In the Name of God

Journal of

Information Systems & Telecommunication

Vol. 6, No. 4, October-December 2018, Serial Number 24

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, ACECR, Iran
Editor-in-Chief: Masoud Shafiee, Amir Kabir University of Technology, Iran

Executive Manager: Shirin Gilaki
Executive Assistants: Mohammad Darzi, Sahar Seidi
Editors: Mahdokht Ghahari, Behnoosh Karimi
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
E-mail: info@jist.ir , infojist@gmail.com
URL: www.jist.ir

Indexed by:

- SCOPUS	www. Scopus.com
- Index Copernicus International	www.indexcopernicus.com
- Islamic World Science Citation Center (ISC)	www.isc.gov.ir
- Directory of open Access Journals	www.Doaj.org
- Scientific Information Database (SID)	www.sid.ir
- Regional Information Center for Science and Technology (RICeST)	www.ricest.ac.ir
- Iranian Magazines Databases	www.magiran.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 & Signal Processing Research Group, ICTRC

Acknowledgement

JIST Editorial-Board would like to gratefully appreciate the following distinguished referees for spending their valuable 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)

- Aghababa, Hossein, University of Tehran, Tehran, Iran
- Cheraghchi, Hamideh Sadat, Shahid Beheshti University, Tehran, Iran
- Dashtbani, Majid, Razi University, Kermanshah, Iran
- Dehkharghani, Rahim, University of Bonab, East Azerbaijan, Iran
- Faraji, Neda, Imam Khomeini International University, Qazvin, Iran
- Farsi, Hassan, University of Birjand, South Khorasan, Iran
- Fasanghari, Mehdi, University of Tehran, Tehran, Iran
- Ghaemmaghami, Shahrokh, Sharif University of Technology, Tehran, Iran
- Ghazalian, Reza, Babol Noshirvani University of Technology, Mazandaran, Iran
- Ghazvini, Mahdieh, Shahid Bahonar University, Kerman, Iran
- Gholamalitabarfirouzjaee, Saeed, Ball State University (BSU), Indiana, USA
- Kasaei, Shohreh, Sharif University of Technology, Tehran, Iran
- Moghavvemi, Mahmoud, University of Malaya (UM), Malaysia
- Mohammad zadeh, Naser, Shahed University, Tehran, Iran
- Mohammadi Zanjireh, Morteza, Imam Khomeini International University, Qazvin, Iran
- Mohammadzadeh, Sajjad, University of Birjand, South Khorasan, Iran
- Rasouli, Abdolreza, Qom University, Qom, Iran
- Sabbaghi Nadoushan, Reza, Islamic Azad University, Tehran, Iran
- Sadeghzadeh, Ramezan Ali, K. N. Toosi University of Technology, Tehran, Iran
- Saeb, Armin, Stellenbosch University, Stellenbosch, South Africa
- Tofighi, Sara, Research Institute for Information and Communication Technology, Iran
- Yousefi, Saleh, Urmia University, West Azerbaijan, Iran
- Zeinali, Esmaeil, Qazvin Islamic Azad University, Qazvin, Iran

Table of Contents

Retinal Vessel Extraction Using Dynamic Threshold and Enhancement Image Filter From Retina Fundus

Erwin*
Department of Computer Engineering, Faculty of Computer Science, Universitas Sriwijaya, Indralaya, Indonesia
erwin@unsri.ac.id
Tomi Kiyatmoko
Department of Computer Engineering, Faculty of Computer Science, Universitas Sriwijaya, Indralaya, Indonesia
tomi.kiyatmoko19@gmail.com

Received: 18/Mar/2019 Revised: 26/May/2019 Accepted: 22/Jun/2019

Abstract

Retinal blood vessels in every human being are important elements of various shapes and sizes, and retinal blood vessels can also determine various types of diseases. Therefore, retinal blood vessel extraction from the retinal fundus image is a key step in the process of recognizing the shape and size of disease patterns in the retina so that it can determine diseases of different types, but the feasibility of retinal blood vessel patterns is important for subsequent processes such as detection, identification, and classification. The previous method that focused on retinal vessel extraction has its own characteristics, especially in the pre-processing, extraction, and post-processing stages. However, there were still many characteristics in previous studies that made it insufficient to meet the needs of ophthalmologists, especially in the segmentation stage, many retinal vessels disappeared at the ends and became thicker, even assuming noise became a retinal blood vessel. Therefore, we conducted an experiment to develop retinal blood vessel segmentation in the medical world using Retina Fundus Dynamic Threshold and Image Enhancement Filter. By using the latest approach in the preprocess namely Butterworth Bandpass Filter as Enhancement Image Filter and the latest segmentation using Dynamic Threshold with a small time value for implementation with low device specification. In this paper we use the databases of DRIVE and STARE. So the proposed method for achieving the average measurement parameters from the DRIVE database is 94.77 percent accuracy and the STARE database is 87.68 percent accuracy.

Keywords: Butterworth Bandpass Filter; Dynamic Threshold; DRIVE; Retinal Blood Vessels; Segmentation; STARE.

1. Introduction

In the diagnosis of retinal disease, retinal blood vessels play an important role in determining certain diseases. Therefore retinal blood vessel extraction from the retinal fundus is a key step in the process of recognizing the shape and size of patterns of disease in the retina with its variety. Images that cover retina, blood vessels, and optic nerve are called retinal image. The retina contains important parts such as blood vessels, macula, cornea, iris, and lenses. Blood vessels have a variety of forms and different types of individuals. The structure of the vessels in the retina image is shaped from the branch line so that retinal blood vessel extraction becomes a line detection problem [1]. Therefore, in dealing with eye syndrome, blood vessels become an important medical object.

Retinal vascular extraction is the main step in detecting blindness-causing eye diseases, including retinopathy [2]. Retinal vascular extraction from digital retinal fundus images is a key step in many computerized diagnostic processes in retinal eye pathology such as diabetes retinopathy, macular degeneration, glaucoma, and occlusion of retinal artery [3]. In high-resolution retinal fundus images can help ophthalmologists diagnose disease automatically by extracting blood vessels, optical disks, and macules [4]. At present, the problem often occurs is a problem found in the image of the retina in the extraction of the blood vessels.

Previous methods of extracting retinal blood vessels have their own concentration, particularly in processing. Blood vessel treatment is characterized by color (redness), shape (curvature), gradient (limit), contrast (background image), etc. However, there are still many characteristics that make it not enough to satisfy as needed. However, there are still many characteristics that make it insufficient to satisfy as needed.

From the description above, the author has observed several existing methods for the extraction of blood vessels with the aim of adjusting medical needs. In steps, the image is changed to grayscale with a certain intensity using gamma correction, the image is enhanced by brightness using Contrast-Limited Adaptive Histogram Equalization, then the Butterworth Bandpass Filter as an image smoothing and sharpening. Then the blood vessels are segmented or extracted using Dynamic Threshold after closing morphology, Median filtering and removing small pixels to clear unwanted noise so that the final results are obtained.

2. Related Work

In this section, we review previous work with different methods on the extraction of retina blood vessels. A lot of research over the past few years has focused on methods that have their own characteristics such as J Dash [5]

* Corresponding Author

propose an approach with three phased segmentation processes where the image is processed using CLAHE in the first step. The segmentation is then performed using ISODATA method. Then, morphological cleaning is done in the third stage to reduce the noise generated during the segmentation process. However, the generated data is still not sufficiently accurate based on the Ground Truth Dataset and is stopped at the ISODATA stage which has a definite value in the cluster number if it finds a threshold value.

Dash et al [6] proposes an approach that is Unsupervised extraction of blood vessel retinal. it has three stages segmentation process is the process of pre-processes, extraction, and post-process. Unsupervised on the segmentation process using the adaptive threshold. Although it has a machine learning system in the paper, it is not explained about the unsupervised system and the pre-process that only relies on gamma correction in order to increase the segmentation parameter results.

Biran et al [2] Proposed an image enhancement filter method such as Gabor, Gauss and Frangi. The results obtained only in the form of images of blood vessels, however, still have some excessive points at the end of the vessels.

Guu et al [7] Proposed automatic retinal vessel extraction method, classified into three tracking-based and filtering-based classification categories. This method has a low level of complexity. These methods are very sensitive to noise, so that the performance is not very good. In this method, the non-vascular structure in the image of the retinal blood vessels also makes the appearance of misclassified pixels.

There are also still images in this method of small segments of blood vessels that are not visible and can not be extracted. According to Soomro et al [8] use CLAHE techniques to provide better images, but diagnostic performance in terms of blood vessel segmentation may decrease due to the loss of small vessels that will disappear.

Several processes discussed in the experiment can identify diabetes in the retina of the eye. Ben Abdallah et al [9] using a multiscale medialness method that shows that the proposed method performs well with contrast but that the retinal fundus image is so low that the green channel is considered because it has the highest contrast between the blood vessels and the background and the end of the slowly disappearing blood vessel also takes a lot of time to get the test results.

Kamble et al [10] using phase stretch transform and the method used by many includes interference after changing the transformation stage to produce a small increase in the width of retinal blood vessels in post-processing operations.

B.Khomri et al [11] using the elite-guided multi-objective artificial bee colony (EMOABC) method as the pre-process segmentation stage with the top hat and green channel, although there are still deficiencies in the focus of retinal vessels, especially for accuracy.

