F_Y(y) - F_0(y)\right)^2 dF_0(y). \qquad (1)$$

It is obvious there is an important problem in CM statistic which is assigning sufficient weights to the sequences of the distribution $F_0(y)$. Anderson and Darling [20] improved the CM statistic by introducing a weighted statistic as follows:

$$A_c^2 \triangleq N \int_{-\infty}^{+\infty} \left(F_Y(y) - F_0(y)\right)^2 \phi\left(F_0(y)\right) dF_0(y), \qquad (2)$$

where $\phi(t)$ is a nonnegative weight function defined over $0 \leq t \leq 1$. A common weighting function for AD statistic is

$$\phi(\mathrm{t}) = \frac{1}{\mathrm{t}(1-\mathrm{t})} \qquad (3)$$

Finally, in the AD test, if t_0 is a threshold or critical point to be selected, the null hypothesis H_0 is rejected if and only if $A_c^2 > t_0$ [11]. Thus, the probability of false alarms under H_0 is:

$$Pr\{A_c^2 \geq t_0 | H_0\} \qquad (4)$$

Now, according to this description, it is clear that the two phases of the Anderson-Darling test are as follows:

1. Calculate AD test statistic using equation (2).
2. Determine t_0 (threshold) according to the probability of false alarm or α.

The calculation of Eq. (2) is not a simple task, so it is not hard to show by breaking the whole integral in (2) into n parts as follows [11]:

$$A_c^2 = -\frac{\sum_{i=1}^{n}(2i-1)(ln z_i + ln(1 - z_{n+1-i}))}{n} - n \qquad (5)$$

where:

$$z_i = F_0(y_i) \qquad (6)$$

As it is observed in Eq. (6), z_i or $F_0(y_i)$ is the cumulative distribution function of noise. It can be shown that the distribution function depends on the variance of the noise. Hence, uncertainty in the noise variance will strongly influence its performance [11].

To overcome this weakness, Blind AD method is proposed which can overcome the weakness of the AD test. In Blind AD method, first of all, one divisor of n (which is the number of samples) denoted by m is chosen. Then samples are divided into $l = \frac{n}{m}$ groups, each containing m samples. So, we will have:

$$\bar{Y}_j \triangleq \sum_{k=0}^{m-1} \frac{Y_{mj-k}}{m} \quad S_j^2 \triangleq \sum_{k=0}^{m-1} \frac{\left(Y_{mj-k} - \bar{Y}_j\right)^2}{m-1} \qquad (7)$$
$$j = 1, \ldots, l$$

In these equations, $\bar{Y}_j$ and S_j^2 are the mean and variance of the samples in jth group, respectively. To remove the uncertainty effects of noise variances in sensing, a key equation is suggested as follows:

$$X_j \triangleq \frac{\bar{Y}_j}{\frac{S_j}{\sqrt{m}}} \quad j = 1,2, \ldots, l. \qquad (8)$$

It can be indicated that the primary user do not send any signal and the received samples only contain noise. X_j is independent of the noise variance which concludes $F_{0,m}(y)$, the cumulative distribution of m^{th} group, is independent of noise variance as well.

Some works have been done with two-sampled GoF tests for modifying them to SS in [12], [11]. But there is a weakness in two sample tests which is the need for prior samples from channel noise. However, this is the first attempt in this paper to use one-sampled GoF tests in CR networks. Superiority of one-sample tests is that they don't need any prior information or sample about the channel noise. In contrast, in two-sample tests, having at least one noise sample is necessary as a prior sample. This requirement is hard to meet in some busy channels where the assessment of empty spectrum is difficult. Following section will introduce three one-sample tests and then modify them for SS.

3. Presentation of Considered GoF Tests

In this section we study three GoF tests, that is, JB, SW and SF. Afterwards, we study their potential for use in SS.

3.1 Jarque-bera test

The first considered test is JB. This test is a one-sample GoF technique for measurement of deviation from Gaussianity and is constructed from the sample kurtosis and skewness. The test is entitled JB due to its pioneers Carlos M. Jarque and Anil K. Bera. Its test statistic is given by [10]

$$JB \overset{\text{def}}{=} \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right), \qquad (9)$$

where N is the number of samples, K denotes the sample kurtosis and S is the skewness of observation samples, defined as:

$$S \overset{\text{def}}{=} \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\hat{\mu}_3}{(\hat{\sigma}^2)^{\frac{3}{2}}} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2\right)^{3/2}} \qquad (10)$$

$$K \overset{\text{def}}{=} \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\hat{\mu}_4}{(\hat{\sigma}^2)^2} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2\right)^2} \tag{11}$$

in which $\hat{\mu}_3$ and $\hat{\mu}_4$ are the estimates of the third and fourth central moments, respectively; $y_i, i = 1, \ldots, N$ are the received samples; $\bar{y}$ is the sample mean and $\hat{\sigma}^2$ denotes the variance estimation [9].

If we have JB > t_0, the null hypothesis is rejected. In contrast, the null hypothesis will be accepted if JB < t_0. The threshold values are computed using the critical values listed in [10].

3.2 Shapiro-wilk test

The S-W test relies on the correlation between "order statistics" of observed samples and a Gaussian distribution. The order statistics is used to represent that the data sample has to be classified, in vector form sorted in an increasing order as $y' = (y_1, \ldots, y_N)$; where the prime $'$ denotes the transpose of a vector. The SW test statistic W is defined as

$$\mathrm{W} = \frac{\left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{a_i y_i}\right)^2}{\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y_i} - \bar{\mathrm{y}})^2} \tag{12}$$

where $\bar{y}$ is the sample mean. W can be defined as a ratio of two estimates of the sample variance, with the estimate in the numerator holding only if the sample is obtained from a Gaussian distribution whereby coefficients a_i are calculated by linear regression to the expected values of standard Gaussian order statistics. Notably, the expected value of W converges to zero when the input signal becomes non-Gaussian and the expected value of W converges to one for Gaussian input signal as the sample size grows [21].

Vector of coefficients $\boldsymbol{a}' = (a_1, \ldots, a_N)$ is normalized so $\boldsymbol{a}.\boldsymbol{a}' = 1$, and in the way that they become symmetrically mirrored, that is, $a_N = -a_1$, $a_2 = -a_{N-1}$, and so on. These coefficients exist in known statistical literatures. However, these tables are limited and the main problem is to get all of the coefficients for all N. Thus, for simplicity, the coefficients are approximated with the following polynomials [21]:

$$a_N = -2.706056u^5 + 4.434685u^4 - 2.071190u^3 - 0.147981u^2 + 0.221157u + c_N \tag{13}$$

$$a_{N-1} = -3.582633u^5 + 5.682633u^4 - 1.752461u^3 - 0.293762u^2 + 0.042981u + c_{N-1} \tag{14}$$

$$a_i = \epsilon^{-1/2}\tilde{m}_i \tag{15}$$

in these equations $u = N^{-\frac{1}{2}}$, and Eq. (15) for $i = 3, \ldots, N-2$ is established.

In (15), $\boldsymbol{\tilde{m}}' = (\tilde{\mathrm{m}}_1, \ldots, \tilde{\mathrm{m}}_\mathrm{N})$ denotes a vector of expected values of order statistics of the standard Gaussian random variables which can be approximated by the following expression:

$$\tilde{m}_i = \phi^{-1}\{(i - 3/8)/(N + 1/4)\} \tag{16}$$

in the above equation, ϕ^{-1} is the inverse standard Gaussian distribution function. Also in Eq. (15)

$$\epsilon = \frac{\boldsymbol{\tilde{m}}'.\boldsymbol{\tilde{m}} - 2\tilde{m}_N^2 - 2\tilde{m}_{N-1}^2}{1 - 2a_N^2 - 2a_{N-1}^2} \tag{17}$$

Finally, the c_i values in Eq. (13) and Eq. (14) are determined from vector $\boldsymbol{c}' = (c_1, \ldots, c_N)$ by

$$\boldsymbol{c} = \boldsymbol{\tilde{m}}/(\boldsymbol{\tilde{m}}'.\boldsymbol{\tilde{m}})^{1/2} \tag{18}$$

Thus, we can calculate according to the available observations and using the above values in Eq. (12) [21].

3.3 Shapiro-francia

Another test based on the correlation of samples is the so-called Shapiro-Francia (SF) [22]. In fact, this test is a modification to the SW test. Assume that the weights for this test are defined as:

$$\boldsymbol{b}' = \frac{\boldsymbol{c}'}{(\boldsymbol{c}'\boldsymbol{c})^{1/2}}. \tag{19}$$

Then, statistic W' can be represented as follows,

$$W' = \frac{(\boldsymbol{b}'\boldsymbol{y})^2}{\sum(y_i - \bar{y})^2} = \frac{(\sum b_i y_i)^2}{\sum(y_i - \bar{y})^2} \tag{20}$$

Note that **c** is the vector of expected values of the N order statistics of the standard Gaussian distribution and $\boldsymbol{y}$ is samples' vector [22]. The elements of the vector $\boldsymbol{c}$ are defined as equation (16) [22].

4. Proposed Spectrum Sensing Methods

Here we apply the presented tests in Section 3 for SS in this section.

In spectrum holes observed samples are obtained independently from the noise distribution. So if we assume that the distribution of channel noise is Gaussian, introduced tests can be used for SS; because, these are Gaussianity tests.

Assume that N received samples $\boldsymbol{y} = \{y_i\}, i = 1, \ldots, N$ are generated from a PU in a CR network and the channel is AWGN. When channel is empty and there is no primary signal transmission, $\boldsymbol{y} = \boldsymbol{N_g}$, where $\boldsymbol{N_g}$ is the noise vector. Without loss of generality, we use real part of received samples. Thus, these series are obtained from a real Gaussian distribution with unknown mean and variance. This is the null hypothesis. On the other hand, in transmission of PU signal we have $\boldsymbol{y} = \boldsymbol{HS} + \boldsymbol{N_g}$, where $\boldsymbol{S}$ shows the PU signal and H represents the channel gain between the PU and CR. When PU signal is present, alternative hypothesis is formed. In absence of PU, received signal is Gaussian, because of the AWGN channel assumption. Thus, we gather a Gaussian data sample.

Notice that this method does not need prior information about PU, and the noise uncertainty does not affect its performance.

The SS problem is to accept or reject the alternative hypothesis in favor of the null hypothesis as follows:

H_1 : $\boldsymbol{y}$ has non-Gaussian distribution=PU is present

H_0 : $\boldsymbol{y}$ has Gaussian distribution=PU is absent (noise only)

Therefore, according to our assumption SS changes based on distribution testing. Thus, we can apply the methods of section 3 to SS by testing the distribution of received signal.

4.1 OFDM signal

OFDM based signals have a good performance on various channels without the need to use sophisticated receivers and are ideal for use in broadband wireless channels. This is the reason for using this modulation in many systems such as WLAN.

In OFDM systems we have:

$$x(t) = \sqrt{\frac{P}{N_p}}\sum_{k}\sum_{n=0}^{N_p-1} c_{n,k} \cdot e^{2i\pi(f_0+n.\Delta f)t} \cdot g(t-kT_s) \qquad (21)$$

In this equation $x(t)$ is a signal with multicarrier modulation where $\{c_{n,k}\}$ is a series of symbols that assume to be i.i.d. N_p, the number of carriers, Δf the frequency offset between carriers, g(t) the pulse function and P is the signal power and T_s is also a OFDM symbol time. Using the central limit theory [23] and according to the above equation, it can be said that the distribution of OFDM will converge to Gaussian.

When the primary user signal is of a Gaussian type, SS will encounter detection problem, because it will detect and confirm the hypothesis H_0 in the presence of Gaussian OFDM signals and as a result, it will not detect the presence of primary users.

To fix this problem we use FFT operation because OFDM signal in the frequency domain has non-Gaussian properties [24].

After receiving signal, it is transformed to frequency domain. This alters Gaussian properties. It should be noted that the Gaussian noise signal has a Gaussian distribution in frequency domain and it does not lose Gaussianity properties in frequency domain. Proposed GoF tests can be applied to output of FFT block.

4.2 Detection algorithms

We use two different scenarios for evaluating presented tests. In the first scenario for JB, SF and SW tests, the test statistic is calculated and then compared with a threshold. In GoF tests thresholds are determined by using Critical Value (CV) tables. CVs depend on false alarm probability and number of received samples. Table 1 includes CVs for JB test in different situations. If $\text{Test Statistic} > \text{CV}$, the absence of PU is decided. In the presence of PU we have $\text{Test Statistic} < \text{CV}$.

Table 1: Critical values for different α values and sample size belong to JB test [25].

Sample size (n)	Significance level (α)		
	0.01	0.05	0.1
100	12.282	5.418	3.680
10	4.821	2.329	1.478

It's should be mentioned that

SS algorithm for JB, SF and SW test can be summarized as follows:

Step1 Transform received signals into Fourier domain using FFT.

Step2 Compute the test statistic.

Step2 reject H_0 if $\text{Test Statistic} < \text{CV}$. In contrast, accepting H_0 and reject H_1 if $\text{Test Statistic} > \text{CV}$.

The proposed sensing diagram is shown in 0.

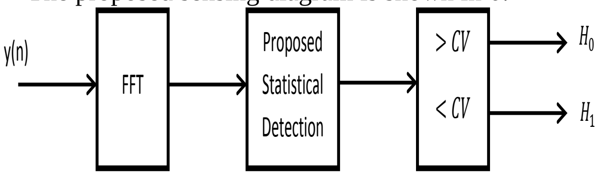


Fig. 1. The proposed spectrum sensing system diagram.

Knowledge of the distribution function of test statistic makes it possible to compare p_value with α value directly instead of comparing test statistic with critical values. In statistics, the p-value of a test is defined as the tail integral of the particular instance of the test statistic over the density of the test statistic which is a random variable itself [14]. Assume that a goodness-of-fit test exists with a test statistic T. Let $Z_T(\tau)$ be the cumulative distribution of T under the null hypothesis. p_value of the test statistic is obtained as follows:

$$p\_value \ = P(T > \tau | H_0) = 1 - Z_T(\tau) \qquad (22)$$

Also, it is obvious

$$\begin{aligned} P_{fa} &= P(T > \tau' | H_0) = 1 - P(T < \tau' \,| H_0) \\ &= 1 - CDF_{(T|H_0)}(\tau') \end{aligned} \qquad (23)$$

According to Eq. (23) and Eq. (24) we will have

$$\text{p} - \text{value} \underset{\text{H}_0}{\overset{\text{H}_1}{\lessgtr}} \text{P}_{\text{fa}} \qquad (24)$$

The *p*-value acts as an indicator of the confidence of the decision reached by the goodness-of-fit test. A low *p*-value (*p*<0.1) shows a high uncertainty about rejecting the hypothesis while a high *p*-value indicates that we are highly confident in rejecting the null hypothesis (*p*<0.9) [26].

In this case, there is no need to obtain critical values from a significance probability of false alarm and it is enough to directly compare the p-value with False alarm probability. So, in this situation, we do not need critical value tables.

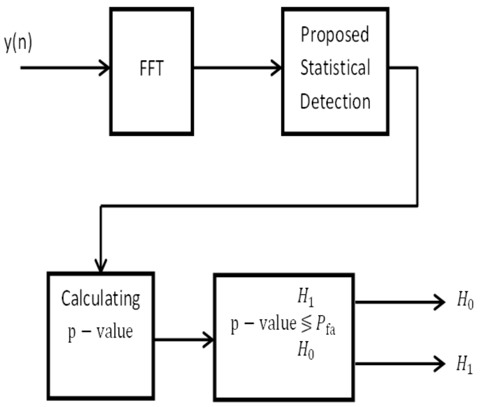

y(n)
FFT
Proposed Statistical Detection
Calculating p − value
$p-\text{value} \underset{H_0}{\overset{H_1}{\lessgtr}} P_{fa}$
H_0
H_1

Fig. 2. The proposed spectrum sensing system diagram for *p*-value.

In the second scenario, SF and SW tests can use p-value in detection procedure; because distribution of them is known and derived in [27], [28]. Therefore, another way to implement SF and SW tests for SS is using p-value. The detection procedure is summarized as follows:

1. Transform the received signals into Fourier domain using FFT.
2. Order the observation samples in an increasing order for calculation of W or W' using Eq.(20) or Eq. (12).
3. Compute Z_T and $p - value$ using the test statistic distribution.
4. Choose an appropriate significance level α.
5. Compare $p - value$ with α. If $p - value < \alpha$, the null hypothesis will be rejected and H_1 will be confirmed.

The proposed detection diagram changes as Fig.2 for SF and SW tests.

4.3 Computational complexity analysis

Here, the computational complexity of the proposed sensing algorithms and AD method as a conventional method are discussed.

Table 2 lists the order of required execution time versus the number of samples for various mathematical operations. Here, to estimate the computational complexity of a test, the highest order of the time complexity is considered.

Table 2: order of computational complexity for different operations [29].