3. Proposed Method and Model

In this paper we are using using Dynamic Threshold and Enhancement Image Filter From Retina Fundus. For extraction of retinal fundus images using Dynamic threhold. The extracted grayscale is improved using gamma correction and CLAHE is applied to extract retinal blood vessels in low quality then a Butterworth Bandpass Filter is performed to sharpen the image as a pre-process. Along with the process, some non-pixel vessels will be removed with the help of the post-processing phase or bwareopen, median filters and morphology closing. The stages of the proposed method of blood vessel extraction are illustrated in Figure 1.

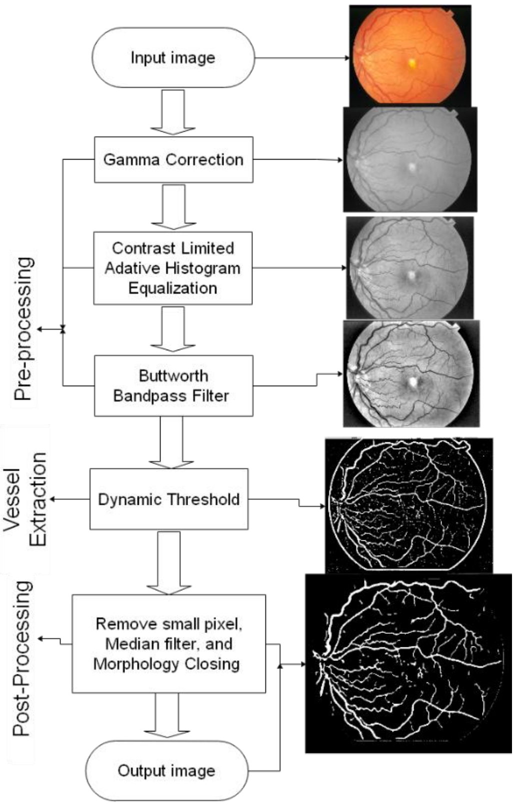

Input image
Gamma Correction
Contrast Limited Adative Histogram Equalization
Buttworth Bandpass Filter
Pre-processing
Dynamic Threshold
Vessel Extraction
Remove small pixel, Median filter, and Morphology Closing
Post-Processing
Output image

Fig. 1. Method Diagram

3.1 Pre-Process

The first step is to prepare a Retina Fundus Image as an Input Image. Fundus Retina images in the form of data originating from DRIVE and STARE database. After getting the data in the Input Image step. Then the Pre-process is done to achieve higher performance accuracy such as image enhancement, filtering, color change, etc. Several techniques are performed in pre-processing as follows:

- Gamma Correction

In the first step of preprocessing, this is Grayscale with the Gamma Correction approach to improve image brightness [12].

The formula used for gamma correction [6] [12] amirin Eq.(1) is as follows:

$$O = C\,I\gamma \qquad (1)$$

where O = output, C = constant, γ = gamma and I = input.

- Contrast-Limited Adaptive Histogram Equalization

After performing the gamma correction then using Contrast-Limited Adaptive Histogram Equalization

(CLAHE). The aim of CLAHE is to improve low contrast image quality [13][14].

CLAHE is an improved version of Adaptive Histogram Equalization (AHE) that divides images into small regions and works on individual constituencies where the contrast of each small constituency is strengthened [6].

- Butterworth Bandpass Filtering

After doing CLAHE, then filtering is Butterworth Bandpass Filtering. The Butterworth Bandpass Filter is a combination of Butterworth Low pass and Butterworth High Pass. Filtering is done to change the fundus image to be sharper by taking high and low frequency data to a certain extent.

With the formula in Eq. (2) Butterworth Low pass Filter [15][16] as follows:

$$H_{LP}(u,v) = \frac{1}{1+[\frac{D(u,v)}{D_L}]^{2n}} \tag{2}$$

With the formula in Eq. (3) Butterworth High pass Filter [15][16] as follows:

$$H_{HP}(u,v) = 1 - \frac{1}{1+[\frac{D(u,v)}{D_H}]^{2n}} \tag{3}$$

With the formula in Eq. (4) Butterworth Band pass Filter [15] as follows:

$$H_{BP}(u,v) = H_{LP}(u,v) * H_{HP}(u,v) \tag{4}$$

Where H_{LP} is Butterworth Low Pass Filter, H_{HP} is Butterworth High Pass Filter, H_{BP} is Butterworth Bandpass Pass Filter, DL is the low pass filter frequencies in Eq. (3) and DH is the frequency of the high pass filter in Eq. (4), n is the sequence of filters D (u, v) is the matrix of u and v.

3.2 Vessel Extraction

After pre-processing, the next step is to do a blood vessel extraction (vessel segmentation). In this process it consists of Dynamic Threshold and complement image as follows. Steps taken in the vessel process this extraction is the Dynamic Threshold [17]. This process converts grayscale images into binary images in the form of 0 and 1 values.

An alternative approach to finding local thresholds is to test statistically the values of the local environmental intensity of each pixel [6]. The most appropriate statistics depend on the input image. Simple and fast functions include the average local intensity distribution, the median value, or the average minimum and maximum values. With the formula [6] [18] in Eq. (5) as follows:

$$L(x,y) = \begin{cases} 1, & \text{if} I_E(x,y) > T(x,y) \\ 0, & \text{if} I_E(x,y) \le T(x,y) \end{cases} \tag{5}$$

where L(x, y) is the result of dynamic threshold and IE(x, y) and T(x, y) are state representations for inputting an image.

3.3 Post-Process

After doing the vessel for extraction. The final step is to post-process. This process will clean up the noise. Cleaning noise using morphology closing, median filtering and bwareopening or removing small pixels from binary images and smoothing out the end of the root of blood vessels resulting from the binary or threshold image. Background cleaning is also done to focus on taking blood vessels by changing the overall black retina and the outer side or apart from the white retina, so that the unnecessary background will disappear through a reduction operation. Median Filter is a nonlinear digital filtering technique, which is often used to eliminate noise from images or signals [19]. This noise reduction is a typical post-processing step to improve the final processing results. The formula used for median filter [16] in Eq. (6) is as follows:

$$y[m,n] = \text{median}\{x[i,j]\}, (i,j) \epsilon\ \omega \tag{6}$$

Where ω represents a user-defined environment, centered around the location [m, n] in the image.

4. Results and Performance Analysis

In this research we used MATLAB with the specifications of device are Celeron processor Dual core 2957U 1.4 GHz laptop, Intel HD graphics with 2 GB RAM.

The Fundus Retina used in this paper for performance of the segmentation process is verified and estimated on the publicly available are data on 20 retinal fundus images obtained from DRIVE [21] and STARE [20] databases.

The DRIVE database is one of the most used databases to focus on segmenting retinal blood vessels. The set of 40 images has been divided into a training and a test set, both containing 20 images. The DRIVE database on the segmentation process of this research using TIF format and the segmentation results using JPG format.

Same as the DRIVE database, STARE was funded by the U.S. National Institutes of Health. During its history, over thirty people contributed to the project, with backgrounds ranging from medicine to science to engineering. STARE is available in full 400 sets of raw images, ground truth with blood vessel segmentation, optical discs, and diagnosis codes.

The following is one image processing from the dataset Figure 2. (a). The image of the initial grayscale from the pre-process with the gamma grayscale correction can be seen in Figure 2. (b). Gamma correction is very important to present the image exactly on the monitor. Corrected images can look either bleached or too dark. The greater the gamma value for grayscale correction, the image will be fainter. In this experiment we used a gamma value of 0.9 seen in Figure 2. (b). Gamma values provide brightness variations in the image can be seen in Figure 2. (d). This variation will have a strong enough impact on the blood vessels, which will either be dimmed or disappeared and will be clearly visible if continued on the next process. So gamma correction is very important in adjusting contrast to focus on the blood vessels by getting the desired intensity so that it has a good impact on the CLAHE process. CLAHE makes the image contrast change sharper than the gamma correction process can be seen in Figure 2.(c) so that the blood vessel area is dark while other than the veins are grayish and bright.

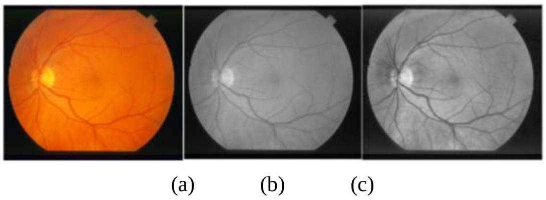
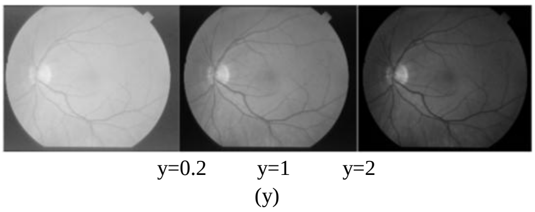


Fig. 2. (a) Original Image. [22]
(b) Gamma correction
(c) CLAHE
(d) Gamma correction with Various Value Gamma (y)



In Figure 3. (a) is a Butterworth Bandpass filter result. Butterworth bandpass filter is a combination of low pass and high pass filter. Low-pass filter itself as smoothing from an image while high-pass itself as an increase in the sharpness of the mind. Therefore this filtering is done to change the fundus image to be sharper and smoother by taking high and low frequency data to a certain extent so that the focus on the blood vessels is more clearly visible due to differences in color with strong density. In this Butterworth Bandpass filter using $D_L = 10$ and $D_H = 5000$ and n = 500 on the order filter, this is the standard value in this study. The variation in the value of the Butterworth Bandpass filter is as follows in Figure 3. (c). The influence that is very strong when using a value that is not in accordance with the testing standard is the result that it is easy to get more noise in the Thresholding process.