Operations	Complexity Order
Multiplication & Division	$O(N^2)$
Summation & Subtraction	$O(N)$
Square	$O(N^2)$
Natural Logarithm	$O(N^2\log(N))$
FFT	$O(N\log(N))$

In the AD test, the natural logarithm is used which is more complex than the proposed methods. The proposed methods have just the four main mathematic operators in which multiplication has the maximum complexity. So, the maximum complexity order for the proposed methods is equal to $O(N^2)$ because of multiplication complexity. However, the highest order of complexity for AD test is belonging to the natural logarithm which is equal to $O(N^2log(N))$.

In the proposed methods, an FFT block operation is used for the detection of the OFDM signals. The order of computational complexity for FFT is $O(Nlog(N))$ and is less than complexity order of multiplication. Therefore, there is no change in the maximum complexity order of the proposed methods since the maximum complexity order is still $O(N^2)$ which is lower than the AD again.

Due to the same complexity order of the proposed algorithms, other methods are also needed to compare their computational complexity. For instance, calculating the number of mathematic operators is another method of measuring the computational complexity of an algorithm in which the operator with the highest complexity is counted. In this method, the maximum complexity order of the proposed methods is related to multiplication (or division).

It is assumed that the number of operators which are not dependent on the number of samples is negligible; since the number of operators increases with the rise of samples numbers.

Table 3 shows the number of multiplication (or division) of the three methods.

Table 3: number of operators for proposed methods.

Proposed method	Operator numbers
JB	7 × N
SW	5 × N
SF	4 × N

In Table 3, N represents the number of received samples. It is shown that the computational complexity of SF test is less than JB and SW tests. Therefore, SS can be done faster.

We need to do interpolation or extrapolation to find some critical values which do not exist in ready tables. Interpolation or extrapolation increases computational complexity of methods. P-value solves the problem and decreases the computational complexity of SF and SW methods.

5. Simulation Results

In this section, simulation results are demonstrated for different scenarios. Our main goal is to compare the performance of the proposed GoF–based sensing methods with each other and with a conventional approach (i.e. AD).

Before describing the achieved results, it should be noted that WLAN signal is used as the reference signal in this study. The applied WLAN signal in this article is simulated based on available standards [30]. Obviously, WLAN uses OFDM modulation which is one of the most deployed methods in wireless communication. Taking into account that WLAN signal is based on OFDM, we can generalize our results to other OFDM based signals such as WiMAX and D-VBT.

In implementing the offered methods, an FFT block is used before OFDM detection like OFDM demodulator. It is assumed that N samples are collected from environment by CRs, written as y_i for $i = 1, \cdots, N$ which are complex valued. When we calculate proposed tests, without loss of generally, we use real-parts of samples.

Fig. 3 depicts the detection probability of five SS methods including Blind AD, SF, JB, SW and AD. The false alarm probability equals $5\% \ (P_{fa} = 5\%)$ and the environment noise is considered to be Gaussian. The number of available samples from the received signals equals N=4000, which is equivalent to 50 OFDM symbols. As shown in the Fig. 3, JB and the SF methods have almost similar performance outperforming the other methods.

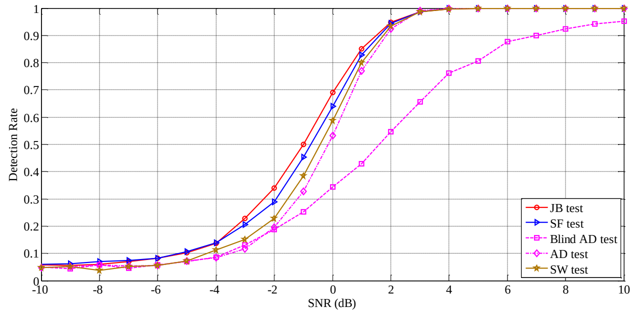


Fig. 3. Detection rate versus SNR value for the simulated WLAN signal (N=4000) over AWGN channels (P_{fa}=0.05).

The computational complexity of the SF method is much less than AD. Also JB method has high complexity order because of calculating high order statistics. On the other hand, statistic calculations are done only twice for the SF method with less complexity. In addition, JB method needs extrapolation and interpolation. Thus finding CV adds more complexity.

In results of Fig. 4 the same parameters are also taken into account. The only difference is that in this situation, the signal has undergone Rayleigh fading. This figure also supports the fact that the offered SF method has much better performance in comparison with the other methods.

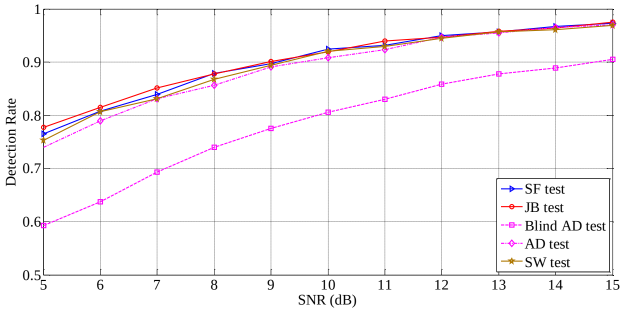


Fig. 4. Detection rate versus SNR for simulated WLAN signal (N=4000) over frequency-flat Rayleigh fading channels (P_{fa}=0.05).

According to Fig. 5, when the number of received samples reduces to N=800 (i.e. 10 OFDM symbols), the performance of all sensing methods are decreased, while the SF method works better than the others in both low and high SNR values. This observation shows the superiority of the SF sensing method over JB-based method, since the JB method requires a relatively large number of received samples for providing good detection performance. Since the CRs should detect PUs as soon as possible, the SF method would be preferred over JB-based SS.

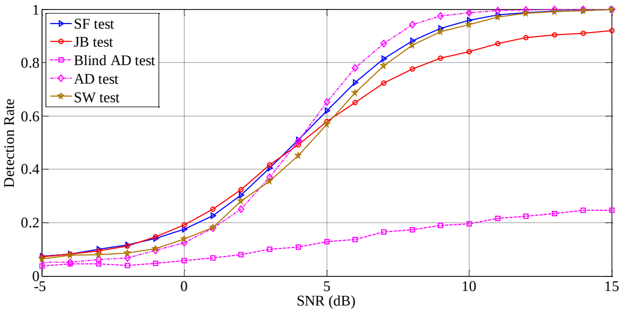


Fig. 5. Detection rate versus SNR value for simulated WLAN signal (N=800) over AWGN channels (P_{fa}=0.05).

Fig. 6 shows the detection performances over frequency-flat Rayleigh fading channels for N=800. As we can see, the Blind AD performance is very poor, while the mixed method performs better than the other methods and has a close performance comparing AD. Considering the acceptable performance of SF method, and the less complexity of SF, it has the best performance due to its less complexity and high detection probability.

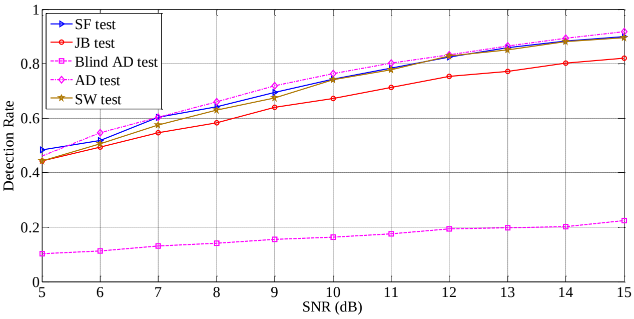


Fig. 6. Detection rate versus SNR value for the simulated WLAN signal (N=800) over Rayleigh fading channels (P_{fa}=0.05).

In Fig. 7, the performance of the proposed methods versus false alarm probability is shown. The ROC of the proposed methods for sample size 3200 (40 OFDM symbol) in WLAN signal is compared with AD method showing improvement in AD algorithm.

Considering Fig. 3 to Fig. 7, SF test has the best performance, so, we chose SF detector to compare with ED.

Fig. 7 shows the detection performance of SF test in comparison to Energy Detector in presence of uncertainty for $P_{fa} = 5\%$ in colored WGN.

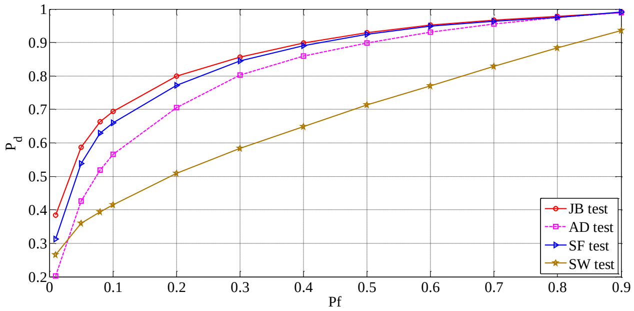


Fig. 7. ROC figure in WLAN for 3200 sample size (40 OFDM symble) over frequency-flat Rayleigh fading channels with SNR = 0 dB.

Fig.8 demonstrates that correlation between noise sample does not affect performance of proposed GoF methods; because there is no any assumption in proposed methods about independent of noise samples. Results of Fig. 8 are verified by our assumptions about the lack of need for independent signal samples.

Usually uncertainty in noise variance exists and ranges about 1 to 2 dB [5]. As it is evident, the SF test works better than the ED against noise uncertainty. This figure demonstrates that correlation between noise sample does not affect performance of proposed GoF methods and lack of need for independent samples in our assumptions verifies results of Fig. 8.

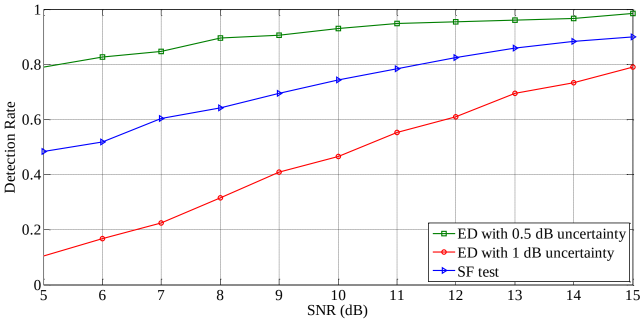


Fig. 8. Signal detection probability versus SNR for WLAN (N=800) Additive colored Gaussian noise in the channel and over Rayleigh fading.

6. Conclusion

In this paper, three one-sample GoF techniques are introduced and modified to detect OFDM-based primary signals. The methods are compared through simulations with each other and a conventional approach (i.e. AD). We showed that the proposed methods' computational complexity is much less than the AD approach. Moreover, Monte-Carlo simulation results for OFDM-based primary signals demonstrate that the SF method outperforms the other techniques in terms of probability of detection. In particular, we have shown that the SF method performs well even in very short sensing durations, in comparison with other GoF-based sensing methods. Furthermore, simulation results show that the SF method outperforms the classical ED in presence of noise uncertainty over different level of SNR. Besides, it is not sensitive to the noise uncertainty which is favorable. Therefore, it can be concluded that the SF method is an appropriate representative of GoF tests in order to be applied in CR networks.

References

[1] M. McHenry. (2005, Aug.). NSF spectrum occupancy measurements project summary. Shared Spectrum Co., Tech. Rep. [Online]. Available: http://www.sharedspectrum.com/

[2] S. Haykin, “Cognitive radio: Brain-empowered wireless communications,”IEEE J. Select. Areas Commun., vol. 23, pp. 201–220, Feb. 2005.

[3] J. Ma, G. Li, B.H. Juang, “Signal processing in cognitive radio,” Proceedings of the IEEE vol. 97, no. 5, pp. 805–823, 2009.

[4] F. Digham, M.-S. Alouini, and M. K. Simon, “On the energy detection of unknown signals over fading channels,” IEEE Trans. Commun., vol. 55, no. 1, pp. 21-24, 2007.

[5] A. Sahai and D. Cabric, “Spectrum sensing: fundamental limits and practical challenges,” in2005 IEEE Int. Symp. New Frontiers DySPAN.

[6] M. Oner and F. K. Jondral, “On the extraction of the channel allocation information in spectrum pooling systems,” IEEE J. Select. Areas Commun., vol. 25, no. 3, pp. 558–565, Apr. 2007.

[7] A. Sonnenschein and P. M. Fishman, “Radiometric detection of spread spectrum signals in noise of uncertainty power,” IEEE Trans. Aerospace Electron. Syst., vol. 28, no. 3, pp. 654–660, July 1992.

[8] Y. H. Zeng and Y.-C. Liang, “Eigenvalue based spectrum sensing algorithms for cognitive radio,” IEEE Trans. Commun., vol. 57, no. 6, pp. 1784–1793, June 2009.

[9] L. Lu and H.-C. Wu, “A novel robust detection algorithm for spectrum sensing,” IEEE J. Select. Areas Commun., vol. 29, no. 2, pp. 305-315, Feb 2011.

[10] Jarque, C. M., and A. K. Bera. "A test for normality of observations and regression residuals." International Statistical Review. Vol. 55, No. 2, pp. 163–172, 1987.

[11] H. Wang, E. Yang, Z. Zhao, and W. Zhang, “Spectrum sensing in cognitive radio using goodness of fit testing,” IEEE Trans. Wireless Comm., vol. 8, pp. 5427–5430, 2009.

[12] G. Zhang, X. Wang, Y.C. Liang, J. Liu, “Fast and Robust Spectrum Sensing via Kolmogorov-Smirnov Test,” IEEE Trans. Commu., vol. 58, no. 12, 2010.

[13] L. Shen, H. Wang, W. Zhang, and Z. Zhao, “Blind spectrum sensing for cognitive radio channels with noise uncertainty,” IEEE Trans. Wireless Commun., vol. 10, no. 6 pp. 1-4, 2011.

[14] N. Kundargi, Y. Liu, A. Tewfik, “A Framework for Inference Using Goodness of Fit Tests Based on Ensemble of Phi-Divergences,” IEEE Trans on Signal processing, vol. 61, no. 4, Feb, 2013.

[15] G. Saporta, Probabilit’es, analyses des donn’ees et statistique, TECHNIP, 1990.

[16] A. N. Mody, Spectrum Sensing of the DTV in the Vicinity of the Video Carrier Using Higher Order Statistics,” IEEE Std. 802.22-07/0359r0, July 2007.

[17] R. Tandra and A. Sahai, “SNR walls for signal detection,” IEEE J. Sel. Areas Commun., vol. 2, no. 1, pp. 4–17, Feb. 2008.

[18] S. M. Kay, Fundamentals of Statistical Signal Processing: Detection Theory, Vol. 2, 3rd edition. Prentice Hall, 1998.

[19] W. Jun, J. Xiufeng, B. Guangguo,C. Zhiping, H. Jiwei, “Multiple Cumulants Based Spectrum Sensing Methods for Cognitive Radios,” IEEE Trans. Commun. vol. 60, no. 12, 2012.

[20] T. W. Anderson and D. A. Darling, “Asymptotic theory of certain Goodness of Fit criteria based on stochastic processes,” Annals of Mathematical Statistics, vol. 23, no. 2, pp. 193-212, 1952.

[21] B. Güner, M. T. Frankford, J.l. T. Johnson, “A Study of the Shapiro–Wilk Test for the Detection of Pulsed Sinusoidal Radio Frequency Interference,” IEEE Trans Geoscience and Remote Sensing, vol. 47, no. 6, 2009.

[22] R.B. D‘Agostino, M.A. Stephens, “GOODNESS-OF-FIT THECHNIQUES”, New York and Basel, 1986.

[23] G. Saporta “Probabilitbs, analyses des donnees et statistique” ed. TECHNIP, 1990.

[24] A. N. Mody, Spectrum Sensing of the DTV in the Vicinity of the Video Carrier Using Higher Order Statistics,” IEEE Std. 802.22-07/0359r0, July 2007.

[25] T. Thadewald, H. Buning, Jarque–Bera Test and its Competitors for Testing Normality – A Power Comparison, Journal of Applied Statistics, vol. 34, no. 1, pp. 87–105, 2007.

[26] H. Sackrowitz and E. Samuel-Cahn, “P values as random variablesexpected p- values,” Amer. Statistician, pp. 326–331, 1999.

[27] P. Royston, "A Pocket-Calaulator Algorithim for the Shapiro-Francia Test for non-Normality: An Application to Medicine" , Statistics in Medicine, vol. 12, 181-184 ,1993.

[28] J. P. Royston, “An extension of Shapiro and Wilk's W test for normality to large samples,” Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 31, no. 2, pp. 115-124, 1982.

[29] J. Borwein & P. Borwein. Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity. John Wiley 1987.

[30] European Telecommunication Standard, “Digital broadcasting systems for television, sound and data services; framing structure, channel coding, and modulation for digital terrestrial television,” Doc. 300 744, 1997.

Seyed Sadra Kashef was born in Urmia, Iran, in 1989. He received the M.Sc. degree in communication systems engineering from Tarbiat Modares University, Tehran, Iran, in 2011. He is currently pursuing the Ph.D. degree in communication systems engineering at Tarbiat Modares University. He is a student member of IEEE. His research interests are in the areas of statistical signal processing and spectrum sensing in cognitive radio networks.