Then use Dynamic Threshold which functions to make certain parts brighter while the other parts become darker or change the image to binary can be seen in Figure 3. (b). The effect depends on the Butterworth bandpass filter process when the color of the retinal blood vessels is darker so the Threshold process will be more visible to the root end, therefore Gamma correction processes, CLAHE and Butterworth bandpass filters are very important in determining the shape, size and length of blood vessels retina.

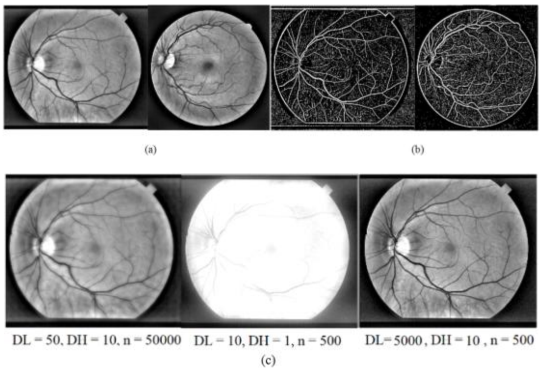


Fig 3. (a). Butterworth Bandpass filter at STARE (left) DRIVE (right). (b). Dynamic Threshold at STARE (left) DRIVE (right). (c). The variation in the value of the Butterworth bandpass filter at STARE

In the post-process it is important to make background deletions that are not needed. Focusing on the retinal vessels means removing the retinal vessels from the other side. Bwareopen image is useful for removing all connected components that are not needed from binary images or thresholds. This also has a Threshold dependency as well as the most major Gamma correction, CLAHE and Butterworth bandpass filters. Not only the role of Bwareopen, but the median filter is also very important as a smoothing of small dots so that the median filter can disappear noise by changing based on the middle pixel value. The illustration can be seen in Figure 4. (a) which is one of the final results of removing the background and connected small pixels assisted by the shape of the retina or mask to remove the background from the outside of the retina in Figure 4. (b) by doing reduction of the final result with a mask so that the final result is obtained in Figure 4. (a).

The results of image processing from DRIVE and STARE database in Table 1. In this case we will do a comparison with the ground truth DRIVE and STARE database. The measurement parameter value will be large if the retinal blood vessels approach in accordance with the dataset starting from the beginning of the retinal blood vessel to the end.

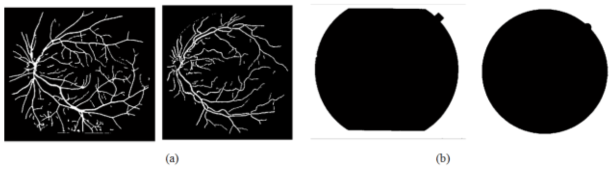
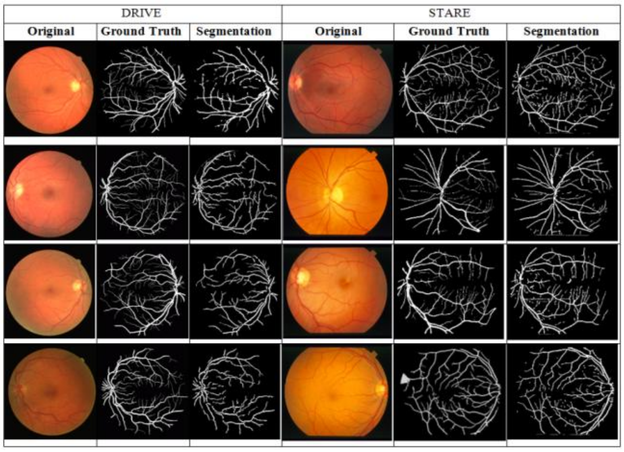


Fig. 4. (a). Morphology and Remove Background at STARE (left) DRIVE (right). (b) Retina Mask for remove background at STARE (left) DRIVE (right).

Table 1. Segmentation result from DRIVE and STARE databases



In measuring the parameters of a research performance is an important thing to do. This will illustrate how well the research is done by matching based on the retinal ground truth dataset. The parameter method used to measure data compatibility with datasets is Confusion Matrix based on binary images. Measurement Parameters can be seen in Table 2 as follows:

Table 2. Measurement Parameters

Parameters	Expresssion
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
Accuracy	$\frac{TP + TN}{TP + FN + TN + FP}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
F1-Measure	$2.\frac{Precision * Recall}{Precision + Recall}$

Where :

TP is True Positive, which is the amount of positive data that is matched based on the dataset correctly by the system.

TN is True Negative, which is the number of negative data that is matched based on the dataset correctly by the system.

FN is False Negative, which is the amount of negative data but is matched based on the wrong dataset by the system.

FP is False Positive, which is the amount of positive data but is matched based on the wrong dataset by the system.

Accuracy values are values that describe the correct matching of results and systems based on the dataset [23]. Specificity value describes the amount of positive category data matched based on the data set correctly divided by the total matching data based on positive data sets [23].

Sensitivity or recall can be defined as the ratio of retinal blood vessel pixels correctly classified in ground truth to the number of retinal blood vessel pixels [23].

Precision is the number of positive samples properly classified as a category divided by the total sample classified as a positive sample [24].

Based on the measurement parameters in Table 2, in this case will get the values of Accuracy, Sensitivity, Specificity Precision and F1 Measure for DRIVE and STARE databases in Table 3 as follows:

Table 3. Results Measurement parameters and Execution time from DRIVE and STARE databases

DRIVE							STARE						
File Name	**Measurement parameters**					**Exe. Time**	**File Name**	**Measurement parameters**					**Exe. Time**
	Acc	**Se (Recall)**	**Sp**	**Prec**	**F1**			**Acc**	**Se (Recall)**	**Sp**	**Prec**	**F1**	
01_test.tif	94.27	54.09	98.36	93.85	68.63	2.90 s	**im0001.ppm**	87.48	29.63	89.40	97.56	45.45	2.47 s
02_test.tif	94.39	54.82	98.41	92.84	68.94	1.93 s	**im0002.ppm**	85.48	28.07	87.64	90.96	42.90	2.43 s
03_test.tif	95.09	54.50	98.75	92.85	68.68	1.93 s	**im0003.ppm**	88.14	34.25	94.92	97.18	50.65	2.32 s
04_test.tif	94.64	54.05	98.57	87.69	66.88	1.94 s	**im0004.ppm**	89.83	28.70	90.51	86.63	43.12	2.46 s
05_test.tif	95.23	54.09	98.77	89.73	67.49	2.02 s	**im0005.ppm**	86.88	33.35	92.30	94.50	49.30	2.38 s
06_test.tif	95.31	56.42	98.06	82.70	67.08	1.92 s	**im0044.ppm**	88.73	35.39	94.66	97.96	52.00	2.64 s
07_test.tif	94.82	52.74	98.64	87.84	65.91	1.94 s	**im0077.ppm**	88.18	33.68	92.69	95.91	49.85	2.67 s
08_test.tif	94.78	54.51	98.37	84.08	66.14	2.00 s	**im0081.ppm**	86.85	29.52	91.34	95.91	45.14	2.47 s
09_test.tif	95.30	54.48	98.66	93.85	68.94	2.32 s	**im0082.ppm**	87.09	29.14	89.79	96.21	44.73	2.64 s
10_test.tif	95.02	54.33	98.61	86.95	66.87	1.84 s	**im0139.ppm**	85.85	28.66	92.21	88.47	43.29	2.45 s
11_test.tif	93.98	54.48	98.15	84.64	66.29	1.87 s	**im0162.ppm**	85.32	29.73	89.01	95.81	45.38	2.50 s
12_test.tif	94.62	52.82	98.41	87.15	65.77	1.93 s	**im0163.ppm**	87.08	31.79	89.23	93.76	47.48	2.45 s
13_test.tif	93.98	55.26	98.13	91.03	68.77	2.02 s	**im0235.ppm**	87.54	29.86	90.01	94.78	45.41	2.57 s
14_test.tif	94.80	55.63	98.22	86.04	67.57	2.02 s	**im0236.ppm**	88.22	31.13	80.02	99.39	47.41	2.61 s
15_test.tif	94.36	55.35	98.10	94.51	69.81	2.06 s	**im0239.ppm**	86.22	27.73	89.05	92.51	42.67	2.45 s
16_test.tif	94.54	56.27	98.17	91.34	69.64	1.98 s	**im0240.ppm**	87.64	24.35	95.20	95.20	38.78	2.54 s
17_test.tif	94.89	55.14	98.37	93.93	69.49	1.89 s	**im0255.ppm**	85.27	29.02	94.30	95.30	44.49	2.31 s
18_test.tif	95.10	54.12	98.50	87.04	66.74	2.05 s	**im0291.ppm**	90.56	25.73	94.95	98.65	40,81	2.41 s
19_test.tif	95.23	52.72	98.80	95.18	67.86	1.92 s	**im0319.ppm**	91.47	26.07	94.33	97.02	41,10	2.36 s
20_test.tif	95.23	53.78	98.71	90.06	67.34	2.15 s	**im0324.ppm**	89.80	24.94	93.10	92.08	39.25	2.37 s
Average	**94.77**	**54.48**	**98.43**	**89.67**	**67.74**	**2.03 s**	**Average**	**87.68**	**29.53**	**91.23**	**94.79**	**44.96**	**2.47 s**

The influence of the measurement parameter values when viewed in image processing originates from the input values at the stage carried out for example such as gamma values at the gamma correction stage, D*L*, D*H* and n values on Butterworth bandpass filters, median filters, bwareopen and threshold input values for dynamic threshold. For bwareopen and the median filter it is very influential on the value of the measurement parameters.

F1 score or F1 measure is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0 [24].

If the value of F1 measure that has been obtained and formed into a graph can be seen in Figure 5. for DRIVE and Figure 6. for STARE.

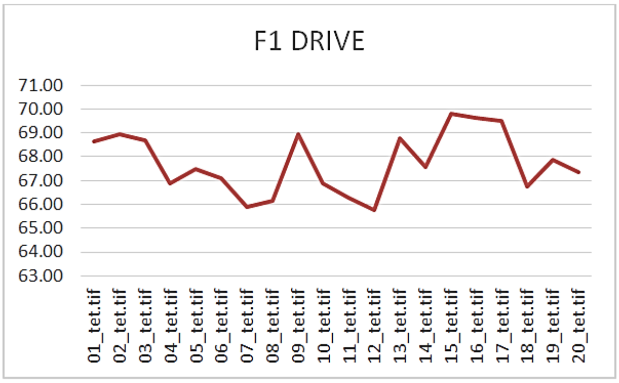


Fig. 5. F1 Measure for DRIVE

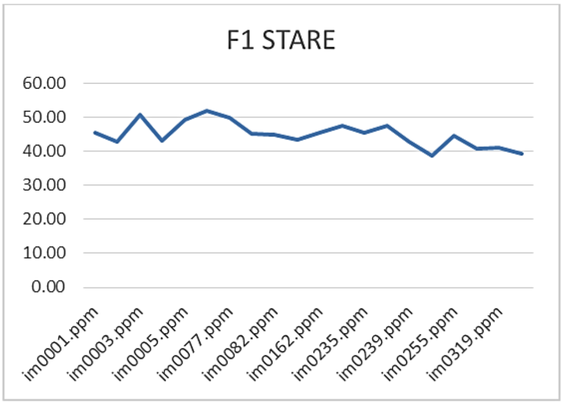


Fig. 6. F1 Measure for STARE

In Figure 5. F1 measure for DRIVE it can be seen that the highest F1 measure value is owned by the file name 15_test.tif while the lowest F1 measure is owned by the file name 12_test.tif. at Figure 6 F1 measure for STARE it can be seen that the highest value of F1 measure is possessed by file name im0044.ppm while the lowest F1 measure is owned by file name im0240.ppm. If seen from the average value of F1 measure the DRIVE value is 67.74 which is better than STARE which is 44.96.

Based on Table 3, the measurement parameter values obtained for DRIVE databases with an accuracy of 94.77%, Sensitivity 54.48%, and Specificity 98.43%. For STARE databases with an accuracy of 87.68%, Sensitivity 29.53%, and Specificity 91.23% so that in this case the values of the measurement parameters in Table 4 can be compared as follows:

Table 4. Results Segmentation comparison from DRIVE and STARE databases between previous methods.

Method		Acc	Se	Sp
DRIVE	- Extended Matched Filter Based on Second Derivative of Gaussian [1]	93.74%	-	-
	- Mathematical Morphology [3]	92%	64%	95%
	- Contrast Enhancement by Top-Hat and Bottom-Hat Transform with Optimal Strel [25]	93.77%	62.96%	98.30%
	- Green Channel, masking, and filtering (top-hat and median filters) [19]	94.02%	77.08%	99.01%
	- The Elite-guided Multi-Objective Artificial Bee Colony Algorithm [11]	94.5%	73.9%	97.4%
	- Dynamic Threshold and Enhancement Image Filter From Retina Fundus	**94.77%**	**54.48%**	**98.43%**
STARE	- Extended Matched Filter Based on Second Derivative of Gaussian [1]	89,31%	-	-
	- Mathematical Morphology [3]	89%	76%	89%
	- Contrast Enhancement by Top-Hat and Bottom-Hat Transform with Optimal Strel [25]	-	-	-
	- Green Channel, masking, and filtering (top-hat and median filters) [19]	-	-	-
	- The Elite-guided Multi-Objective Artificial Bee Colony Algorithm [11]	94%	73.2%	96.2%
	- Dynamic Threshold and Enhancement Image Filter From Retina Fundus	**87.68%**	**29.53%**	**91.23%**

Based on Table 4 the comparison of parameter values with the previous methods that the values in the proposed method are better than the previous methods, both the parameters of accuracy and specifity. Even so for the value of sensitivity is still smaller than the previous methods. But for the STARE database it still has values are smaller than previous methods on parameters measuring accuracy, sensitivity, and specificity.

Based on the execution time contained in Table 3 in which the implementation time stamp for each file so for average values obtained for DRIVE is Execution Time 2.03 seconds and for STARE is Execution Time 2.47 seconds, it can be compared with the time of implementation and the system requirements of device that used in previous methods. Here Comparison Execution Time and System requirements of device with previous methods in Table 5.

Table 5. Comparison Execution Time and System requirements of device from DRIVE and STARE databases with previous methods.

Method		Exe. Time	System
DRIVE	- Extended Matched Filter Based on Second Derivative of Gaussian [1]	2.61 s	1.65 GHZ, 2 GB RAM
	- Mathematical Morphology [3]	-	-
	- Contrast Enhancement by Top-Hat and Bottom-Hat Transform with Optimal Strel [25]	2 - 3 s	1.8 GHZ, 3 GB RAM
	- Green Channel, masking, and filtering (top-hat and median filters) [19]	1.3 s	3.7 GHZ, 8 GB RAM
	- The Elite-guided Multi-Objective Artificial Bee Colony Algorithm [11]	2.21 s	2.6 GHZ, 4 GB RAM
	- Dynamic Threshold and Enhancement Image Filter From Retina Fundus	**2.03 s**	**1.4 GHZ, 2 GB RAM**
STARE	- Extended Matched Filter Based on Second Derivative of Gaussian [1]	2.40 s	1.65 GHZ, 2 GB RAM
	- Mathematical Morphology [3]	-	-
	- Contrast Enhancement by Top-Hat and Bottom-Hat Transform with Optimal Strel [25]	2-3 s	1.8 GHZ, 3 GB RAM
	- Green Channel, masking, and filtering (top-hat and median filters) [19]	-	3.7 GHZ, 8 GB RAM
	- The Elite-guided Multi-Objective Artificial Bee Colony Algorithm [11]	3.14 s	2.6 GHZ, 4 GB RAM
	- Dynamic Threshold and Enhancement Image Filter From Retina Fundus	**2.47 s**	**1.4 GHZ, 2 GB RAM**

Based on a Table 5 at the Execution time of the proposed method has a section DRIVE database has small value compared with the previous method with the system requirements device are too low, the device Celeron 2957U processor 1.4 GHz Dual core laptop, Intel HD graphics with 2 GB of RAM also for STARE compared with previous method has enough value Execution time. Although there is a smaller time [19] but used system requirment device core i3 6th gen CPU 3.7 GHz device, 8 GB of RAM.

5. Conclusion

The extraction produced in this paper is quite good starting with the pre-process to obtain performance accuracy with image enhancement, filtering, color change, etc. and then optimized post-process. Retinal image processing experiments available for data set blood vessel extraction from the DRIVE and STARE database using accuracy, sensitivity, and specificity parameters. Thus the proposed method achieves average value of measurement parameters from DRIVE database is accuracy of 94.77 percent, sensitivity of 54.48 percent, specificity of 98.71 percent, and execution time of 2.03 seconds, and STARE database is accuracy of 87.68 percent, sensitivity of 29.53 percent, specificity of 91.23 percent, and execution time of 2.47 seconds. For DRIVE database, the comparison of parameter values with previous methods that the values in the proposed method are better than the previous methods, both the accuracy and specificity parameters. Even though the sensitivity value is still smaller than previous methods, and the STARE database still has values smaller than previous methods for parameters measuring accuracy, sensitivity, and specificity. When viewed in image processing, the influence of the measurement parameter values originates from the input values at the stage, such as gamma values at the gamma correction stage, DL, DH and n values on Butterworth bandpass filters, median filters, bwareopen and dynamic threshold input values. The value of the measurement parameters is very influential for bwareopen and the median filter. However, this method still has no match between the results of processing with the ground truth or the causes of reference such as noise and the root roots of lost or increased blood vessels.

Acknowledgement

This paper is partly supported by Rector of University of Sriwijaya. The author would like to thank United States Agency for International Development (USAID) for the training and mentoring support in writing this article, through the Sustainable Higher Education Research Alliance (SHERA) Program for Universitas Indonesia's Scientific Modelling, Application, Research and Training for City-centred Innovation and Technology (SMART CITY) Project.

References

[1] N. Pratap and S. Rajeev, “Extraction of Retinal Blood Vessels by Using an Extended Matched Filter Based on Second Derivative of Gaussian,” Proc. Natl. Acad. Sci. India Sect. A Phys. Sci., 2018.

[2] A. Biran, P. S. Bidari, A. Almazroa, and K. Raahemifar, “Blood Vessels Extraction from Retinal Images Using Combined 2D Gabor Wavelet Transform with Local Entropy Thresholding and Alternative Sequential Filter,” IEEE Can. Conf. ECE, pp. 1–5, 2016.