Paeiz Azmi was born in Tehran-Iran, on April 17, 1974. He received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Sharif University of Technology (SUT), Tehran-Iran, in 1996, 1998, and 2002, respectively. Since September 2002, he has been with the Electrical and Computer Engineering Department of Tarbiat Modares University, Tehran-Iran, where he became a professor on December 2011. From 1999 to 2001, he was with the Advanced Communication Science Research Laboratory, Iran Telecommunication Research Center (ITRC), Tehran, Iran. From 2002 to 2005, he was with the Signal Processing Research Group at ITRC. He is senior member of IEEE now.

Hamed Sadeghi was born in Tehran, Iran, in 1983. He received the M.Sc. degree in communication systems engineering from Tarbiat Modares University, Tehran, Iran, in 2008. He is currently pursuing the Ph.D. degree in communication systems engineering at Tarbiat Modares University. He is a student member of IEEE. His research interests are in the areas of statistical signal processing and its applications, including detection, estimation, data fusion, and spectrum sensing.

An Ultra-Wideband Common Gate LNA With Gm-Boosted And Noise Canceling Techniques

Amin Jamalkhah*
Department of Electrical Engineering, Shahid Bahonor University of Kerman, Kerman, Iran
a.jamalkhah@eng.uk.ac.ir
Ahmad Hakimi
Department of Electrical Engineering, Shahid Bahonor University of Kerman, Kerman, Iran
hakimi@uk.ac.ir

Received: 30/Nov/2013 Accepted: 12/Apr/2014

Abstract

In this paper, an ultra-wideband (UWB) common gate low-noise amplifier (LNA) with g_m-boosted and noise-cancelling techniques is presented. In this scheme we utilize g_m-boosted stage for cancelling the noise of matching device. The bandwidth extension and flat gain are achieved by using of series and shunt peaking techniques. Simulated in .13 um Cmos technology, the proposed LNA achieved 2.38-3.4dB NF and S11 less than -11dB in the 3.1-10.6 GHz band. Maximum power gain (S21) is 11dB and -3dB bandwidth is 1 .25-11.33 GHz. The power consumption of LNA is 5.8mW.

Keywords: G_m-Boosted, Low Noise Amplifier, Noise- Cancelling, Ultra-Wideband, Shunt and Series Peaking.

1. Introduction

In recent years, the demand for ultra-wide band systems has increased. These systems are a new wireless technology, which have the ability to send data over a wide spectrum of frequency bands [1]. The advantages of this technology include high data rate, low power, reduced interference and low-cost, that are critical for broadband wireless communication. An UWB LNA is a first block in an ultra-wide band transceiver and Its performance can affect the overall performance of the transceiver. The UWB LNA must be able to provide several basic requirements, such as broadband input matching, low noise figure (NF), sufficient gain to reduce the noise of the mixer, small die area and low power consumption.

Several techniques have been reported in published literature to design UWB LNAs with high performance. Based on the characteristics of the noise and input matching, the UWB LNA can be divided into two main groups, the common gate (CG) LNA, and the common source (CS) [2], [3]. Although the noise figure of CG LNA is more than CS configuration, but the low input impedance of CG LNA in a wide spectrum of frequency band, makes it attractive for LNA design. Thus, the common gate LNA is suitable to achieve broadband input matching [2]. The noise figure of the CG LNA $(1+\gamma/\alpha)$ [4], depends on the process parameters and device size, and almost remains constant with frequency. Also, the NF and input impedance of the CG LNA have a tight relationship with each other. To avoid tight relationship between the input resistance and the noise figure of the CG LNA, we can use g_m-boosting technique [5]. Also to release this trade-off in the CG LNAs a noise-cancelling technique [6] has been widely used. From a feed-forward path the noise of the CG transistor is reduced, while the input impedance is matched simultaneously.

In this work we will combine these two techniques. The g_m-boosting stage made with a common source amplifier. In order to cancel the noise of CG transistor, instead of using feed-forward path, we utilize g_m-boosting stage to cancel the noise of the input matching transistor. However this technique needs extra PDC budget for g_m-boosting stage, in general the total power consumption is low and NF will improve. For achieving the bandwidth extension and flat gain, we use shunt-series peaking and stagger compensation techniques [7]. In section II the proposed LNA and considerations of it will be described. The simulation results are presented in section III.

2. Proposed UWB CG LNA With g_m Boosted and Noise Cancelling Techniques

2.1 Basic idea

The principle of noise canceling is to generate the noise signals with the opposite phase polarities in different paths and adding them at the output. The conceptual diagram of the proposed technique is shown in Fig.1. This LNA is composed of a transistor M_1, two auxiliary amplifiers (A_1 and A_2), and R_s is source impedance. Thermal noise of the matching device M_1, $I_{n,M1}$, generates a voltage noise at node x. A_1 act as g_m-boosting stage and the generated noise voltage at node x, amplified by a factor of $-A_1$ and applied to node y.

The noise current due to the channel of M_1 (I_1) goes to the output. For canceling the noise of M_1, the voltage noise in node *y* must be converted to current noise with the opposite phase polarity to the polarity of I_1. Therefore

* Corresponding Author

to generate such current, an auxiliary amplifier is required. A_2 act as an auxiliary amplifier and convert the voltage noise in node *y* to current noise (I_2). For canceling the noise of the input transistor M_1, I_1 and I_2 must be equal with the opposite phase polarities. A_1 and A_2 in addition to canceling the noise of the M_1, increase the effective transconductance and the total gain of the LNA will be increased. In summary the combination of A_1 and A_2 provide another current path, which result in reduced noise and increased overall gain. The noise canceling condition is as below

$$A_1A_2R_s = 1 \tag{1}$$

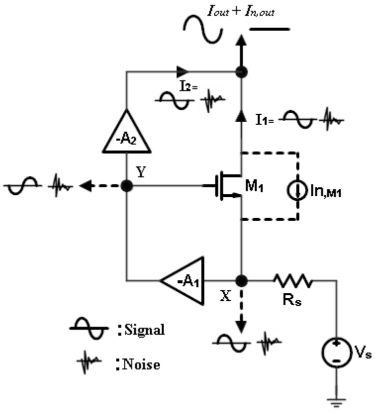


Fig. 1. Conceptual diagram of the proposed LNA with g_m-boosted and noise-canceling techniques

As shown in (1), A_1 and A_2 are inversely related to each other, and the power consumption of LNA will be reduced.

The proposed LNA is shown in Fig. 2. The "main CG amplifier", made with M_1 and R_1, is enhanced by the "A_1" made with M_2 and R_2. M_3 act as A_2. According to voltage noise in node y, We need a CS configuration (M_3) to eliminate the noise of the input matching device. M_4 and M_5 act as a highly linear output voltage buffer. The output buffer is used to drive a 50-Ω load for simulation results. L_0 provides Biasing of M_1 and also resonate with input parasitic capacitance to provide wideband input matching. In order to achieve an UWB LNA, we utilize $L_{1\text{-}3}$ for bandwidth extension and flat gain (shut-series peaking techniques) [7]. All of inductors are on chip with low Q. $C_{0\text{-}4}$ are coupling capacitors.

The noise canceling condition is as below

$$g_{m2}g_{m3}R_2R_s = 1 \tag{2}$$

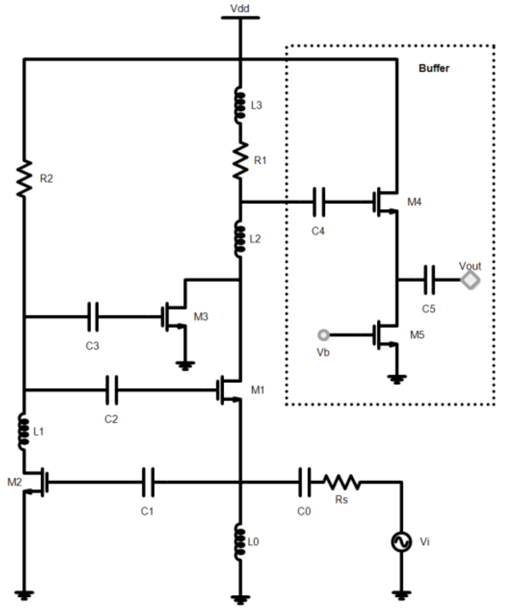


Fig. 2. Complete circuit schematic of the proposed UWB LNA with the output buffer.

2.2 Input matching

G_m-boosted technique employs feed-forward to avoid tight relationship between the input resistance and the noise figure [8]. As shown in Fig.1, for an input voltage change of ΔV, the gate-source voltage changes by -(1+A)ΔV and the drain current by -(1+A)g_m ΔV. Thus the g_m increased by a factor of 1+A1 [8]. By using this technique the input impedance of the proposed UWB LNA is as below

$$Z_{in} = \frac{1}{g_{m1}(1+A_1)} \| sL_0 \| \frac{1}{sC_{in}} \tag{3}$$

Where A_1=$g_{m2}R_2$ and C_{in} is the total parasitic capacitance at the input. By increasing the operating frequency of the circuit, the parasitic input capacitance C_{in} starts playing critical roles, and performance of the LNA at high frequencies reduced. For the UWB LNA, S11<-10dB is needed from 3.1–10.6 GHz. With proper selection of L_0 and C_{in} values, the parasitic capacitance C_{in} absorbed by L_0 and the imaginary part of Z_{in} is negligible within the bandwidth. A low quality factor (Q) for the input matching network suggest a possible broadband impedance match.[9]

2.3 Gain and bandwidth

The effective transconductance of proposed LNA is as below

$$G_{m,eff} = (g_{m1}(1+g_{m2}R_2) + g_{m2}g_{m3}R_2)/2 \tag{4}$$

Under input-impedance-matching and noise canceling conditions, the effective transconductance become $1/R_s$. As frequency increases, the parasitic capacitances at the nodes of the circuit, reduce the gain and bandwidth of the circuit. One way to increase the bandwidth is utilizing shunt and series peaking techniques [7]. In shunt peaking technique an inductor (L_3) connected in series to the load to broaden the bandwidth. In nodes with significant parasitic capacitance, the series peaking technique can be used. In these nodes, an inductor ($L_{1,2}$) is inserted to separate total capacitance into two constituent capacitance. By inserting the inductor, a peaking will be created above the -3db frequency and then by compensating this peaking, larger bandwidth will be achievable. To reduce peaking in the frequency response and in order to broadens the magnitude response, we utilize $L_{1,2}$ with low quality factor [7]. At the resonance frequency of the input matching network (ω_0), the input matching is the best and degrades on the either side, because the quality factor (Q) of the matching network in the input is low. When the load is resistive, for a broadband response there is a significant roll-off in the transconductance gain after ω_0. By using this roll-off, canceling the peaking due to series peaking technique in the transimpedance gain at the output and flatten the overall gain of the amplifier will be achieved. This is done through proper staggering of the resonance frequencies of input matching and series peaking technique [7].

2.4 Noise analysis

The dominant noise source in a common-gate LNA, is channel noise of the input matching transistor, which is equal to ($\gamma/\alpha g_m R_s$), where γ and α are the process dependent parameters. The g_m can be increased to reduce the noise, since the noise and input matching of the CG stage are inextricably related, increasing the g_m degrades the input matching. In order to avoid tight relationship between the noise figure and input matching, we use g_m-boosted and noise canceling techniques together. Also the effective transconductance increased, which result in increased gain and reduced NF. In order to simplify the calculation, it is assumed that the output impedance of transistors is infinite and input bias current is ideal. Furthermore only the thermal noise of the channel transistors and resistors are considered. The noise factor F (NF=10logF) can be derived from

$$F = \frac{V^2{}_{n,out}}{4KTR_S A_V{}^2} = \frac{I^2{}_{n,out}}{4KTR_S G_m{}^2} \tag{5}$$

$I_{n,out}{}^2$ and $V_{n,out}{}^2$ are the output current and voltage noise, $A_v{}^2$ and $G_m{}^2$ are the voltage gain and transconductance under input-impedance-matching and noise canceling conditions. The low frequency noise factor components can be derived as

$$F_{M2} = \frac{\frac{\gamma}{\alpha} g_{m2}(g_{m3}R_2)^2}{R_S G_m{}^2} = \frac{\gamma}{\alpha} \times \frac{1}{R_S g_{m2}} \tag{6}$$

$$F_{M3} = \frac{\frac{\gamma}{\alpha} g_{m3}}{R_S G_m{}^2} = \frac{\gamma}{\alpha} \times \frac{1}{g_{m2} R_2} \tag{7}$$

$$F_{R2} = \frac{R_2 g_{m3}{}^2}{R_S G_m{}^2} = \frac{1}{g_{m2}{}^2 R_2 R_S} \tag{8}$$

$$F_{R2} = \frac{1}{R_1 R_S G_m{}^2} = \frac{R_S}{R_1} \tag{9}$$

The noise due to terms (6-8) are inversely related to g_{m2}, so g_{m2} can be increased to lowering the total NF. By increasing the g_{m2}, to establish input impedance matching and noise canceling conditions, g_{m1} and g_{m3} must be reduced, therefore the power consumption of the $M_{1,3}$ will be reduced. For large g_{m2}, the width of M_2 must be increased. However, this increases the parasitic capacitances in drain and gate of M_2, deteriorating the gain and NF at high frequencies. Also due to the effect of parasitic capacitance in the drain of M_3, the gain decrease, and the NF will be increased at higher frequencies. To mitigate these effects and increasing the bandwidth, series peaking inductors ($L_{1,2}$) insert to these nodes. Fig.4 show the NF and power gain (S21) with and without the inductors L_1 and L_2. Without L_1 and L_2, NF and gain of the LNA degrade at high frequencies.L_1 and L_2, lowering the effects of parasitic capacitances of the circuit at high frequencies, and result in improved NF and gain at high frequencies. Also the effective transconductance $G_{m,eff}$ in the input matching and the noise cancelling conditions, become $1/R_s$, compared to conventional CG LNA is higher and NF will be reduced. However large g_{m2}, requires an extra PDC budget, but the total power consumption of LNA is low.

2.5 Process variations

Inter-die and intra-die variations are two sources of variations in CMOS technologies [10]. Inter-die variation is typically accounted in circuit design as a shift in the mean of some parameter ("e.g.,V_{th} , μ_n ,..") equally across all devices or structures on any one chip. This type of variation can be compensated by proper biasing of the circuit [11]. Intra-die variation (mismatching) is the deviation occurring spatially within any one die that can be solved by layout techniques and proper sizing. In this work we focus on the inter-die variations.

Inter-die variations modeled as worst case process corners. In CMOS manufacturing process there are five process corners (TT, FF, SS, FS, SF). For instance in fast corner, all process variation deviate toward a device with increase current, threshold voltage V_{th} and gate oxide thickness T_{ox} decrease and mobility μ_n increase. Fig .3 shows the constant bias scheme that used in the circuit to reduce the effect of inter-die variations.

Threshold voltage variation has a greater impact on the drain current, thus the effect of V_{th} variations on V_b will be studied. In fast corner V_{th} decrease and result in increase drain current (I_d), voltage drop across R increased and V_b will be reduced. The following expressions are achieved from [10].

$$v_b = V_{dd} - R\ I_d \tag{10}$$

$$I_d = K'(V_b - V_{th})^2 \tag{11}$$

where k'=(1/2)($\mu_n C_{ox}$W/L) is the transistor parameter. By combining of (10) and (11), V_b can be derived as

$$V_b = V_{th} + \frac{\sqrt{2\,K'R\,(V_{dd} - V_{th}) + 1} - 1}{K'R} \tag{12}$$

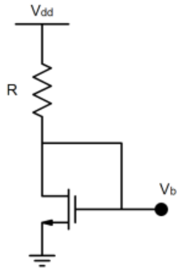


Fig.3. Bias network

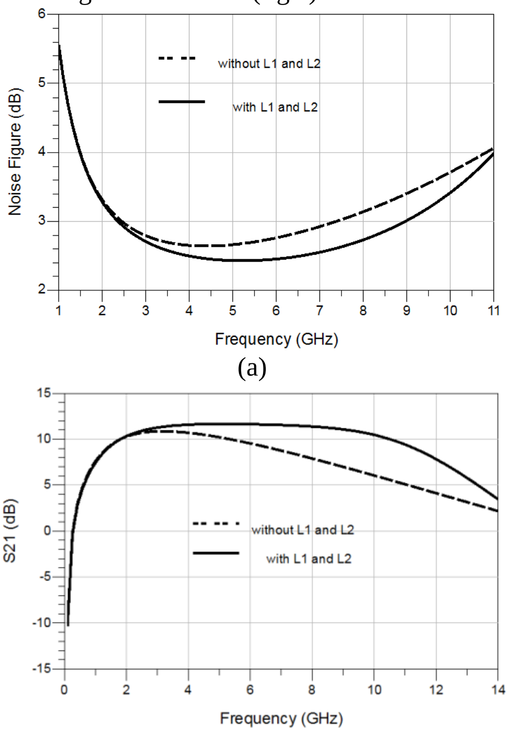
V_{th} variation result in V_b variation as follows

$$\begin{aligned} \delta V_b &= \frac{\partial V_b}{\partial V_{th}} \partial V_{th} \\ &= \delta V_{th} - \frac{\delta V_{th}}{\sqrt{2\,K'R\,(V_{dd} - V_{th}) + 1}} \end{aligned} \tag{13}$$

In (13) by choosing a large R the second term will be negligible and V_b variation will be approximately equal to the V_{th} variation “i.e., the decreases in the V_{th}, will decreases the voltage bias V_b and vice versa.” As mentioned before, in inter-die variation all devices suffer from same variation on the chip. The drain current “I_d=k(V_b-V_{th})^2” of the transistors which are biased with the constant bias circuit (Fig.3), almost remains unchanged against the V_{th} variations. Thus LNA works properly in different process corner.