[3] A. L. Pal, S. Prabhu, and N. Sampathila, “Extraction of Retinal Blood Vessels from Retinal Fundus Image for Computer Aided Diagnosis,” Canar. E.college, pp. 400–403, 2015.

[4] Z. Yavuz and C. Köse, “Blood Vessel Extraction in Color Retinal Fundus Images with Enhancement Filtering and Unsupervised Classification,” Hindawi Int. J., vol. 2017, 2017.

[5] J. Dash, “Retinal Blood Vessels Extraction from Fundus Images Using an Automated Method,” 2018 4th Int. Conf. Recent Adv. Inf. Technol., pp. 1–5, 2018.

[6] J. Dash and N. Bhoi, “An Unsupervised Approach for Extraction of Blood Vessels from Fundus Images,” J. Digit. Imaging, 2018.

[7] D. Güu, “A Novel Retinal Vessel Extraction Method Based on Dynamic Scales Allocation,” Int. Conf. image, Vis. Comput., pp. 145–149, 2017.

[8] T. A. Soomro, “Retinal Blood Vessel Extraction Method Based on Basic Filtering Schemes,” IEEE Int. Conf. Image Process., 2018.

[9] M. Ben Abdallah et al., “Automatic Extraction of Blood Vessels in the Retinal Vascular Tree Using Multiscale Medialness,” Hindawi Int. J., vol. 2015, 2015.

[10] R. Kamble, “Automatic Blood Vessel Extraction Technique Using Phase Stretch Transform In Retinal Images,” 2016 Int. Conf. signal Inf. Process., 2017.

[11] B. Khomri, A. Christodoulidis, L. Djerou, and M. C. Babahenini, “Retinal blood vessel segmentation using the elite-guided multi-objective artificial bee colony algorithm,” IET Image Process., vol. 12, pp. 2163 – 2171, 2018.

[12] S. A. Amiri, “A Preprocessing Approach For Image Analysis Using Gamma Correction,” vol. 38, no. 12, 2012.

[13] F. K. P, D. Saepudin, and A. Rizal, “Analisis Contrast Limited Adaptive Histogram Equalization (Clahe) Dan Region Growing Dalam Deteksi Gejala Kanker Payudara Pada Citra Mammogram,” elektro, vol. 9, pp. 1–14, 2014.

[14] A. L. I. M. Reza, “Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement,” J. VLSI Signal Process., vol. 38, pp. 35–44, 2004.

[15] D. Govind, B. Ginley, and B. Lutnick, “Glomerular detection and segmentation from multimodal microscopy images using a Butterworth band-pass filter,” SPIE Med. Imaging, vol. 1058114, no. March, 2018.

[16] R. C. Gonzalez, Digital Image Processing Third Edition. 2006.

[17] Diptoneel Kayal Sreeparna Banerjee, “Dynamic Thresholding with Tabu Search for Detection of Hard Exudates in Retinal Image,” in Industry Interactive Innovations in Science, Engineering and Technology, 2017, pp. 553–560.

[18] Z. H. Chan FH, Lam FK, “Adaptive thresholding by variationalmethod,” IEEE Trans. Image Process., vol. 7, pp. 468–473, 1998.

[19] A. Ray, A. Chakraborty, D. Roy, B. Sengupta, and M. Biswas, “Blood Vessel Extraction from Fundus Image,” Emerg. Technol. Data Min. Inf. Secur., pp. 259–268, 2018.

[20] Michael Goldbaum, “STARE database,” 2003. [Online]. Available: http://cecas.clemson.edu/~ahoover/stare/.

[21] J. J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, “Ridge based vessel segmentation in color images of the retina,” IEEE Trans. Med. Imaging, vol. 23, no. 4, pp. 501–509, 2004.

[22] M. D. Michael Goldbaum, “STructured Analysis of the Retina,” 1975. [Online]. Available: http://cecas.clemson.edu/~ahoover/stare/.

[23] H. B. Wong and G. H. Lim, “Measures of Diagnostic Accuracy: Sensitivity , Specificity , PPV and NPV,” in

Proceedings of Singapore Healthcare, 2011, vol. 20, no. 4, pp. 316–318.

[24] D. M. W. Powers, “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation,” J. Mach. Learn. Technol., pp. 37–63, 2011.

[25] R. K. B, H. Kabir, and S. Salekin, “Contrast Enhancement by Top-Hat and Bottom-Hat Transform with Optimal Structuring Element : Application to Retinal Vessel Segmentation,” Springer Int. Publ. AG 2017, pp. 533–540, 2017.

Erwin was born in Palembang, Indonesian, in 1971. He received the Bachelor degree in Mathematics from the University of Sriwijaya, Indonesian, in 1994, and the M.Sc. degrees in Actuarial from the Bandung Institute of Technology (ITB), Bandung, Indonesian, in 2002. He recently completed his Ph.D. degrees in 2019 in Informatics Engineering at University of Sriwijaya. In 1994, he joined, University of Sriwijaya, as a Lecturer. Since December 2006, he has been with the Department of Informatics Engineering, University of Sriwijaya, where he was an Assistant Professor, became an Associate Professor in 2011. Since 2012, he has been with the Department of Computer Engineering, University of Sriwijaya. His current research interests include image processing, and computer vision.

Tomi Kiyatmoko was born in Gelumbang, Indonesian, in 1996. He has been accepted as a student college since 2015 and became part of the Department of Computer Engineering University of Sriwijaya, Indonesia. He is presently working on a project for his undergraduate degree at the Department of Computer Engineering, Faculty of Computer Science, University of Sriwijaya. His research interests included image processing, computer vision, and pattern recognition.

A 2-bit Full Comparator Design with a Minimum Quantum Cost Function in Quantum-Dot Cellular Automata

Davoud Bahrepour *
Department of Computer, Mashhad Branch, Islamic Azad University, Mashhad, Iran
bahrepour@mshdiau.ac.ir
Negin Maroufi
Department of Computer, Khorasan Razavi, Neyshabur, Science and Research branch, Islamic Azad University, Neyshabur, Iran
Maroofi.negin@mail.um.ac.ir

Received: 08/Nov/2018 Revised: 08/May/2019 Accepted: 16/Jul/2019

Abstract

In recent years, reduction of the complementary metal-oxide-semiconductor (CMOS) circuit's feature size has posed significant challenges, such as current loss and leakage and high power consumption. Consequently, further size reduction of CMOS technology is not feasible. As an emerging nanoscale technology, quantum-dot cellular automata (QCA) can be utilized in the near future for designing computers and very-large-scale integration (VLSI) circuits. QCA technology makes it possible to design low-power, high-performance, and area-efficient logical circuits. A comparator function is a digital logical function which compares and evaluates whether or not a bit is greater than, smaller than or equal to the other bit (half comparator). A full comparator has a third input which shows the result of the previous step. Half and full comparators play an essential role in CPU architecture. The current paper proposes a full comparator circuit based on QCA and a new quantum cost function. In addition, a 2-bit comparator is presented based on the introduced full comparator. Employing the new quantum cost function, the present study compares the proposed full comparator design with previously presented designs in terms of area, delay, and complexity. Comparisons show that the proposed design occupy less area and produces less delay and so is more suitable for usage in CPU design.

Keywords: Quantum-Dot Cellular Automata; Full Comparator; Cost Function; QCA Cell; Majority Gate; NOT Gate.

1. Introduction

In recent years, complementary metal-oxide-semiconductor (CMOS) technology has encountered different challenges, such as high power consumption and current leakage. Minimizing CMOS circuits involves particular problems, including the occurrence of various physical phenomena, namely a particular mass of each element and quantum effects, all of which pose obstacles to proper operation of transistors. Consequently, researchers are seeking smaller technologies with lower power consumption and less current leakage [1-3].

Quantum- dot cellular automata (QCA) is considered to be among the six emerging technologies with higher performance. Even though QCA circuits are associated with more challenges than are faced by CMOS, the simple structure of QCA circuits has led researchers to further study their implementation. With its low-power, high-performance, and area-efficiency features, QCA technology can redesign fundamental circuits, such as comparators. As a result, the circuit industry is impacted by the need to propose chip designs that are more practical and provide better performance in QCA technology [4-8].

Section 2 reviews the fundamental blocks in QCA technology and its clocking. Section 3 covers the background of comparators and provides an overview and comparison of QCA designs. Section 4 proposes a 2-bit QCA comparator design while Section 5 introduces a new cost function for the evaluation of QCA circuits. Based on this cost function, the proposed design is then compared to previous designs.

2. Background

This section first provides an overview of the QCA cell structure, QCA wire, and concept of the clock in QCA circuits. The NOT, three-input majority, and five-input majority gates in QCA technology are then introduced.

2.1 QCA Cell

In QCA technology, each cell consists of four dots and two electrons which can move freely between the holes or dots. The two electrons have six different states to fill the dots, but not all of these states are stable. The electrons are as far as possible from each other due to Coulomb’s repulsive force (i.e., electrostatic gravity and repulsion). As stated earlier, stable states appear when the electrons fill the dots diagonally, which is known as polarization. These two states (Figure 1) demonstrate -1 and +1 polarizations, which are assigned the logical values of 0 and 1, respectively [9,10].