3. Simulation Results

The LNA is simulated by Advanced Designed System (ADS) in 0.13um CMOS. Figures 5-7, show the S11, power gain (s21), and NF of LNA at different process corners respectively. Since the circuit has only NMOS transistors, only three corners are shown. The S11 is shown in Fig.5 and is less than -11 dB at 3.1-10.6 GHz band in different corners. L_0 and C_{in} resonate with each other to provide wideband input matching. Also we cannot choose large L_1, because L_1 affects on the input matching, thus the value of L_1 is selected so that the input matching of the circuit remains below than -10dB in the bandwidth of the circuit. Fig.6 shows S21 of the LNA at different process corners. In TT, FF, and SS corners, maximum power gain (S21) achieved at 5 GHz, 5.4 GHz and 4.75 GHz respectively which is 11.6dB, 12.5dB, and 10dB respectively. The -3dB bandwidth is 10.46 GHz (1.2 to 11.66 GHz) in TT corner, 11 GHz (1 to 12 GHz) in FF corner, and 9.76 GHz (1.1 to 10.86 GHz) in SS corner. The NF of the LNA is lower than 3.7dB in TT corner, 3.4 dB in FF corner, and 4.5dB in SS corner from 3.1-10.6 GHz and is shown in Fig.7. Minimum NF of LNA in TT, FF, and SS corners respectively is 2.48 dB at 5.44 GHz, 2.31 dB at 5.5 GHz and 2.7 dB at 5.3 GHz. Simulating of IIP3 is done by Two-tone test with 10 MHz spacing, which is shown in Fig.8. At 6GHz, the simulated IIP3 is -13 dBm.1dB compression gain is -25 dBm (fig.9).



(a)

(b)

Fig. 4. NF and Power Gain (S21) with and without inductors L1 and L2

Table 1. COMPARISON WITH UWB CMOS LNAs

Process	-3dB Band [GHz]	NF [dB]	S11* [dB]	S21* [dB]	S12* [dB]	P1dB [dBm]	PDC [mW]	Vdd [V]	Circuit Topology	Reference
0.18µm	3.1-10.6	4.5-6.2	-9.5	13.2	n.a.	-11	28	1.8	CG+ double resonant load	[12]
0.18µm	1.5-11.7	3.74-4.74	-8.6	12.26	-26	-22	10.34	1.8	Current-reuse technique	[13]
0.18µm	.5-11	3.9-4.5	-9	10.2	n.a.	n.a.	14.4	n.a.	Dual-channel shunt technique	[14]
0.18µm	3-10.35	3.3-11.4	-8.3	12.5	n.a.	-14	7.2	1.2	CG+ resonant load	[15]
0.18µm	1.2-11.9	4-5-5.1**	-11	9.7	-35	-16	20	1.8	Broadband noise-cancelling	[16]
0.13µm	**1.25-11.34**	**2.38-3.4****	**-11**	**11**	**-35**	**-25**	**5.8**	**1.2**	**G_m-boosted+noise cancelling techniques**	**This work**

*Maximum values,**In the 3.1-10.6 GHz band.

The power consumption of the LNA is 4.14mW without the output buffer and 5.8mW with the output buffer. Table I shows the comparison of this work with the previous published UWB LNAs.

4. Conclusions

An UWB CG low noise amplifier (LNA) based on G_m-boosted and noise-cancelling techniques, was presented. By using these two techniques low NF achieved. By employing an inductor with low Q in the input of LNA, broadband impedance match achieved. The bandwidth of LNA improved by utilizing series and shunt peaking techniques. By using stagger compensation technique flat gain over the bandwidth achieved. By utilizing the constant bias circuit, LNA worked properly in different process corners. Simulation results for the proposed LNA realized in 0.13um CMOS demonstrate 11-dB maximum power gain, 2.38-dB minimum NF while dissipating 4.85mA from 1.2-V supply.

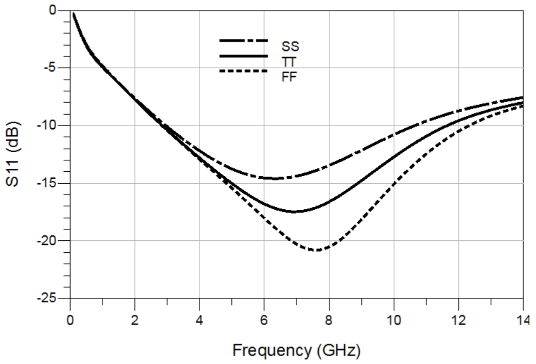


Fig. 5. S11 at different process corners

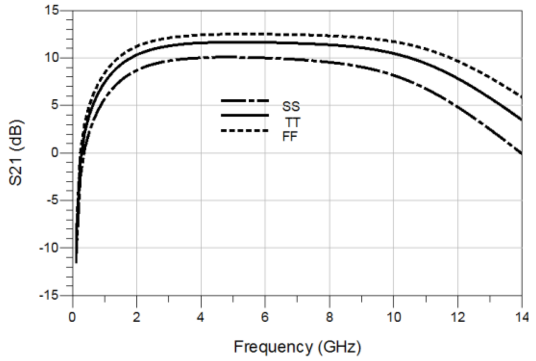


Fig. 6. S21 at different process corners

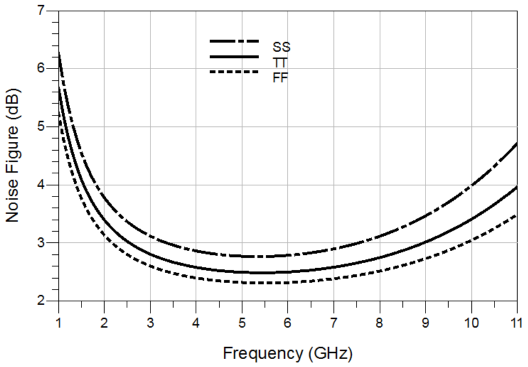


Fig. 7. NF at different process corners

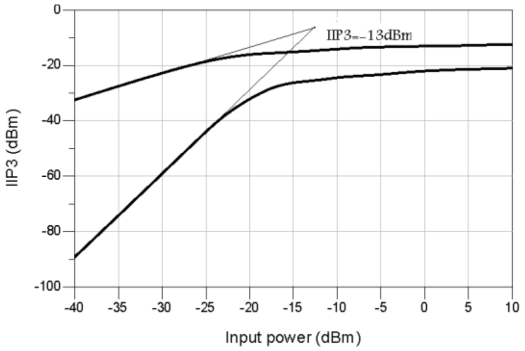


Fig. 8. IIP3 of the LNA at 6 GHz

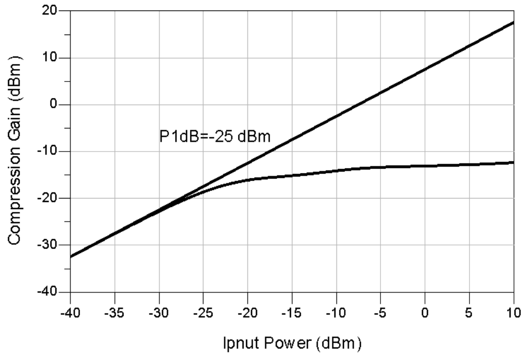


Fig. 9. 1dB Compression point

References

[1] Chang-Ching Wu, Mei-Fen Chou, Wen-Shen Wuen and Kuei-Ann Wen,"A Low Power CMOS Low Noise Amplifier for Ultra-wideband Wireless Applications," IEEE, ISCAS, 2005.

[2] H. Zhang, X. Fan, and S. Sinencio, "A low-power, linearized, ultrawideband LNA design technique," *IEEE J. Solid-State Circuits*, vol.44, no.2, pp.320–330, Feb. 2009.

[3] D. Ponton, P. Palestri, D. Esseni, L. Selmi, M. Tiebout, B. Parvais, D.Siprak, and G. Knoblinger, "Design of ultra-wideband low-noise amplifiers in 45-nm CMOS technology: Comparison between planar bulk and SOI FinFET devices," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol.56, no.5, pp.920–932, May 2009.

[4] D. J. Allstot, X. Li, and S. Shekhar, "Design considerations for CMOS low-noise amplifiers," in *Proc. IEEE Radio Freq. Integr. Circuit Symp*., 2004, pp. 97–100

[5] W. Zhuo, X. Li, S. Shekhar, S. H. K. Embabi, J. P. de Gyvez, D. J. Allstot, and E. Sanchez-Sinencio, "A capacitor cross-coupled common gate low noise amplifier," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol.52, no.12, pp.875–879, Dec. 2005.

[6] F. Bruccoleri,E. A. M. Klumperink, and B. Nauta, "Wide-band CMOS low noise amplifier exploiting thermal noise canceling," *IEEE J. Solid- State Circuits*, vol.39, no.2, pp.275–281, Feb. 2004.

[7] Shekhar, S.; Walling, J.S.; Allstot, D.J. "Bandwidth Extension Techniques for CMOS Amplifiers," *IEEE J. Solid- State Circuits*, vol.41, no.11, pp.2424 - 2439, Nov. 2006.

[8] X, Li. S, Shekhar. and D. J. Allstot. "Gm -BooSted Common·Gate LNAand Differential ColpittS VCO/QVCO in 0, 18-um CMOS." *IEEE J. Solid, State Circuits*. vol.40. PP.2609--2618. DEC.2005.

[9] D. J. Allstot, X. Li, and S. Shekhar, "Design considerations for CMOS low-noise amplifiers," in *Proc. IEEE Radio-Frequency Integrated Circuits Symp*., Jun. 2004, pp.97–100.

[10] Y. Liu and J.S. Yua "CMOS RF Low-Noise Amplifier Design for Variability and Reliability," IEEE Trans on device and material reliability, vol.11, no.3, pp.450-457 ,Sep 2011.

[11] D. Gómez, M. Sroka, and J. L. G. Jiménez, "Process and Temperature Compensation for RF Low-Noise Amplifiers and Mixers," IEEE Trans. Circuits Syst. I, Reg. Papers, vol.57, no.6, pp.1204-1211, june 2010.

[12] B. Park, S. Choi, and S. Hong, "A low-noise amplifier with tunable interference rejection for 3.1- to 10.6-GHz UWB systems," *IEEE Microw. Wireless Compon. Lett*., vol.20, no.1, pp.40–42, Jan. 2010.

[13] Y.-S. Lin, C.-Z. Chen, H.-Y. Yang, C.-C. Chen, J.-H. Lee, and G.-W.Huang, "Analysis and design of a CMOS UWB LNA with dual-RLC branch wideband input matching network*," IEEE Trans. Microw. Theory Tech*., vol.58, no.2, pp.287–296, Feb. 2010.

[14] Q.-T. Lai and J.-F. Mao, "A 0.5–11 GHz CMOS low noise amplifier using dual-channel shunt technique," *IEEE Microw. Wireless Compon. Lett*., vol.20, no.5, pp.280–282, May 2010.

[15] C.-Y.Wu, Y.-K. Lo, and M.-C. Chen, "A 3–10 GHz CMOS UWB low noise amplifier with ESD protection circuits," *IEEE Microw. Wireless Compon. Lett*., vol.19, no.11, pp.737–739, Nov. 2009.

[16] C.-F. Liao and S.-I. Liu, "A broadband noise-canceling CMOS LNA for 3.1–10.6-GHz UWB receivers," *IEEE J. Solid-State Circuits*, vol.42, no.2, pp.329–339, Feb. 2007.

Amin Jamalkhah received the B.Sc. degree in electrical engineering from Lorestan University, Khoramabad, Iran in 2011, and M.Sc. degree in electrical engineering from Shahid Bahonar University of Kerman,Iran in 2014.His general research interests include CMOS RF circuits for wireless communications.

Ahmad Hakimi received the Ph.D. degree from the Istanbul Technical University. His research interests are analog CMOS integrated circuit design, nonlinear circuit theory and applications, RF microelectronics.

Load Balanced Spanning Tree in Metro Ethernet Networks

Ghasem Mirjalily*
Department of Electrical and Computer Engineering, Yazd University, Yazd, Iran
mirjalily@yazd.ac.ir
Samira Samadi
Department of Electrical and Computer Engineering, Yazd University, Yazd, Iran
s.samadi@stu.yazd.ac.ir

Received: 28/Sep/2013 Accepted: 12/Apr/2014

Abstract

Spanning Tree Protocol (STP) is a link management standard that provides loop free paths in Ethernet networks. Deploying STP in metro area networks is inadequate because it does not meet the requirements of these networks. STP blocks redundant links, causing the risk of congestion close to the root. As a result, STP provides poor support for load balancing in metro Ethernet networks. A solution for this problem is using multi-criteria spanning tree by considering criterions related to load balancing over links and switches. In our previous work, an algorithm named Best Spanning Tree (BST) is proposed to find the best spanning tree in a metro Ethernet network. BST is based on the computation of total cost for each possible spanning tree; therefore, it is very time consuming especially when the network is large. In this paper, two heuristic algorithms named Load Balanced Spanning Tree (LBST) and Modified LBST (MLBST) will be proposed to find the near-optimal balanced spanning tree in metro Ethernet networks. The computational complexity of the proposed algorithms is much less than BST algorithm. Furthermore, simulation results show that the spanning tree obtained by proposed algorithms is the same or similar to the spanning tree obtained by BST algorithm.

Keywords: Metro Ethernet Network, Spanning Tree, Load Balancing, Shortest Path Selection.

1. Introduction

Ethernet technology is widely accepted in enterprise deployments and it is said that today more than ninety percent of data traffic is Ethernet encapsulated. Currently, the Ethernet technology is applicable in all levels from local area to metropolitan and wide area network environments [1].

In an Ethernet network, only one active path can exist between two nodes, because multiple active paths create loops in the network. Existence of the loops in the network topology confuses the forwarding and learning algorithms. Current Ethernet networks rely on IEEE STP [2] or IEEE Rapid Spanning Tree Protocol (RSTP) [3]. These are link management protocols that provide path redundancy while preventing undesirable loops in the network.

In both STP and RSTP, all of the traffic will be routed on the same spanning tree. Furthermore there isn’t any mechanism for load balancing. These result in unbalanced load distribution and bottlenecks, especially close to the root and therefore, cause inefficient utilization of resources in metropolitan area networks [4].

Traffic engineering in metro Ethernet networks is a widely researched topic and some improvements have been proposed in the literature in order to solve this problem [5-8].

SmartBridge [5] is a bridged network architecture that addresses the problems associated with spanning trees in Ethernet networks. In SmartBridge architecture, packets will be forwarded along the shortest paths. Although shortest path provides low latency, it does not solve the problem of load balancing in the network. Furthermore, in this architecture, all of the bridges must be SmartBridge compliant.

K. Lui et. al [6] propose an approach named STAR (Spanning Tree Alternate Routing). STAR finds paths that are shorter than their corresponding tree paths; therefore it reduces latency between source and destination pairs. However, STAR is complex and it risks overloading of critical links.

Tree-Based Turn-Prohibition (TBTP) [7] is another approach to load balancing in Ethernet networks. TBTP constructs a spanning tree by blocking some pairs of links around nodes, such that all cycles in the network will be broken. However, TBTP does not consider the best spanning tree and switch load balancing.

Multiple Spanning Tree Protocol (MSTP) [8] is another related standard that is defined in IEEE 802.1s. MSTP improves the load balancing and failure recovery capabilities. However, maintaining multiple spanning trees adds a large complexity to network management and causes high control overheads.