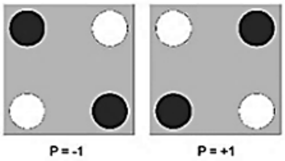


Fig. 1. Two stable states of the basic QCA cell (the cell on the right logical 1; the cell on the left logical 0)

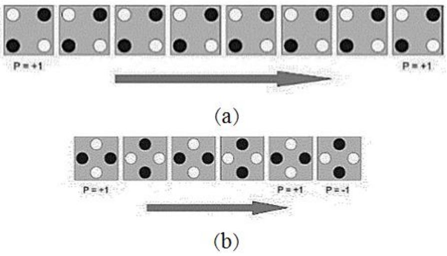
* Corresponding Author

2.2 Wire Structure in QCA

Coulomb’s repulsive force not only works between the electrons in one cell, but it also causes the electrons in each cell to affect the electrons in adjacent cells. The difference between the two is that, when two cells are considered as neighbors, the electrons are placed in a certain state so as to minimize Coulomb’s repulsive force as much as possible. Moreover, an array of lateral cells can be utilized as a wire to propagate information. Figure 2 presents two different models of QCA wires. In the second model (complement chain), the cells are rotated 45 degrees so that the input signal propagates in the odd cells and its complement propagates in the even cells. Placing these two wire models on each other creates a crossing wire model (Figure 3). Due to the difference in cell polarization in the crossing wire model, the two wires do not affect each other [1, 11].

2.3 QCA Clock

Because of QCA's structural features, the clock acts as an electronic factor for controlling the movement of electrons within a cell. The clock synchronizes the different parts of the circuit.



(a)

(b)

Fig. 2. (a) QCA standard wire, (b) QCA complement wire

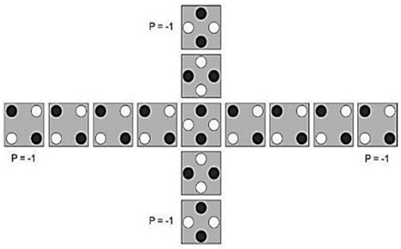


Fig. 3. QCA crossing wire model

Each clock cycle in QCA has four phases (Figure 4): switch, hold, release, and relax. In the switch phase, the polarization of each cell is affected by the adjacent cells. The hold phase places the electrons at the maximum distance from each other so that they enter the stable state. Cells in the hold phase are able to detect the polarization of the adjacent cells in their switch phase. In the release phase, electrons are gradually released and the barrier force declines. In the relax phase, there is no polarization and the electrons may move freely inside a cell [1].

One of the critical issues in QCA clock design is the setting of the four areas in each clock cycle. In fact, misallocation of each clock cycle's areas causes errors in circuit operation. In QCA clock design, coincidences in entering the inputs, wire length, and number of cells in each phase should be considered. Evidently, data flow control must also be specifically examined in the structural design of QCA and increasing the number of cells in each clock phase achieves this. To prevent noise, a minimum of two cells in each clock phase is necessary. In addition, there should be a threshold value for the maximum number of cells in each clock phase due to the consequences of increasing the number of cells in each phase. Not only does the clock frequency decrease by doing so, but some of the cells may also enter uncertain states due to specified limitations on the energy required to polarize each cell [1,2,8].

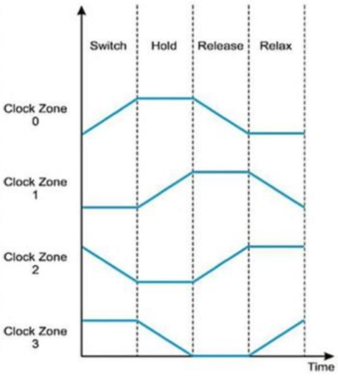
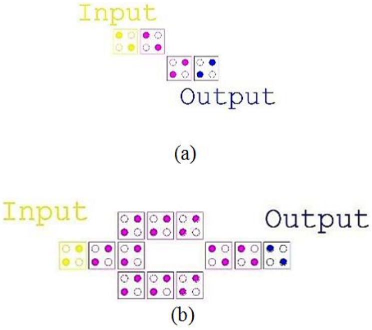


Fig. 4. Clock phase in QCA

2.4 Not Gate

The inverter is one of the two fundamental building blocks of each QCA circuit. Figure 5-a presents the basic and simple QCA inverter gate and Figure 5-b provides another NOT gate design, in which the signal enters from the left side and divides into two QCA wires. These signals then merge on the right side. The complement of the entered signal is calculated at the merge time and released on the right side. As a result of Coulomb’s repulsive force in this structure, the stable state of the output is the complement of the input, which thus places the electrons at the maximum distance from each other [1,3,9].



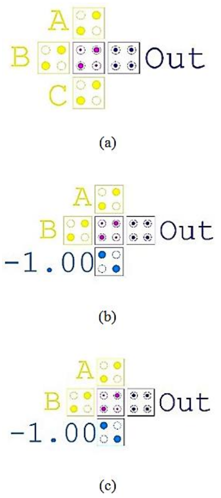
(a)

(b)

Fig. 5. Two different designs of NOT gate in QCA

2.5 Majority Gate

Another fundamental building block of a QCA circuit is the majority gate. Since the majority gate is programmable, it can be used for designing different digital logic structures. As depicted in Figure 6-a, the three-input majority gate has three inputs, an output, and a work cell. The work cell is polarized according to the majority polarizations of the cells, as well as the repulsive force among the three input cells [1].



(a)

(b)

(c)

Fig. 6. (a) Three-input majority gate in QCA, (b) Two-input AND gate in QCA, (c) Two-input OR gate in QCA

The logical function of this gate is as follows [1]:

$$M(A, B, C) = AB + AC + BC \quad (1)$$

According to the logical function of the majority gate, if the constant value of -1 (logical zero) is assigned to the C input, it operates as the two-input AND gate (Figure 6-b).

$$M(A, B, 0) = AB + (A)(0) + (B)(0) = AB \quad (2)$$

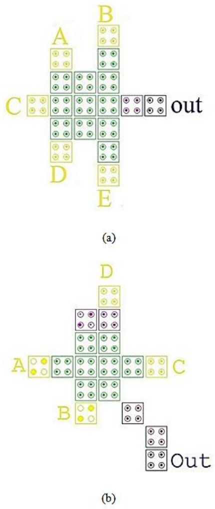
Moreover, if the constant value of +1 (logical 1) is assigned to the C input, it operates as the two-input OR gate (Figure 6-c).

$$M(A, B, 1) = AB + (A)(1) + (B)(1) = A + B \quad (3)$$

A full comparator design utilizes two different designs of the five-input majority gate. Equation 4 presents the five-input majority gate [1, 14].

$$M(A, B, C, D, E) = ABC + ABD + ABE + ACD + ACE + ADE + BCD + BCE + BDE + CDE \quad (4)$$

Figure 7 provides both designs of the five-input majority gate [12, 14].



(a)

(b)

Fig. 7. (a) Five-input majority gate in QCA with 18 cells [15], (b) Five-input majority gate in QCA with 17 cells [12, 13]

As shown in Figure 7-b, A, B, C, and D are the inputs of the gate. In this design, input D acts as two similar inputs. In fact, the two inputs are linked and considered as input D. This majority gate consists of 17 cells [12, 13].

3. Comparator Background

Comparators are one of the essential parts of digital logic circuits and have been widely used in CPUs and microcontrollers. Consequently, any progress made in circuit design will improve CPU performance [16-18]. The half-comparator function compares two inputs and produces the result as an output. This output specifies whether a bit is greater than, smaller than or equal to the other bit. Equation (5) depicts the logical function of a half-comparator.

$$\begin{aligned} F_{A>B} &= A\overline{B} \\ F_{A<B} &= \overline{A}B \\ F_{A=B} &= \overline{F_{A>B}}\,.\,\overline{F_{A<B}} \end{aligned} \quad (5)$$

where A and B are the inputs and $F_{A>B}$, $F_{A<B}$, and $F_{A=B}$ are the outputs [12]. In full comparators, there is a third input (C) which maintains the result of the previous step at each stage. Equation (6) provides the full comparator functions. Clearly, by assigning a constant value of 1 to input C, the full comparator operates as a half-comparator [12, 16-18].

$$\begin{aligned} F_{A>B} &= A\overline{B}C \\ F_{A<B} &= \overline{A}BC \\ F_{A=B} &= \overline{F_{A>B}}\,.\,\overline{F_{A<B}} \end{aligned} \quad (6)$$

In 2008, Y. Xia and K. Qiu proposed the full comparator shown in Figure 8. In this design, a Universal Logic Gate (ULG) is used to build the full comparator and the ULG circuit can operate as any n-input function [15]. With the utilization of ULG, this design exhibits a better performance than that of previous designs employing majority and inverter gates (MI) (Figure 9) [15].

As depicted in Figures 8 and 9, the full comparator design utilizing ULG has fewer crossing wires and its number of cells increased compared to the full comparator design using MI. As a result, circuit performance improved [15].