In our previous works, some solutions to the problem of load balancing in metro Ethernet networks have been proposed. In [9], a novel algorithm is introduced to select the Best Spanning Tree (BST) for a given network topology based on the load balancing criterions. BST algorithm finds the best spanning tree by calculating the defined score for each possible tree and by selecting the spanning tree with highest score as the Best Spanning

* Corresponding Author

Tree. The defined score is a function of shortest path criterion and load balancing on links and switches. In summary, BST algorithm is an exhaustive search algorithm that finds the Best (Optimal) spanning tree. This is done by finding all spanning trees of the network, evaluating them based on the defined criteria and then selecting one with greatest score. In [10], a Multiple BST (MBST) approach is proposed that is the application of BST algorithm in a multiple spanning tree scheme. MBST considers all of the possible edge-disjoint spanning trees and all of the possible VLANs grouping and finds the best solution. Note that a set of spanning trees are edge-disjoint if they have not any common edges (links). Using edge-disjoint spanning trees enhances load balancing and resiliency, because in the case of a link failure, only one spanning tree will be failed and the failure has no impact on the spanning trees not including the failed link.

Although, BST and MBST algorithms can find the best answer for small networks, but their complexity is too large in large-scale networks.

In [11], in order to reduce the complexity of BST algorithm, a simple approach that finds the sub-optimal spanning tree in small metro Ethernet networks is introduced. In [12], the algorithm is improved by introducing Shortest Path Selection criterion. The new algorithm is applicable in realistic large-scale metro Ethernet topologies. In this paper, that is an extended version of [12], we introduce Load Balanced Spanning Tree (LBST) algorithm. LBST is a simple algorithm that finds the load balanced spanning tree by using the same criterions used in BST algorithm. The computational complexity of LBST algorithm is much less than BST. LBST simplifies the process of finding spanning tree by using an iterative algorithm with zero initial values for link loads. Although setting zero is the simplest way for specifying the initial values, but it is very far from the real values and this can degrades the performance of the LBST algorithm. In order to improve the performance of the algorithm, a modified version of the algorithm called Modified LBST (MLBST) will be introduced. In MLBST, more accurate estimation of the initial values of link loads will be used.

In fact, LBST and its modified version MLBST find a near-optimal spanning tree for a given metro Ethernet network in a simple manner based on shortest path criterion and load balancing on links and switches. In this way, three major criterions are introduced: load balancing over links, load balancing on switches and shortest path selection. Also, three coefficients α, β and γ corresponding to above criterions are defined. This allows the network managers to weight the importance of each criterion based on their defined goal. These criterions and corresponding coefficients will be used to assign weights to the links and then algorithm finds the shortest path between each node pair by using Dijkstra's algorithm [13]. During this process, the link weights must be updated to reflect the effects of adding new traffic demands at each step.

Although the constructed spanning tree may be not the best, but it is minimum weight and is a good approximation to the BST result.

The rest of the paper is organized as follows: Section 2 introduces some definitions and notations. In section 3, the LBST algorithm is explained in detail. In Section 4, some numerical simulation results are presented. Section 5 proposes a modified version of LBST algorithm, and finally, some conclusions are drawn in section 6.

2. Definitions and Notations

In this paper, a metro Ethernet network is modeled by a graph with N nodes representing switches and a set of M links connecting nodes. Here, for simplicity assume symmetric links and symmetric traffic demands. The traffic demands between nodes are represented by a N-by-N matrix D, that its component $d_{i,j}$ $(i ,j =1, 2, ...N ,i \neq j)$ represents the mean traffic rate between nodes i and j. Also $b_{i,j}$ $(i ,j =1, 2, ...N ,i \neq j)$ represents the bandwidth of the link between nodes i and j and c_j $(i =1, 2,... N)$ represents the switching capacity of ith node.

The goal of this paper is finding the best spanning tree based on three defined criterions. These criterions are links load balancing (LLB), switches load balancing (SLB) and shortest path selection (SPS). Here, the shortest path between two nodes is defined as a path with maximum aggregated bandwidth and minimum hop counts.

In this work, three coefficients α, β and γ, are defined to indicate the importance of each criterion. Note that $0\leq\alpha, \beta, \gamma \leq 1$ and $\alpha+\beta+\gamma=1$. For example, if the main criterion is LLB, assign $\alpha=1, \beta=\gamma=0$, but if the goal is to find the best tree based on LLB and SPS criterions but not SLB, assign $\alpha=0.5, \gamma= 0.5, \beta=0$.

In the proposed algorithm, the link weight is a major component in finding the best spanning tree. Here, the link weight is defined as a linear function of three components: the utilization of the link, the average utilization of switches that the link is between them, and the inverse of normalized bandwidth of the link. Therefore, one can write the link weight between nodes i and j ($W_{i,j}$) as:

$$W_{i,j} = \alpha u_{l_{i,j}} + \beta u_{s_{i,j}} + \gamma u_{b_{i,j}} . \qquad (1)$$

In above Equation,

$$u_{l_{i,j}} = \frac{l_{i,j}}{b_{i,j}} \qquad (2)$$

is the utilization of the link between nodes i and j and $l_{i,j}$ is the traffic flowing through it. Also,

$$u_{s_{i,j}} = \frac{1}{2}\left(u_{s_i} + u_{s_j}\right), \qquad (3)$$

is the average utilization of switches i and j, where:

$$u_{s_i} = \frac{s_i}{c_i} , u_{s_j} = \frac{s_j}{c_j} \quad . \qquad (4)$$

Here s_i and s_j are the traffics flowing through switches i and j respectively. Also,

$$u_{b_{i,j}} = \frac{b_{min}}{b_{i,j}} \quad (5)$$

is the inverse of normalized bandwidth of the link between nodes i and j, where $\mathrm{b_{min}}$ is bandwidth of the link with minimum bandwidth in the network. Note that the $\mathrm{b_{min}}$ must be selected over all active links of the network.

For a network graph with N nodes, each spanning tree has *N-1* links. For each spanning tree, the variance of link utilizations (σ_l^2) can be defined as [9]:

$$\sigma_l^2 = \frac{1}{N-1}\sum_{k=1}^{N-1}\left(\frac{l_k}{b_k} - \bar{l}\right)^2 , \quad (6)$$

where

$$\bar{l} = \frac{1}{N-1}\sum_{k=1}^{N-1}\frac{l_k}{b_k} \quad (7)$$

is the average of link utilizations, $\mathrm{l_k}$ is traffic load on kth link and $\mathrm{b_k}$ denotes the bandwidth of *k*th link. For each spanning tree, σ_l^2 indicates the degree of link load balancing, therefore in LLB criterion, the main goal is to find a spanning tree with minimum σ_l^2.

Similar to (6), for each spanning tree, the variance of switch utilizations (σ_s^2) can be defined as [9]:

$$\sigma_s^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{s_i}{c_i} - \bar{s}\right)^2 , \quad (8)$$

where

$$\bar{s} = \frac{1}{N}\sum_{i=1}^{N}\frac{s_i}{c_i} \quad (9)$$

is the average of switch utilizations, $\mathrm{s_i}$ is traffic load cross *i*th switch and $\mathrm{c_i}$ denotes the switching capacity of *i*th switch. For each spanning tree, σ_s^2 indicates the degree of switch load balancing, therefore in SLB criterion, the goal is to find a spanning tree with minimum σ_s^2.

In SPS criterion, the goal is to find a spanning tree with maximum bandwidth links and minimum hop count paths. In this way, the parameter L is defined as:

$$L = \frac{\sum_{k=1}^{N-1} l_k}{\sum_{k=1}^{N-1} b_k} \ . \quad (10)$$

The goal is to select the spanning tree with minimum L. To clarify the definition of parameter L, note that $\sum_{k=1}^{N-1} l_k = \sum_{i=1}^{N}\sum_{j=1,\, j\neq i}^{N} d_{ij}h_{ij}$ indicates the aggregated load on links. Where h_{ij} is the number of hops from node i to node j and d_{ij} is the traffic demand between these two nodes that is a fixed known value. It is clear that if traffic demands pass on shortest (minimum hop count) paths, $\sum_{k=1}^{N-1} l_k$ will be minimum. On the other hand, $\sum_{k=1}^{N-1} b_k$ indicates the aggregated bandwidth of links. Therefore, minimizing L guarantees each traffic demand travels on shortest path with maximum bandwidth and minimum hop count.

The proposed algorithm also uses a parameter named Minimum Function (MF) to break loops. This parameter is defined for a given spanning tree as:

$$MF = \alpha\sigma_l^2 + \beta\sigma_s^2 + \gamma L \ . \quad (11)$$

In next section, the details of the new algorithm are explained.

3. LBST Algorithm

In previous section, some definitions that will be used here are described. This section describes the proposed algorithm named LBST in detail. This algorithm is simple and easy to implement in comparison to BST that is a computationally complex algorithm.

LBST algorithm first sorts node pairs based on their traffic demand and builds shortest path for each pair sequentially.

Here the highest-demand-first technique will be used because inserting lower traffic demand into the network can be done without loss of much performance [4]. After finding the best path for each node pair, the algorithm loads the traffic on this path and then updates the weights of links on the path. This process will be continued for other node pairs in descending order of their traffic demands.

The LBST algorithm is implemented in two different methods. In first version named LBST I, loops will be broken step by step. The steps of this algorithm are described as follows:

1. Assign the initial value of link weights using Equations (1)-(5). The initial spanning tree is a null graph. Note that in the first step, there is no load on links and switches, therefore $u_{l_{i,j}}$ is zero for all links, but about the switches, algorithm uses the demand matrix to find the initial load on switches. To do this, add the traffic demands that the *i*th switch is their origination and assign the result to calculate the initial value of u_{s_i} by using Equation (4).
2. Sort the node pairs based on the traffic demands in descending order and set *k=1*.
3. Select the *kth* node pair and find the shortest path between them.
4. Check whether or not the links and switches of this path have enough capacity. If yes then load the corresponding traffic demand on path. If not, select next shortest path, then go back to step 4.
5. Concatenate the discovered path to the spanning tree.
6. Update the link loads and switch loads by adding the corresponding traffic demand to the loads of the links and switches located on the path. Then, update the link weights according to the Equations (1)-(5).
7. Check whether is any loop in the constructed tree or not. If not go to step 8, else break the loop:
   a. For selected loop, determine the links which formed the loop.
   b. For each link check if this link is removed, do other links and switches in the loop have enough capacity to handle the excess traffic? If yes, then this link is a candidate, otherwise this link cannot be removed.

   c. If there is not any candidate link, ignore this path and select next shortest path, then go back to step 4. Otherwise, by having the list of candidate links for removal, remove the link which contributes the lowest MF to the tree. In other words, after removing each link in the loop, you get a different tree. Now keep the tree with the lowest Minimum Function defined in Equation (11).
   d. Update parameters and go back to step 7.

8. Set *k=k+1* and go back to step 3. Continue this process for all node pairs.

Note that if during the process, some links or switches are fully used; in the rest, ignore them in selecting the best paths.

As the first issue, breaking loops step by step in LBST I is a time consuming process that decreases the speed of the algorithm considerably. One approach to speed up the algorithm is breaking all the loops at the end of the algorithm instead of breaking them step by step. The new approach is named LBST II algorithm. It is clear that the speed of the LBST II is much more in comparison to the LBST I. The steps of the LBST II algorithm are described as follows:

1. Assign the initial value of link weights using Equations (1)-(5) as described in LBST I algorithm. The initial spanning tree is a null graph.
2. Sort the node pairs based on the traffic demands in descending order and set *k=1*.
3. Select the *kth* node pair and find the shortest path between them.
4. Check whether or not the links and switches of this path have enough capacity. If yes then load the corresponding traffic demand on path. If not, select next shortest path, the go back to step 4.
5. Concatenate the discovered path to the spanning tree.
6. Update the link loads and switch loads by adding the corresponding traffic demand to the loads of the links and switches located on the path. Then, update the link weights according to the Equations (1)-(5).
7. Set *k=k+1* and go back to step 3. Continue this process for all node pairs.
8. Check whether is any loop in the constructed tree or not. If not go to step 9, else break the loop:
   a. For selected loop, determine the links which formed the loop.
   b. For each link in the selected loop, check if this link is removed, do other links and switches in the loop have enough capacity to handle the excess traffic? If yes, then this link is a candidate, otherwise this link cannot be removed.
   c. By having the list of candidate links for removal, remove the link which contributes the lowest MF to the tree. In other words, after removing each link in the loop, you get a different tree. Now keep the tree with the lowest MF. If there is not any candidate link, ignore this sub-step and go to sub-step d. (Note that in very rare situations, the algorithm may be not able to break loops. In these rare cases, we must use LBST I algorithm to find the solution).
   d. Update parameters and go back to step 8.
9. End.

The proposed algorithms are trying to find the shortest possible paths between nodes and in the same time, they are trying to balance the utilization of link bandwidths and switching capacities.

4. Simulation Results

In this section, the performance of the proposed algorithms will be compared with BST algorithm proposed in [9]. The algorithms are implemented in MATLAB. The input parameters for a network with *N* nodes and *M* links are network graph, bandwidth vector *B* which elements are $b_{i,j}$ *(i=1,2,...,N)*, switching capacity vector *C* which elements are c_i *(i=1,2,...,N)*, traffic demands matrix *D* which elements are $d_{i,j}(i=1,2,...,N, d_{i,i} = 0)$, and α , β, γ $(0\leq\alpha,\beta,\gamma\leq 1, \alpha+\beta+\gamma=1)$.

For simulation, a typical popular topology for metro Ethernet networks [14], is considered. A metro Ethernet network usually consists of a core part and several aggregation and access regions. The task of the core part is to forward the traffic load of the aggregation regions toward the edge nodes. The shape of the core is usually one or more rings formed by high speed switches and links. The aggregation part aggregates the traffic of access parts to several internal switches that are connected to the core rings. For aggregation part, usually topologies such as rings or dual homing structures are used. The access parts are usually tree shaped, because the cost of building the interconnections are high.

Figure 1 shows a typical metro Ethernet network. Here, the core consists of four switches with switching capacity of 8 Gbps interconnected to a ring formed by 2 Gbps Ethernet links. Also, two edge nodes with switching capacity of 8 Gbps are connected to core switches with 2 Gbps links and four aggregation nodes with switching capacity of 4 Gbps are connected to two core nodes using dual homing with 1 Gbps links. In this topology, there are eight access nodes with switching capacity of 1 Gbps are connected to aggregation switches with 1Gbps links.

In this typical network, for simplicity consider constant bit rate traffic demands between access nodes and edge nodes. These bidirectional traffic flows are shown in Table 1 where notation “Ac” stands for Access.

As you can see from Figure 1, the physical structure of the access part is tree. Therefore, only the core, aggregation and edge parts of the network are considered in the load balancing algorithm as shown in Figure 2. In this figure, the labels indicated on nodes are switch names; where, the notation “Ed”, “Co”, and “Ag” stands for Edge, Core and Aggregation, respectively.

First consider LLB criterion (α=1, β=γ=0). The spanning trees selected by BST, LBST I and LBST II algorithms are shown in Figures 3.a, 3.b and 3.c, respectively. In this case, the variance of link utilizations (σ_l^2) is 0.0133 for BST tree, 0.0230 for LBST I tree and 0.0179 for LBST II tree. The selected spanning tree by LBST I is ranked third best (3rd) tree by BST algorithm and selected spanning tree by LBST II is ranked second best (2nd) one.

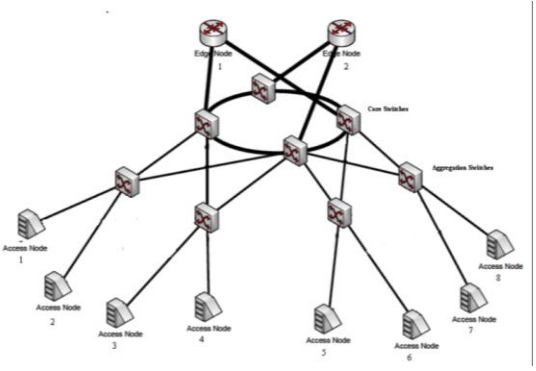


Fig. 1. A typical metro Ethernet network.

Table 1. Traffic demands (Mbps)

	Ac1	Ac2	Ac3	Ac4	Ac5	Ac6	Ac7	Ac8
Edge1	50	50	100	200	200	200	200	100
Edge2	100	100	300	300	200	100	100	200

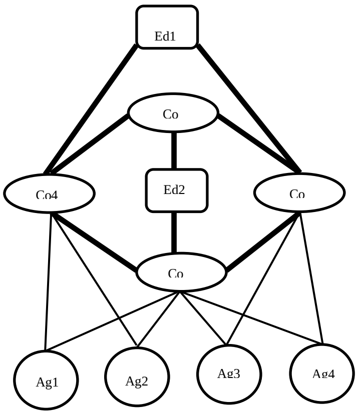
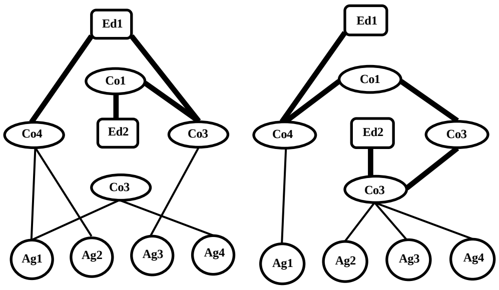
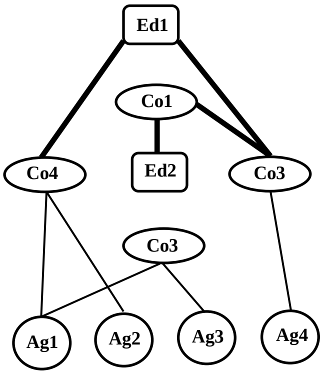


Fig. 2. Metro Ethernet network graph.



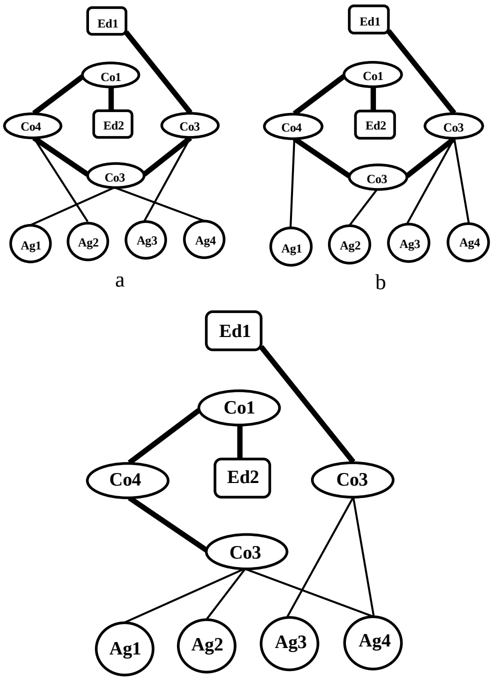
a

b



c

Fig. 3. Spanning tree selected based on LLB criterion by: a) BST algorithm b) LBST I algorithm c) LBST II algorithm.



a

b

c

Fig. 4. Spanning tree selected based on SLB criterion by: a) BST algorithm b) LBST I algorithm c) LBST II algorithm.

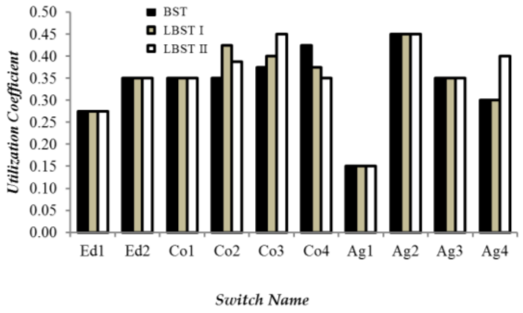


Fig. 5. Utilization coefficient of switches for SLB criterion.

Now consider SLB criterion (*α*=0, *β*=1, *γ*=0). The spanning trees selected by BST, LBST I and LBST II algorithms are shown in Figures 4.a, 4.b and 4.c respectively. In this case, the variance of switch utilizations (σ_s^2) is 0.0063 for BST tree, 0.0066 for LBST I tree and 0.0069 for LBST II tree. The selected spanning tree by LBST I is ranked third best (3rd) tree by BST algorithm and selected spanning tree by LBST II is ranked fourth best (4th) one.

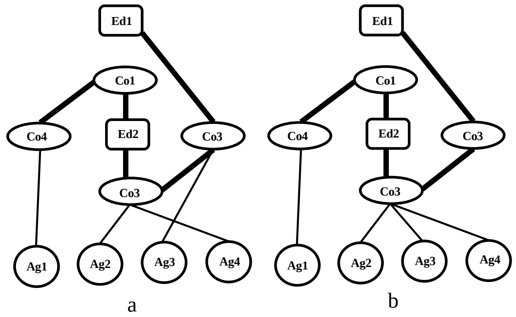


Fig. 6. Spanning tree selected based on SPS criterion by:
a) BST algorithm b) LBST I and LBST II algorithms.

The utilization coefficients of switches for SLB criterion are shown in Figure 5. This Figure shows that the load balancing obtained by using LBST I and LBST II algorithms is very close to the results obtained by using BST algorithm.

In last scenario, consider SPS criterion (*α*=*β*=0, *γ*=1). The selected spanning trees are shown in Figure 6. As a numerical comparison, the trees selected by LBST I and LBST II algorithms are the same and are ranked third best (3rd) tree by BST algorithm with L equal to 0.471 while this value is 0.464 for the best spanning tree selected by BST. These values are very close to each other.

From above simulation results, the following statements can be concluded:

Although LBST algorithms are not the best but their results are the same or similar to the results obtained by using BST algorithm.

The computational complexity of LBST algorithms is much less than BST algorithm. As a roughly comparison, the typically run time of BST algorithm for a network with tens of switches and links on a new high speed computer is tens of minutes, while the run time of our new approaches for the same network is only several seconds.

LBST II algorithm breaks all of the loops in the last step, while the LBST I algorithm breaks loops step by step. Therefore, LBST II is much faster than LBST I. Furthermore, simulation results show that the output of LBST II is the same or similar to the output of LBST I algorithm.

5. Modified LBST Algorithm

As described in section 3, the initial values of link weights in the first step of the LBST algorithms are assigned using Equations (1)-(5). In the beginning, there is no load on links, therefore in simulations done in previous section, the initial value of $u_{l_{i,j}}$ was set to zero for all links. About the switches, algorithm uses the demand matrix to find the initial load on switches. To do this, add the traffic demands that the ith switch is their origination and assign the result to calculate the initial value of u_{s_i} by using Equation (4). Although this is a simple method for specifying the initial values, but it is far from the real values. In this section, we want to study the effects of initial values on the performance of the LBST algorithm. In this way, a different method for calculating the initial values of link loads and switch loads is introduced. This modification enables the algorithms to estimate the initial link weights with more accuracy. For future references, name the new algorithm, Modified LBST (MLBST).

Note that MLBST is useful for LLB and SLB criterions. For SPS criterion, the output of LBST and MLBST algorithms are the same.

In the following, the new algorithm is described, and then by driving some simulations, the effectiveness of using accurate initial values on the performance of the algorithm is showed.

As mentioned in previous section, the output of LBST II is the same or similar to the output of LBST I algorithm, but its computational complexity is less. For this reason, in the rest, only the LBST II algorithm is considered.

The MLBST algorithm, first sets the initial link loads to zero and then runs the LBST II algorithm described before once without loop breaking (by ignoring the 8th step). After that, the traffic loads on links and switches are known. By using this information, algorithm obtains the link weights using Equations (1)-(5). Now, using these new initial values, it runs the LBST II algorithm again without loop breaking. This process can be repeated several times. In the last run, the MLBST algorithm uses LBST II algorithm exactly as described before (without ignoring the 8th step).

For performance evaluation of MLBST, the simulation scenarios described in previous section are run again by using MLBST. The spanning tree obtained based on LLB criterion with one run for obtaining initial values is ranked second best (2nd) tree by BST algorithm with σ_l^2 equal to 0.0179, while the spanning tree obtained by using two runs for obtaining initial values is the Best Spanning Tree that is shown in Figure 3.a.

By repeating the simulation for SLB criterion, the obtained tree for MLBST with just two runs for obtaining initial values, is the Best Spanning Tree.

The above results show that MLBST algorithm can find the best spanning tree by repeating the algorithm steps several times.

6. Conclusions

In this paper, first a new simple approach named LBST algorithm is introduced for finding the best load balanced spanning tree in metro Ethernet networks. LBST is an iterative algorithm that finds the spanning tree based on load balancing on links and switches. The criterions used are the same as criterions used in our previous algorithm named BST. BST algorithm is an exhaustive search algorithm that finds the Best Spanning Tree. Although, BST algorithm can find the best answer but its complexity is too large in large-scale networks.

Simulation results showed that the output of LBST algorithm is close to the output of BST algorithm while its computational complexity is much less than it.

In the first step of the LBST algorithm, the initial value of link loads must be assigned. In LBST, for simplicity, these initial values were set to zero for all links. For performance improvement, the Modified version of LBST algorithm named MLBST is introduced. MLBST calculates the initial values of link loads and switch loads in a more accurate way. Simulation results showed that the MLBST algorithm can find the best spanning tree by repeating the algorithm steps several times. In summary, the following conclusions can be derived from simulation results:

1. The computational complexity of the proposed algorithms is much less than BST.
2. Although LBST algorithms are not the best but their results are the same or similar to the results obtained by using BST algorithm.
3. The performance of LBST II is very close to the performance of LBST I algorithm, but its computational complexity is much less.
4. MLBST can find the best spanning tree for LLB and SLB criterions by repeating the algorithm steps several times. For this reasons, in practice we prefer to use MLBST for finding the best spanning tree in metro Ethernet networks.

The proposed approaches can be used offline in the design process of a new metro Ethernet network or online during the operation of the network. In online mode, a central node (for example an edge node) is responsible for collecting the traffic and topology information and calculating the best spanning trees.

Note that even though our new algorithms have a good performance in the considered scenarios, other popular realistic traffic models such as multimedia traffic in more realistic metro Ethernet topologies can be considered in future works.

Acknowledgment

This work is supported by Iranian Telecommunication Research Center (ITRC).

References

[1] M. Huynh, and P. Mohapatra, “Metropolitan Ethernet network: A move from LAN to MAN,” *Computer Networks*, Elsevier, pp. 4867-4894, 2007.

[2] CISCO SYSTEM, “Understanding spanning tree protocol,” CISCO white paper, www.cisco.com, 1997.

[3] CISCO SYSTEM, “Understanding rapid spanning tree protocol,” CISCO white paper, www.cisco.com, 2006.

[4] P.M.V. Nair, “Quality of service in metro Ethernet,” *Ph.D. thesis*, Southern Methodist University, 2006.

[5] T. Rodeheffer, C. Thekkat, and D. Anderson,“Smartbridge: A scalable bridge architecture”, *ACM Computer Communication Review*, Vol. 30, No. 4, pp. 205-218, 2000.

[6] K. Lui, W.C. Lee, and K. Nahrstedt, ”STAR: a transparent spanning tree bridge protocol with alternate routing”, *ACM Computer Communication Review* , Vol. 32, No. 3, pp. 33-46, 2002.

[7] F. De Pellegrini, D. Starobinski, M. G. Karpovsky, and L. B. Levitin, "Scalable, distributed cycle-breaking algorithms for gigabit Ethernet backbones", *Optical Networking*, Vol. 5, No. 2, pp. 122-144, 2006.

[8] IEEE 802.1s, *Standards for local and metropolitan area networks*, 2002.

[9] G. Mirjalily, H. Karimi, and S. Rajai, “Load balancing in metro Ethernet networks by selecting the best spanning tree”, *Journal of Information Science and Engineering*, Vol. 27, No. 5, pp. 1747–1759, 2011.

[10] G. Mirjalily, F. Akhavan Sigari, and R. Saadat, “Best multiple spanning tree in metro Ethernet networks”, Proc. of International Conference on Computer and Electrical Engineering, Dubai, UAE, pp. 117-121, 2009.

[11] S. Samadi, G. Mirjalily, and S.M.T. Almodarresi, “A simple algorithm to find the proper spanning tree in metro Ethernet networks”, Proc. of International Conference on Software Technology and Engineering, KLL, Malaysia, pp. 185-189, 2011.

[12] S. Samadi, G. Mirjalily, “Load balanced spanning tree in metro Ethernet networks”, Proc. of Iranian Conference on Electrical Engineering, Mashhad, Iran, 2013.

[13] Y. W. Bang, and C. Kun-Mao, “Spanning trees and optimization problems”, Chapman & Hall, 2004.

[14] A. Kern, “Traffic-driven optimization of resilient QoS-aware metro Ethernet networks”, *Ph.D. Thesis*, Budapest University of Technology and Economics, 2007.

Ghasem Mirjalily received his Ph.D. degree in Telecommunication Engineering from Tarbiat Modarres University, Iran in 2000. He has been a visiting researcher at the Communications Research Laboratory, McMaster University, Canada in 1998. Since 2000, he has been with Yazd University, Iran, where he is an associate professor. Dr. Mirjalily is senior member of IEEE. His current research interests include Traffic Engineering, Metro Ethernet Networks, and wireless Networks.

Samira Samadi received the B.S. degree in Electrical Engineering (Electronics) from Isfahan University of Technology (IUT) in 2006 with honors and M.S. degree in Electrical Engineering (Communication Systems) from Yazd University in 2011. Now, she is working on industrial Ethernet networks. Her interest is working on Data Communication Networks, especially on Traffic Engineering and Metro Ethernet Networks.

Node to Node Watermarking in Wireless Sensor Networks for Authentication of Self Nodes

Hasan Farsi*
Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran
hfarsi@birjand.ac.ir
Seyed Morteza Nourian
Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran
nourian_morteza@yahoo.com

Received: 03/Aug/2013 Accepted: 12/Apr/2014

Abstract

In order to solve some security issues in Wireless Sensor Networks (WSNs), node to node authentication method based on digital watermarking technique for verification of relative nodes is proposed. In the proposed method, some algorithms with low computational for generation, embedding and detection of security ID are designed. The collected data packets by the nodes are marked using security ID. In the proposed method, header is used to mark the packets. Since the nature of the sensor networks is cooperative, using the head of the packets is proposed for authentication. Also using the marked head can prevent from sending and receiving fake data in the other nodes. Simulations have been performed in environments with imposing unrealistic data and having a probability from 1% to 10%. Comparing the proposed method with other methods shows that the proposed method in term of security, reducing traffic and increasing network lifetime is more effective.

Keywords: Watermarking, Wireless Sensor Network, Packets Head, Node, Information Security.

1. Introduction

Wireless Sensor Networks (WSNs) are developing networks that can play an important role in collecting and transmitting data. Application of these networks, especially in strategic areas such as border military districts that are not easily accessible, is highly in progress. Traffic control, healthcare, biotechnology and pharmaceuticals, rescue and military reports - security and espionage cases are some areas that WSNs are used [1,2]. WSNs are multi hop that have the shared nature and internal power management without any direct supervision [1,2]. These networks are constructed by many sensors that are called nodes in this article. These nodes are capable of sensing, gathering and transmission of information and communication together. Depending to some applications of these networks, the transmitted data between the nodes can be very sensitive and crucial. Due to the performance of these networks, there are many security challenges. This corresponds to emerging attacks resulted by bogus nodes and unrealistic information. (Some attacks are node capture and impose false routing information) [1]. Therefore, it is necessary to adapt appropriate procedures to ensure the information security is guaranteed.

In order to prevent the transmission of unrealistic information, a method is required to be applied to distinguish factual information. In fact, certification and authentication methods have to be used. In WSNs due to limitation of computation capabilities, energy, bandwidth and storage, conventional and traditional encryption methods using secret codes may not be implemented. Therefore, it is needed to use low complexity and more appropriate methods to verify the original information.

In this paper, a digital watermarking method for authentication of packets in WSNs has been proposed. Generally, watermarking is used to achieve two main objects; recognition of intangible property information and transmission of confidential information. In recent years, much research in the field of digital watermarking based on data such as audio, image, video and text and information resources, has been proposed in regular networks [3,5]. However, the limited research in this area has been conducted on WSNs [6,10].

In [6], an algorithm for generating the security ID for outgoing packets has been introduced. This identifier is embedded using the pre-defined embedded algorithm in the send-data packets. These packets are transmitted in the whole network and in central receiver station, the main packets are identified and their information is extracted. In [7], a watermarking technique is combined by imposing a series of restrictions during the operation of sensing and processing the information performed by the node. These limitations are appropriate with the encrypted identifier embedded in the data. In some methods, watermarking is only used to prevent unauthorized users to access the information [8], or to guarantee the completeness and detection of the received bits [9]. In some cases, the secure transmission of confidential information is discussed. This confidential data in accordance with the proposed algorithm in

* Corresponding Author

conventional data, before sending the message is embedded [10].

In all the presented methods, with some additional processing on each node, it is tried to increase data security but most of these methods have not considered the shared nature of network performance. In fact, only the sender node is considered and the presented methods are provided independently from other nodes. However, by considering the entire network, it is possible to propose more effective methods providing better performance.

In this paper, we propose a method to verify authentication and to indentify real packets sensed, received and transmitted by original nodes. In the proposed method, considering the nature and performance of WSNs, the packet header is used to embed the security ID. Using certain spaces of the packet header and a new embedded algorithm, the proposed watermarking is performed. At the receiver node, the packet header is examined and verified by the rest of the received packets. In some environments with having a possibility of forced fake packets, the proposed method increases network security, reduces energy consumption and network traffic, effectively.

In following section, the structure of used WSNs is briefly explained. In third section, the procedure of collection, packaging and delivering information in WSNs' nodes are explained. In section 4, the process of watermarking in the proposed method is explained. In section 5, the procedure of verification of packet authentication is demonstrated. Finally, in section 6, the obtaining results by applying the proposed method on WSNs are presented and compared with other methods.

2. Structure of Wireless Sensor Networks

WSNs consist of many sensors called nodes. These nodes randomly or based on a specific design are distributed in an environment [1,2]. These nodes are virtually classified based on their transmission range in each area.