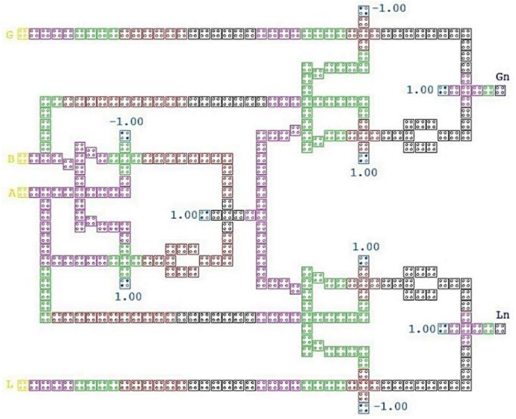


Fig. 8. Full comparator circuit based on QCA utilizing ULG gate [15]

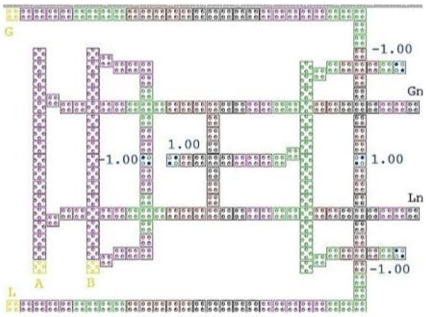


Fig. 9. Design of full comparator in QCA utilizing MI [15]

In 2014, S. S. Anuradha et al. proposed a new full comparator circuit [14]. This circuit introduced a new design for the five-input majority gate (see Figure 7) that was more fault-tolerant in comparison with previous designs. Figure 10 presents the full comparator based on the five-input majority gate. It is notable that the two inputs have a constant value of zero.

As demonstrated in Figure 10, two five-input majority gates and one three-input majority gate were employed in designing this full comparator. There is a significant reduction in the complexity and number of cells in comparison with previous full comparator designs [14].

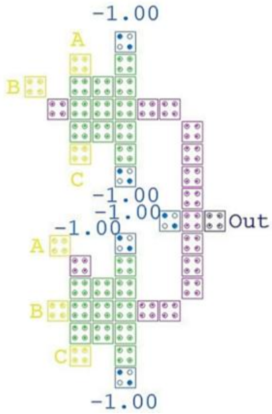
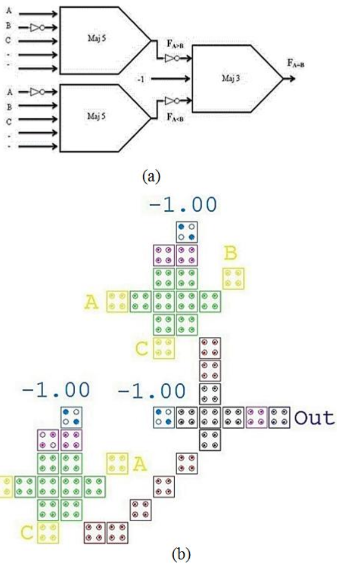


Fig. 10. Full comparator based on the five-input majority



(a)

(b)

Fig. 11. (a) Schematic (b) Circuit of a full comparator using the five-input majority gate [12]

Figure 11 presents another design for the full comparator with the minimum number of cells in its circuit [12]. This design features the five-input majority gate with 17 cells (Figure 7-b).

This full comparator features two five-input majority gates and one three-input majority. Figure 11-b shows that A and B are the inputs, C is the result of the comparison in the previous stage, and the output is one of three states, namely $\mathrm{F_{A>B}}$, $\mathrm{F_{A<B}}$ or $\mathrm{F_{A=B}}$. Evidently, if input C takes the value of 1, the circuit changes to the half-comparator state. This circuit simultaneously produces two outputs ($\mathrm{F_{A>B}}$ and $\mathrm{F_{A<B}}$) and output $\mathrm{F_{A=B}}$ is produced 0.25 clock cycles later [12].

Figure 12 provides the simulation results of the full comparator. According to the simulation, the first two signals represent inputs A and B and the third signal is input C, which is assigned the constant value of zero. In addition, the fourth signal is the output. As shown, the delay of this full comparator is equal to 1.25 clock cycles. In fact, this full comparator has the minimum number of cells as well as less delay in comparison to the other designs.

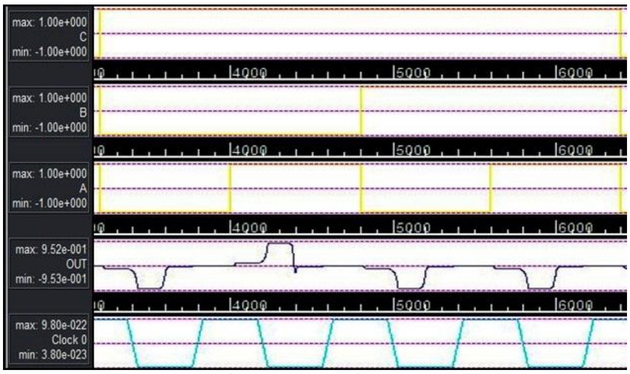


Fig. 12. Simulation of the proposed full comparator using QCA designer software

4. Proposed 2-bit Comparator

To compare inputs with a higher number of bits, efficient comparators are needed to evaluate and determine whether the two strings of bits are lower than, greater than, or equal to each other. For this purpose, the present study proposes a 2-bit comparator based on the comparator in [12].

Figure 13 presents the block diagram of the proposed model.

As seen, three 1-bit full comparators compare the two bits. The bits with the same value for each number enter the 1-bit comparators at the same time. Then, in the next stage, the results propagate through the 1-bit full comparator. Figure 14-a and Figure 14-b show the schematic and QCA design respectively. Figure 15 provides the simulation results of the proposed 2-bit comparator. The output of the circuit clearly indicates whether the 2-bit numbers are equal or not. The delay of the proposed circuit is equal to one clock cycle.

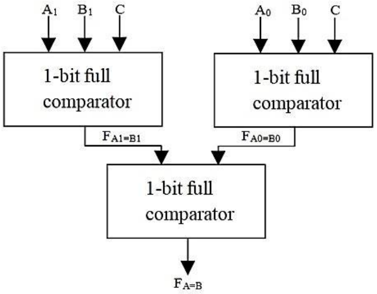
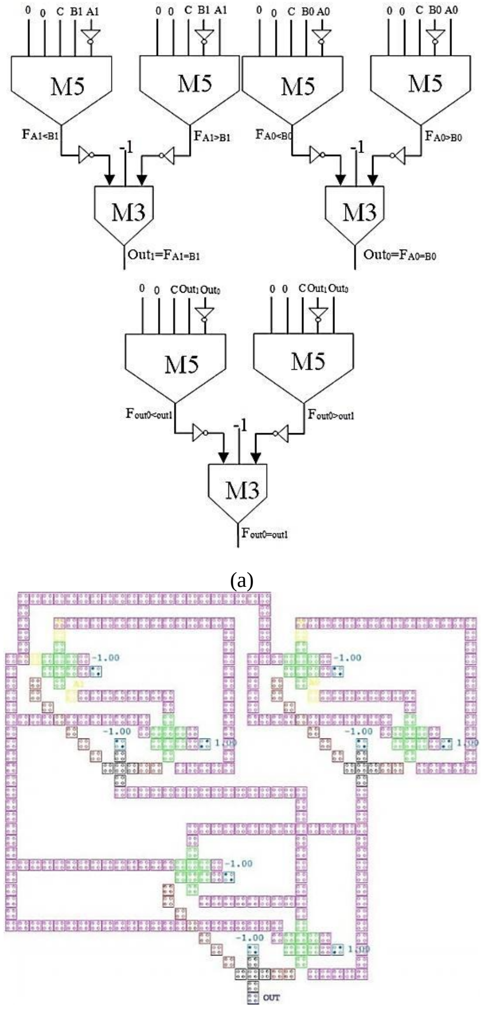


Fig. 13. Block diagram of the proposed 2-bit comparator based on [12]



(a)

(b)

Fig. 14. (a) Schematic (b) QCA design of the proposed 2-bit comparator based on [12]

5. Comparison of Full Comparators by the Cost Function

It is very common to introduce a function or an index for evaluating different designs within the same technology as this assists designers to develop better circuits and to compare their design with previous ones [19].

This section proposes a quantum cost function according to QCA features. Then, by the employment of this function, the cost of each full comparator design is calculated.

[17] and [20] introduce a function to evaluate the cost and quality of a QCA circuit. This function shows that the area occupied by a QCA circuit plays a key role in evaluating circuits. One of the advantages of QCA over earlier technologies is its small size. The complexity of a QCA-based circuit is determined based on the number of cells used in the circuit. As the number of cells increase, the polarization of each cell depends on more cells. Furthermore, in most cases, as the number of cells rises, the crossing wires become efficient and so a layered design is beneficial. All of the above points cause the design to occupy more area. The proposed cost function has a direct correlation with the area of a circuit. Another parameter in the evaluation of QCA circuits is power. Lower power is associated with less power dissipation and so it can be claimed that the lower power of QCA circuits produces a better design. Undoubtedly, power has a direct relationship with the cost function as well. Another parameter considered in the evaluation of circuit design is the delay of a circuit. Circuit delay is associated with the complexity of a circuit in certain aspects. Maximum delay occurs when the longest path from the input to the output (i.e., critical path) includes more cells. On the other hand, less delay in the critical path indicates a better design and cost function [17, 20-21].

Equation (7) presents the cost function [21].

$$\mathrm{Cost} = \mathrm{Area} \times \mathrm{Power} \times \mathrm{Delay} \tag{7}$$

Reference [21] indicates that power has a direct and consistent association with complexity.

$$\mathrm{Power} \equiv \mathrm{Complexity} \tag{8}$$

The two parameters affect each other; if complexity increases, power will also grow and vice versa. Overall energy dissipation of a circuit is equal to the amount of dissipated energy in each cell. The energy dissipation of each cell is approximately the same in all cells. Therefore, the amount of energy dissipation in each cell, which is considered to be a constant value for all QCA cells, can be overlooked [17]. To express the overall power consumption in a circuit, the number of cells may be sufficient, eliminating the need to multiply the dissipated energy in each cell by the number of cells [21]. In addition, the power consumption in QCA circuits is so low that it can be ignored.