In each cluster, according to the amount of energy belonging to the node, the center of the cluster is selected and other nodes either directly or via other nodes of the cluster are connected to the center of the cluster.

These clusters and the centers are not fixed and during the different business cycle of the network may vary. Network information is transmitted to the main station and sink through the centers of the clusters. In Figure 1, a simple exponential segmentation of different clusters is shown. These clusters are connected to each other through the centers or the boundary nodes [1,2].

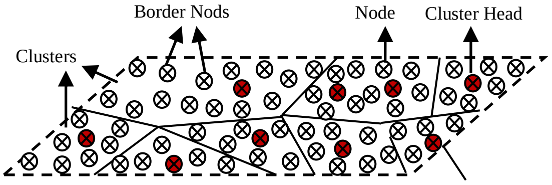


Fig. 1. WSNs' nodes divided into different clusters

3. Collecting, Packaging and Sending the Data

The standard "Zigbee / IEEE 802.15.4" is normally used in WSNs. With respect to this standard and the models presented in [6,10], we propose a new model for packaging the information in the network nodes. A new pattern packet by nodes is given by:

$$\text{Packet} = (\text{Head, Send-Data})$$

Each packet is divided into two parts; the first part is packet header or Head and the second part is called transmission information or Send-data. The packet header contains some information such as the routing information, packet length and other parameters in the standard "Zigbee/IEEE, 802.15.4". The send-data contains the pre-defined information which is needed to be transmitted by the nodes in each time period. It is assumed that each node collects the information in the "kth" cycle. This information is collected and stored in two forms by the node; first, the sensed information by the self-node from its surrounding environment and second, the received packets from other nodes and their authenticity has been verified. If the information is sensed by self-node, each packet is in form of Si= (t, d1, d2, ...,dn). In this packet, "i" refers to the cycle of working nodes in the sense of [(i=1,2, ..., m), m≤ k. "T" refers to the time of sensing information. [Dj, (j=1,2, ..., n)] indicates the parameters that must be sensed by the node. The general pattern of the packets is in form of (N, S1, S2, ..., Sk, D), where "N" indicates identification of the node in the network or ID, and the "D" specifies the length of each packet. If the packets are received from other nodes and their authenticity is verified, they have different parameter of N. In fact, these packets are packaged in the other nodes with the same pattern, except that N contains the node ID which has sensed its surrounding information, primarily.

The packet transmission time by the nodes has to be defined for all nodes. For example, in [6,10], it is assumed that the transmission is performed when the storage is fully loaded. However, due to defining a new model for the packets, the packet transmission time can be performed in specific time intervals. By sending packets scheduled times, network can plan better programming. In addition, some nodes in the network can sense and transmit information both. Therefore, the criteria of fully loading storage can decrease network performance. The reason is that the nodes located in optimum direction of information transmission are unable to receive new packets from other nodes while their storages are not fully loaded even they have performed their duties in time intervals. This problem is due to being the same nodes in WSNs, and therefore the storage capacity is the same for all nodes. If transmitter node sends the information when its storage is fully loaded, then the storage of receiver node should be unloaded to be able to receive the information completely; otherwise the information has to wait until the receiver node located in optimum direction of information transmission sends its data and then by unloading the storage it could receive the information. This problem, due to being several nodes in the network, may destroy or create unreasonable delay in the sending the packets and the network performance will be degraded. Meanwhile, it increases the traffic in optimum direction of transmission. However, if the transmission time is scheduled, then the nodes are able to sense and store the surrounding information and also to receive the packets transmitted by other nodes at the same time.

At a period of time until reaching to retransmission time by the nodes, if the packets are completely received by the node they are packaged and sent on time. But if the packets are not completely received, they are stored in a part of storage and after complete receive, they are packaged and transmitted in the next period time. The transmission time periods are selected such that the transmission is performed before the storage is fully loaded. Moreover, the time period should be set so that the nodes have appropriate transmission rate based on the length of the header and the amount of transmission data packet. For example, the time period should not be considered too low such that the amount of transmission data packet compared to the length of the packet header is unusual. Thus, the transmission time periods are determined based on the type of nodes, the network workspace, request information from the network nodes and the storage capacity. With regards to appropriate timing, it is possible reduce network traffic and delay. By achieving the coordination between the transmission and receive time of packets at the nodes, the appropriate timing is obtained.

If the measure of information transmission time at the nodes is full-load storage, sometimes the traffic increases around the nodes located in optimal routes and especially in the cluster centers. However, it is possible to control the traffic by proper timing. In Table 1, two measures of information transmission time are compared. For this comparison, a working period, "T", is considered for information transmission for each node, and a fixed amount of data is sent by the network. These nodes send data at the same rate for the first stage by the measure of full-load storage and in the second stage with defined schedule. This comparison has been performed based on this reality that in different time periods in the grid, the number of packets that have been transmitted but not yet received, both criteria are calculated. It has been assumed that there are 35 nodes and one cluster center. In first step, the length of each bit stream, 100Kbit has been assumed and in the second step, for each 0.1T, the nodes transmit the data in 8 stages. During the 8 stages, the size of the transmitted data is as same as the first step. The transmission rate is 50kb/s for each node and bit transmission is based on Poisson distribution. For convenience in computing, the total information sent to the network is limited and the traffic in both steps has been considered to be fixed. In the first step, the time of full-load storage is considered to have an exponential distribution and totally production and distribution of the packets on the network trend to Gaussian distribution. Therefore, the nodes located in transmission path receive higher information and include higher traffic. Thus, these nodes have to send the information in shorter time intervals compared to the other nodes and they are time scheduled according to the network conditions. This comparison is intended to provide a time 5 minutes and the number of packets is counted in every 30 seconds.

Table 1. Number of packets not received on the first and second tests

Number of packets in the second stage	Number of packets in the first stage	Calculation time of un-received packets
0.1 T	4	10
0.2 T	22	16
0.3T	29	14
0.4 T	38	15
0.5 T	36	12
0.6 T	39	15
0.7 T	22	13
0.8 T	10	14
0.9 T	3	12
T	-------	-------

According to the table, by the assumption that none of the packets has not been lost by traffic and the delay, it is observed that the transmission and receive of information has better coordination in case of scheduled times information. It should be noted that the number of packets has the same length in the first stage but in the second stage they have variable lengths and in some cases much smaller than the first step. For example, at "0.1T", the number of bits of 10 undelivered packets is 350kbit which is approximately equal to 3.5 packets in first step.

As observed, the first stage has more changes compared to the second stage. These changes sometimes correspond to create traffic and sensible delay in the network, and in some cases can cause data is lost. The traffic in wireless sensor networks is one of the main factors for identification of key nodes, especially center of the clusters. By identification of the location of such nodes, communicating and accessing information is easily performed. It also increases the chance of attacks to the network which even incapacitates the entire network [1,2]. By scheduled times, packaging and proper timing between the sending and receiving packets, the packet congestion around the center of the cluster and probability of location identification decrease. In section 6, the impact of this type of packet scheduling operations in watermarking is fully investigated.

According to the new paradigm of packaging and sending data, the packets sent by each node contain multiple nodes. After receiving the packets, the base station extracts and saves the information of each node according to characteristics of "N" and "D" intended for the node in the network and the business cycle.

4. Process of Watermarking at Receiver

In this section, before indicating the proposed method, in order to generate and embed the security ID of data packets, a brief discussion on use of packet header as the location of embedding security ID is presented.

4.1 Using packet headers for watermarking

In this article, regards to the packet header differs in different nodes and the packets lose their headers by passing through different and they are replaced by new packets with new packet header, the packet headers are used for authentication of packets.

The main reason for using the packet headers is that due to nature of WSNs it is not required the authentication to be only performed in the base station. In most proposed method so far, the transmitter node has been only considered and the nature of communal belonging to the WSNs has not been used. Since each node placed in transmission of the packets can perform the authentication. Because the data is sent through a node, if the client node is able to detect the original information from the fake information then it can remove fake information, receive the original data and save it in transmission information part of the new packets. Using this trick result in reduction of traffic and congestion of the packets around the nodes, especially center of clusters located in optimum pathways and prevents occurrence of the attacks resulted from the network traffic, such as, traffic analysis. Moreover, the consumed energy for receiving one bit in each node is about 0.5µj and for transmission is about 1 µj [1, 6, 10]. Therefore, if the authentication of the information is checked in base station and the information is fake, in this case, all the nodes placed in communication path to base station have to consume a lot of energy for receiving and transmission of the information which is actually fake. This is unacceptable due to energy constraints in the nodes. On the other hand, the transmission information part of the packet may contain several hundred bytes and if the security ID is supposed to be embedded in the transmission information part and it is assumed that the authentication is checked node-to-node, all the nodes have to receive many bytes for authentication of real or fake information. In this case, the consumption of the energy by the nodes increases. For better performance of the network, it is required to propose a strategy in which the client nodes can decide to send or block the information by consumption of minimum energy. Although using packet header for watermarking in other networks except WSNs seems inappropriate due to its change in different transmission stages, in WSNs the transmission of the packet is in cooperation with the other nodes and the overall patterns are similar, using the packet headers for watermarking can be appropriate. Thus, it is possible to use the packet header as a proper choice to verify the ownership. In each node, firstly, the header part of the packet which has a length much smaller than the transmission information part is received and then the authentication operation is performed. If it is right and original, the rest of the packet is received, otherwise the packet is fake and it is discarded.

4.2 Watermarking process

As mentioned in the previous section, incoming packets with sensed data are stored in the data packet transmission part, and a new header with the new security ID is sent. Figure (2) shows the watermarking process in the sender node. According to Figure 2, the sensed information in different working cycles of the node with the received packets from other nodes is packaged. Packaging in each period is stopped when there is enough time for executing watermarking algorithms until arriving the transmission time. According to standard "Zigbee/IEEE802.15.4", packet header in stop time of packaging with network information is formed. By forming the packet header, the algorithms (1) and (2) (which are defined in section 4.2.1) are performed. The result of the algorithm (2) is a new header which contains the security identifier, and proves the authenticity of packet ownership in the client nodes. Embedding the security ID is performed such that in watermarked header, the original header information is maintained. It is obvious that since the security ID of the packets change in every stage the network security considerably increases.

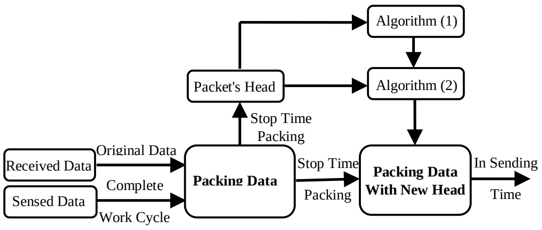


Fig. 2. The watermarking process for sender node.

By new definition of information packaging in the nodes and similarity of the nodes in WSNs, some defined fields in the header seem to be insignificant. For example, the field of the sender node, according to the standard “zigbee”, has allocated 2 bytes for header. As noted, the transmission information part contains the sender's node ID. Meanwhile, according to the standard, some header fields can provide some void bits. In this study, we try to embed the security ID using these fields. This is achieved such that the information in the header is maintained. In fact, in this methodology, the header information is not manipulated and only security ID bits are embedded in the void and least significant bits of the header which are defined in both the transmitter and receiver nodes.

Now, we are encountered with two issues; first, how to produce secure ID, and second, how to embed the Security ID.

4.2.1 Generation of security ID

First, the packet headers are divided into two parts according to the standard used. The first part consists of some fields which remain unchanged and its information has been used generate the security identifier. In this article, these fields are called fixed fields of header. The second part consists of some fields such as the field related to the sender node or tab ratings, which are either less important or can include more number of bits compared to the fixed fields of header. In these fields it is possible to define a number of void bits, which their places are known in the receiver and the transmitter. These fields are called variable fields. This part is used to embed the security ID bits.

In order to generate the security ID bits, the proposed algorithm in [6], has been used but with a slight change in the algorithm, a “Hash” function is applied. The “Hash” functions generate a unique output by receiving the specific inputs under predefined conditions. One of the inputs to the hash function is the number of bits, "L", which is considered as a security ID and it is shared in all nodes. Other input is a security key, "k", which is a prime number and it is shared automatically in all nodes. The certain bits of the fixed fields of header which are known in both sender and receiver are the other input. This input causes that the probability of generation of fake security ID considerably reduces due to being difference of headers in different stages of transmission. Note that this algorithm is executed only when the header is received. Because the fixed fields and other parameters are transmitted without any changes in all nodes. The hash function generates a bit string as a security ID only by simple operations. The security ID generation algorithm has not to be simply detectable, to be low complexity.

Algorithm 1. Generation of security ID bits

```
/ / Initial data k, L
1) fix = constant values in header fields
2) w = LSB (fix)
3) WM = Hash [L, k, w]
```

4.2.2 Embedding the security ID

The output of Hash function, WM, is a L-bit string. This bit string has to be placed in variable fields of header. First, the bits in the variable header fields that are marked to embed the bits of security identifier are numbered sequentially. This numbered bit string is indicated by "d", where d(i) refers to ith bit of the bit string. The length of WM-bit string can reach to tens of bits. However, as will be shown in Table 2, an 8-bit string can also meet the security expectations.

Algorithm 2 shows the process of embedding the security ID bits. The algorithm includes the property of rotational placement and security ID may not be easily detected and recovered. An important feature of this algorithm is non-interference in the placement of information in the header.

After construction of the new bit string, d, the sign bits are placed in the variable fields of header according to the specified numbers, and the information with the new header is then transmitted.

Algorithm 2. Embedding WM bits in “d”

```
/ / Initial data k, L
1) len = length (d)
2) For j = 1: L
       p = [j * k / len]
       mod = j * k mod len
       if p + mod <len
           d (p + mod +1) = WM (j)
       else
           d (p + mod-len +1) = WM (j)
       End if
   End
```

5. Verification and Authentication of Packet Ownership

After transmission of the packets by the sender node, it is firstly needed to verify the authentication of packet ownership in client node and then to be fully downloaded. Figure 3 shows the receive process in client node. In client node, the header of the packet is fully received and then the algorithm 1 is executed. Since the fixed fields of header remain unchanged, the algorithm 1 in the receiver and the transmitter has similar results. After receiving the header, the sign bits in the variable fields of header are extracted.

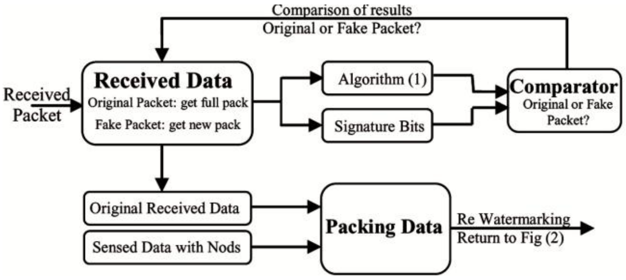


Fig. 3. Block diagram related to verification of authentication of packet ownership in client node.

According to the algorithm 1, WM-bit string is generated based on fixed fields of header. Then, based on pre-defined location of the sign bits in the receiver and the transmitter, the bit string, d, is extracted. Next, the bit string of WM' is extracted from bit string of d' by reverse operations. By comparing WM and WM ', the received packet is verified whether original or fake. This procedure is performed based on algorithm 3. If the packet is original, the remaining information of packet is received otherwise the packet is removed from the receive cycle and discarded. In order to compare WM and WM ', we use a measure which indicates virtual difference between WM and WM ' [6]. This measure is practically obtained by consideration of environment and probability of error created by the environment. After verification of being original or fake packet, a command is sent to the receiver node. If the packet is original, it is commanded to the receiver node to continue receiving the packet. However, in case of detection of fake packet the remaining information of packet is not received and it is commanded to the client node to receive another packet header. After receiving the original data, if the packet transmission information is fully received, the information is sent to the packaging part of the node. Meanwhile, the node information sensed in different business cycles is transferred to packaging part of the information. From this stage onwards, if the information is needed to be embedded, the watermarking is re-performed as shown in Fig. 2.

Algorithm 3: Security ID bits extracted

```
/ / Initial data k, L and t values of allowable error
1) The Algorithm (1) WM Production
2) Len '= length (d');
3) For j = 1: L
        p = [j * k / len ']
        mod = j * k mod len '
              if p + mod <len '
                  WM '(j) = d' (p + mod +1);
              Else
                  WM '(j) = d' (p + mod-len +1);
          End if
End

4) E = sum (Xor (WM, WM '))
5) If E <t
        Accept data
   Else if E> t
        Reject data
   End if
```

6. Applying the Proposed Method on Wireless Sensor Networks

As noted earlier, the proposed method provides acceptable results in the environments containing fake packets. In this section, the existing security in the proposed method is discussed. Then, network traffic analysis and energy consumption will be discussed. In the analysis, it is assumed that all the nodes are imposed by 0.01 to 0.1 fake packets. The process and transmission rate of the packets are similar to the assumptions in section 3.

6.1 Security of the proposed method

In this method, a harmonic function is used such that its process is not simply recognizable. Since the input of the function is the security key which is inaccessible, the probability of fake WM generation is quite low. In addition, the state of placement regards to varying WM in different nodes causes to increase the security of the received packets. If an undesired system is going to produce fake header and ID which is undistinguishable by the network, it is required to consider huge aspects. It should be noted that since the header information of each packet belonging to the packet itself, is used for generation of ID, the ID of the packets differs even they are generated in one node. Since undesired system is unaware about the parameters of "l" and "k" and the process of ID generation in the nodes, it can assume that the probability of correct generation of ID for each bit is 0.5. Since the information of the fixed header fields is used for generation and authentication of ID at receiver, due to being difference between the headers in various stages of sending headers, the probability of generation of fake ID highly reduces. If the total length of the header is n-bit, the probability of generation of fake ID compatible with the original ID is multiplication of the probability of finding the L-bit location of ID in header and the probability of correct ID. This is given by:

$$p(n,L) = \frac{1}{\binom{n}{L}} 0.5^{L} = \frac{L!(n-L)!}{n!} 0.5^{L} \qquad (1)$$

Table 2 shows the probability of fake ID for different values of n.

Table 2. The probability of fake ID with length of 8 bits

n	L	P(n,L)
48	8	1.04×10-11
56	8	2.75×10-12
64	8	8.83×10-13
72	8	3.27×10-13
80	8	1.35×10-13
88	8	6.08×10-14
96	8	2.95×10-14
104	8	1.52×10-14
112	8	8.22×10-15
120	8	4.65×10-15
160	8	4.38×10-16
200	8	≅ 0

It is observed that by choosing only 8 bits for the ID and watermarking the probability of generation of fake ID considerably reduces. It seems that due to variation of the ID during the transmission process, 8 bits provides the network security expectations. However, using less bits results in lower complexity and lower costs for transmission.

The most important factor in authentication methods is the accuracy of correct verification of original packets such that neither the original packet is discarded nor the fake packet is received. Figure 4 represents the percentage of detection for the original and fake packets. The upper curve represents the percentage of received original packets to the total original packets and the lower curve illustrates the percentage of received fake packets to the total fake packets. In this experiment, n=72 and the results for different values of L are examined.

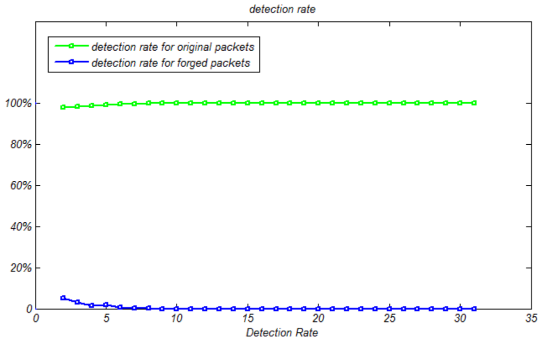


Fig. 4. Percentage of received packets for original and fake cases.

It is observed that if the number of bits for WM is more than 8 bits, ideal results are achieved. Compared to [6], although the methods are very different but they have the same results for more than 4 bits.

Therefore, according to Tab. 2 and Fig. 4, the proposed method can provide the required network security expectations. Meanwhile, the security identifier produced and embedded by the proposed method contains lower number of bits compared to other methods [6].

6.2 Network traffic analysis

One of the main methods used by the foreign system to detect the location of the nodes is network traffic analysis. By the network traffic analysis the location of the nodes placed in strategic areas or the position of cluster centers in the optimal transmission paths are identified. This provides different attacks such as wormholes, sinkhole attacks and node compromising to be imposed to the network [1,2]. The two-step proposed method can reduce network traffic. The first step consists of the new definition of packaging and scheduling packets and second step includes using the headers for watermarking and authentication of the packets. It is obvious that traffic reduction causes reduction of delay in transmission of the information in the network.

Fig. 5 shows the proposed method against the method in [6,10] for packaging and information transmission in terms of the number of packets existing in queue in different time intervals. These results have been obtained by computation of the number of packets produced in the node for each time interval independent of previous time interval. In this procedure, some nodes obtain the necessary bandwidth to send the packets in the time interval generated and some other nodes, due to limitation of bandwidth, are unable to send the information. In this case, the generated packets which have not been transmitted, stay in transmission queue and the node, as soon as accessing to the necessary bandwidth send them. The number of packets stayed in the queue for each time interval is:

$$P_P - P_T = P_Q \tag{2}$$

Where the PP indicates the number of packets generated in each time interval, PT, is the number of generated packets which have been transmitted at the same time interval and PQ is the number of generated packets which have been stayed in queue. PQ is calculated at each time interval. The percentage of ratio of PQ to PP has been shown in Fig. 5. These values represent the percentage of packets which stay in queue at every time transmission. Obviously, higher values represent higher network traffic and delay. These results have been only obtained for the impact of packaging and information transmission patterns in network traffic and the watermarking applied in the proposed method and in [6] has not been considered in the simulation.

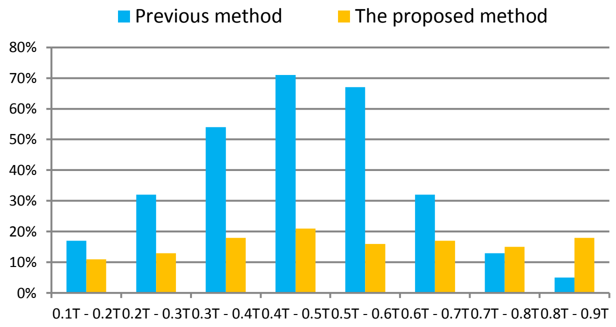


Fig. 5. The amount of queued packets per time slot

According to Fig. 5, if a criterion is filling storage [6.10], the ratio of PQ to PP increases in some time intervals. The reason is the packet congestion in some intervals and the limited bandwidth. The limitation of bandwidth has to be considered as one of the fundamental challenges in this type of network where the transmission delay is important. Inattention to bandwidth limitation causes packet loss and node information. The congestion of nodes which do own transmission as soon as filling the storage increases and consequently the number of packets prepared for transmission increases. In this case, some packets are sent at generated time and some packets stay in queue to be sent at next time intervals with having the required bandwidth. Regards to assumption of Gaussian distribution, the congestion of generation and transmission of the packets increases in median time intervals and non-transmitted packets affect the transmission in the next time intervals. Therefore, for median time intervals between (0.3T-0.6T), the number of queued packets greatly increases. Since the traffic load and the amount of information for transmission have been considered the same for the proposed method and the method presented in [6,10], as observed in Fig. 5, the method in [6,10] packs and transmits the desired information and therefore the generation and the congestion of the packets reduces in the final time intervals and the nodes have higher bandwidth for transmission of remaining packet in the network. The important issue is that the network traffic has to be controlled such that key nodes cannot be identified and with the fluctuations in the density of traffic and the congestion of packets for transmission, the network may be disordered.

According to the new model packaging as shown in Fig. 5, it is observed that with having the timing for transmission and appropriate pattern for packaging, the number of queued packets for transmission, even in high congestion time, is almost unchanged. In other words, with accurate timing for information transmission, it is possible to reduce the delay and prevent fluctuation of network traffic at different time intervals.

Figure 6 compares the proposed method with the method presented in [6]. The comparison is based on the freed bandwidth in each area of the network, without applying the watermarking. The number of nodes is variable in each area and the results have been obtained regards to different number of nodes. For each number of nodes, the first column and the third column are related to the method presented in [6] and the proposed method, respectively. In the second column, the watermarking method presented in [6] in combination with the new model of packaging and information transmission has been shown. In fact, the second column indicates the effect of the new model applied on [6] in reduction of network traffic. For example, if there are 50 nodes in center of a cluster, the proposed method regards to imposing the percentage of the fake packets, is able to free 14% of occupied bandwidth compared to without applying watermarking case. This reduction is more sensitive in strategic areas. For two reasons, the proposed method provides better performance; first the procedure of packaging and transmission and second prevention of fake packets distribution in the network. For example if the fake packets are received and distributed in an area with 50 nods (sending the packets in the network broadcast is assumed as broadcast), they are received and sent many times until reach to cluster center several times and they are finally discarded in base station. However, the proposed method is able to verify and discard the fake packets in primary step from the send-receive cycle and therefore increases the bandwidth for transmission of the original packets.

The calculation is based on computation of the amount of information which has been sent and received in the network regards to the amount of fake information for different time intervals. The ratio of the obtaining values to the case without watermarking are calculated and shown in Fig. 6. The different values in different methods can indicate as free bandwidth by applying the related methods.

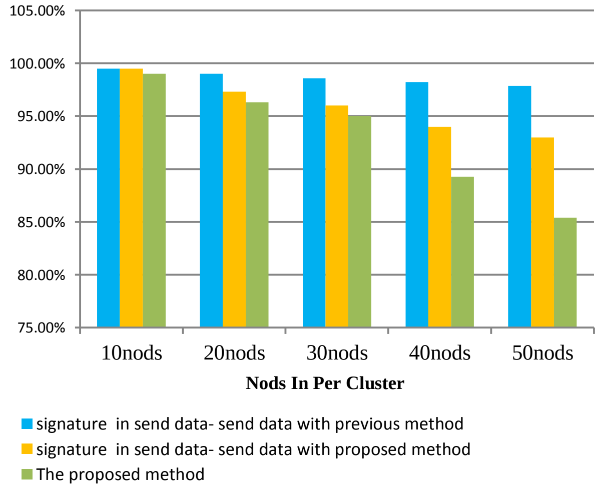


Fig. 6. Occupied bandwidth in different methods with and without watermarking

6.3 Evaluation of energy consumption

One of the important issues that have a large impact on the performance of each method in WSNs is the problem of energy consumption. As referred before, the proposed method is appropriate when the probability of imposing fake packets exists. As the probability increases, the proposed method provides better performance, especially in terms of energy consumption and traffic reduction. Since the energy consumption of the nodes for sending and receiving the bits is more than internal process applied on the bits, although extra processing is performed by the proposed method in each node, by discarding fake packets from send-receive cycle not only the consumed energy resulted by extra processing is compensated but also some energy is stored in the node. This increases the longevity of the network and provides better performance. Reduction of the traffic caused by the proposed method also has a direct effect on reducing energy consumption by the nodes. Figure 7 indicates the amount of consumed energy by the proposed method in comparison with the method presented in [6]. This has been achieved based on computational models discussed in [6.10]. The only difference is that this has not been achieved for only one node and has been obtained by averaged total impact of traffic in each area. For authentication of packets in [6], energy consumption of each node slightly increases compared to the case without watermarking but in the proposed method, due to accurate transmission timing and the number of nodes in each area, the energy consumption of each node even decreases. In Figure 7, for different scheduling in the network, the amount of consumed energy for variable number of nodes in each zone has been calculated. The effect of traffic reduction on reducing energy consumption is quite clear. The experiment assumptions are similar to the previous example. There are several points which are briefly mentioned:

First, for low number of nodes in each area, the consumption of energy is slightly higher by the proposed method. The reason is reduction of packet numbers and low transmission rate. Due to being smaller length of packets compared to other methods, it is required to send a packet header for each packet which causes low rate transmission of information and for similar amount of information; it is needed to consume more energy. However, as observed in Fig. 7, this problem is slightly removed by increasing the transmission time.

The second point is the effect of traffic on reducing the consumption of energy. As observed, by increasing the number of nodes in each zone, the averaged energy consumption of the nodes reduces compared to the case of without watermarking. According to the description given in section 6.2, the proposed method can reduce the network traffic compared to the case without using watermarking. Obviously, by increasing the numbers of nodes in the network, the number of send and receive packets increases. If the fake packets are distributed in the environment containing wireless sensor network and there is not a criterion for verification of the fake packets, due to increasing the number of nodes and the information consequently, the number of fake packets and undesired information increases in case of unused watermarking and therefore more energy is consumed by the network. As shown in Fig. 7, by increasing the number of the nodes, the proposed method is able to decrease the consumed energy approximately 15% compared to the case without using watermarking.

The third point is that by increasing the transmission time such that the performance of the network is not been affected, the energy consumption more reduces. The major reason is increasing the packet length with increasing duration of time. As a result, the information transmission rate is higher and the amount of desired information is transmitted by lower number of packets. It has to be mentioned that increasing the time duration more than a certain threshold results in more delay and more congestion of the packets in the network. Therefore, it has to select the most appropriate procedure for transmission of information by a balance between all conditions and restrictions.

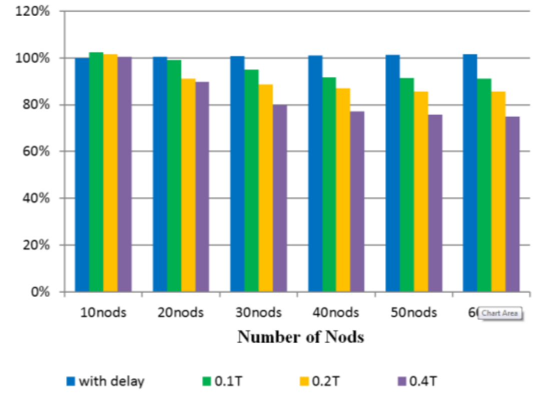


Fig. 7. Analysis of energy consumption in each area

7. Conclusions

In this paper, due to the shared nature of sensor networks, we proposed a method which provides acceptable results in terms of safety, traffic and energy consumption.

The proposed technique in the environments with a high risk of attacks provides better results than existing methods. In environments with very low probability of attacks and secure environment, the proposed method imposes additional processes leading to reduction of network performance. However, for the non-secure environments, the proposed method through relatively light processing and therefore low complexity provides considerable security for the network. Furthermore, in spite of general methods, the proposed method results in positive effects on the network lifetime. The new model for packaging and transmission of information applied in the proposed method provides the network the capability of time programming for sending and receiving the information. This reduces delay in data transmission.

References

[1] Misra, S.; Woungang, I.; Misra, Ch.S. “Guide to Wireless Sensor Networks.”;Springer-Verlag London Limited 2009.

[2] Tse, D.; Viswanath, P. “Fundamentals of Wireless Communication.”; Cambridge University Press 2005.

[3] Yin, H. C.; Lin, F. Q.; Ding, R. “A Survey of Digital Watermarking.”; Journal of Computer Research and Development, vol.42, no. 7, pp. 1093-1099, 2005.

[4] Agrawal, R.; Kiernan, J. “Watermarking relational databases.”; in Proceeding of the 28th VLDB Conference, Hong Kong, China: VLDB Press, 2002, pp. 155–166.

[5] Sion, R.; Atallah, M.; Prabhakar, S. “Rights protection for relational data.”; in Proceedings of ACM SIGMOD, San Diego, CA, USA: ACM Press, 2003, pp. 98–109.

[6] Dong, X.; Li, X. “An Authentication Method for Self Nodes Based on Watermarking in Wireless Sensor Networks.”; Networking and Mobile Computing, 2009. WiCom '09. 5th International Conference 24-26 Sept. 2009; pp: 1- 4.

[7] Feng, J.; Potkonjak, M. “Real-time watermarking techniques for sensor networks.”; in SPIE Security and Watermarking of Multimedia Contents, Santa Clara, CA, USA: SPIE Press, 2003, pp. 391–402.

[8] Sion, R.; Atallah, M.; Prabhakar, S. “Resilient rights protection for sensor streams.”; in Proceeding of the 30th VLDB Conference, Toronto: VLDB Press, 2004, pp. 732–743.

[9] Guo, H.; Li,Y.; Jajodia, S. “Chaining watermarks for detecting malicious modifications to streaming data.” Information Sciences, no. 177, pp. 281–298, 2007.

[10] Xiao, X.; Sun, X.; Yang, L.; Chen, M. “Secure data transmission of wireless sensor network based on information hiding.”; in Proceedings of The Fourth Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services. Philadelphia, PA, USA: IEEE Press, 2007 ,pp. 1–6.

Hasan Farsi received B.Sc and M.sc degrees from Sharif University of technology, in Communication engineering in 1993 and 1995, respectively. Since 1995, he was employed in university of Birjand as an academic staff. He was entered as a Ph.D student in Center for Communications Systems research (CCSR), University of Surrey, UK, in 1999 and received the Ph.D degree in communication Eng. in 2003. Since 2003, he was re-employed in university of Birjand and at the moment he works as associate professor in department of electrical and computer eng., university of Birjand. So far he has published more than 20 journal papers, 25 conference papers and a book titled "Time variable PDF presented by neural network" in 2013. Also, he is a member of scientific committee in three international journals and reviewer of IET. The interest areas are signal processing, speech processing, image and video coding.

Seyed Morteza Nourian received B.Sc and M.Sc degrees from Islamic Azad university of Najaf abad and University of Birjand, in 2010 and 2012, respectively. At the present time, he works as an invited lecturer in university of Najaf abad, Isfahan in department of electrical and computer engineering. So far he has published one journal and two conference papers. The interest areas are speech processing and wireless sensor networks.