Thus, another parameter affecting the performance of a QCA circuit can replace that of power. As a result, according to Equation (8), a new cost function is introduced as follows:

$$\mathrm{Cost} = \mathrm{Area} \times \mathrm{Complexity}\ \times \mathrm{Delay} \tag{9}$$

Therefore, it can be claimed that Equation (9) is another and easier method for correctly evaluating the cost function based on the complexity of a QCA circuit rather than the previously mentioned cost function.

In the following, all the previously full comparators are simulated with QCA Designer software, version 2.0.2, and the cost function is calculated for each of them as well. Moreover, the Bistable simulation engine is employed in the simulation and Table 1 specifies the set values.

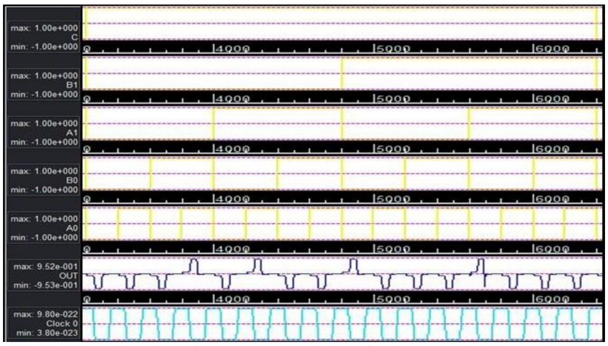


Fig. 15. Simulation of the proposed 2-bit full comparator using QCA designer software

Table 1. Parameters in the QCA Designer Bistable Simulation Engine

Number of Samples	12800
Convergence Tolerance	0.0010
Radius of Effect (nm)	65.0
Relative Permittivity	12.9
Clock High	9.80e-022
Clock Low	3.80e-023
Clock Shift	0.0e+0
Clock Amplitude Factor	2.0
Layer Separation	11.5
Maximum Iterations Per Sample	100
Cell Size (nm)	18×18
Cell Distance (nm)	2
Quantum-Dot Diameter (nm)	5

Table 2 provides the comparison of the full comparator circuits in terms of area, complexity, delay, and cost.

Table 2. Comparison of Full Comparator Circuit Designs

Proposed Design	Area (μm2)	Complexity (number of cells)	Delay (clock cycle)	Cost
ULG [15]	0.65	353	2.25	516.2625
MI [15]	0.29	222	2	128.76
Five-input Majority Gate-based Full Comparator [14]	0.09	48	1.25	5.4
Proposed Full Comparator	0.08	43	1.25	4.3

According to the information in Table 2, there is a significant difference between the full comparator cost introduced in [15] that utilizes ULG and that in [15] using MI. This may be due to the number of cells, as well as the smaller area, when compared to the MI design. Although the delay in [14] and in the proposed design are the same, their area and number of cells differ, thereby resulting in a better cost function. As the cost function considers all the important parameters in QCA-based full comparator designs, it is a favorable benchmark for the comparison of various designs.

6. Conclusion

Comparator circuits play a pivotal role in computational operations and are widely used in designing microcontrollers and CPUs. The current paper introduced a full comparator circuit and also presented a cost function to compare the proposed design with other designs. The proposed cost function can serve as a proper benchmark for the overall comparison of different designs. Furthermore, the results of the simulation indicated that the proposed full comparator features the best area and complexity (i.e., number of cells) as well as the best cost value. Therefore, it can be inferred that the introduced full comparator is optimal.

Acknowledgments

This study was partially supported by Islamic Azad University, Mashhad Branch, Mashhad, Iran. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the host institution or funders.

References

[1] J. C. Das, and D. De, "Novel low power reversible binary incrementer design using quantum-dot cellular automata. Microprocessors and Microsystems", vol. 42, 2016, pp. 10-23.

[2] R. Jayalakshmi, and R. Amutha, "A Theoretical Study on the Implementation of Quantum Dot Cellular Automata", Fourth International Conference on Advances in Electrical, Electronics, Information, Communication, and Bio-Informatics (AEEICB), Chennai, 2018, pp. 1-6.

[3] R. Sarma, and R. Jain, "Quantum Gate Implementation of a Novel Reversible Half Adder and Subtractor Circuit," International Conference on Intelligent Circuits and Systems (ICICS), Phagwara, 2018, pp. 72-76.

[4] M. M. Rahman, N. M. Nahid, and M. K. Hassan, "Energy dissipation dataset for reversible logic gates in quantum dot-cellular automata" Data in brief, vol. 10, 2017, pp. 557-560.

[5] M. Walter, R. Wille, D. Grobe, F. S. Torres, and R. Drechsler, "An exact method for design exploration of quantum-dot cellular automata," Design, Automation and Test in Europe Conference and Exhibition (DATE), Dresden, 2018, pp. 503-508.

[6] V. S. Kalogeiton, D. P. Papadopoulos, O. Liolis, V. A. Mardiris, G. C. Sirakoulis, and I. G. Karafyllidis, "Programmable Crossbar Quantum-Dot Cellular Automata Circuits," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 36, no. 8, 2017, pp. 1367-1380.

[7] G. Cocorullo, P. Corsonello, F. Frustaci and S. Perri, "Design of Efficient BCD Adders in Quantum-Dot Cellular Automata," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 64, no. 5, 2017, pp. 575-579.

[8] M. Mohammadi, S. Gorgin and M. Mohammadi, "Design of non-restoring divider in quantum-dot cellular automata technology," IET Circuits, Devices and Systems, vol. 11, no. 2, 2017, pp. 135-141.

[9] K. Walus, T. J. Dysart, G. A. Jullien, and R. A. Budiman, "QCADesigner: A rapid design and simulation tool for quantum-dot cellular automata," IEEE transactions on nanotechnology, vol. 3, no. 1, 2004, pp. 26-31.

[10] D. Bahrepour, and J. Forouzanfar, "A Novel Robust Macrocell Based on Quantum Dot Cellular Automata. Quantum Matter," vol. 5, no. 5, 2016, pp. 689-696.

[11] P. D. Tougaw, and C. S. Lent, "Logical devices implemented using quantum cellular automata," Journal of Applied Physics, vol. 75, no. 3, 1994, pp. 1818-1825.

[12] D. Bahrepour, "A Novel Full Comparator Design Based on Quantum-Dot Cellular Automata. International Journal of Information and Electronics Engineering," vol. 5, no. 6, 2015, pp. 406.

[13] S. Hashemi, M. Tehrani, and K. Navi, "An efficient quantum-dot cellular automata full-adder. Scientific Research and Essays," vol. 7, no. 2, 2012, pp. 177-189.

[14] S. S. Anuradha, B. D. Ravi, and M. Pasar Vishal, "Design of five input majority gate full comparator using Quantum-Dot Cellular Automata," International Journal of Ethics in Engineering and Management Education, vol. 1, no. 4, 2014, pp. 326-328.

[15] Y. Xia, and K. Qiu, "Design and application of universal logic gate based on quantum-dot cellular automata" In Communication Technology, ICCT. 11th IEEE International Conference on, 2008, pp. 335-338.

[16] S. Perri, P. Corsonello, and G. Cocorullo, "Design of efficient binary comparators in quantum-dot cellular automata," IEEE Transactions on Nanotechnology, vol. 13, no. 2, 2014, pp. 192-202.

[17] M. Gladshtein, "Quantum-dot cellular automata serial decimal adder," IEEE Transactions on Nanotechnology, vol. 10, no. 6, 2011. pp. 1377-1382.

[18] S. Saravanan, I. Vennila, and S. Mohanram, "Design and Implementation of an Efficient Reversible Comparator Using TR Gate," Circuits and Systems, vol. 7, no. 9, 2016, pp. 2578.

[19] M. J. Sharifi, and D. Bahrepour, "Introducing a technology index concept and optimum performance design procedure for single-electron-device based circuits," Microelectronics Journal, vol. 42, no. 7, 2011, pp. 942-949.

[20] M. Chabi, A. Roohi, R. F. DeMara, S. Angizi, K. Navi, and H. Khademolhosseini, "Cost-efficient QCA reversible combinational circuits based on a new reversible gate," In Computer Architecture and Digital Systems (CADS), 18th CSI International Symposium on, 2015, pp. 1-6.

[21] Oklobdzija, V. G. (Ed.). (2001). The computer engineering handbook. CRC press.

Davoud Bahrepour was born in Mashhad. He received the M.S. and Ph.D. degree in Computer Architecture from Islamic Azad University, Science and Research Branch, Tehran, Iran in 2007 and 2012 respectively. He is an Assistant Professor in Department of Computer Engineering, Islamic Azad University, Mashhad Branch. His current research interests are Computer Architecture, Nano circuit design and cloud computing. His email address is: bahrepour@mshdiau.ac.ir.

Negin Maroufi received the B.Sc. degree in computer engineering – hardware in Ferdowsi University of Mashhad, in 2013. She received the M.Sc. degree in software engineering from Khorasan Razavi, Neyshabur, Science and Research branch, Islamic Azad University, Neyshabur, Iran, in 2017. She designed a database for a research organization named Neyshabur Longitudinal Study of Ageing (Nelsa) and she is currently working as a database administrator, there. Her research interests are electronics, Nano-sized and QCA circuits, also data analysis and data mining. Her email address is maroofi.negin@mail.um.ac.ir

Long-Term Spectral Pseudo-Entropy (LTSPE): A New Robust Feature for Speech Activity Detection

Mohammad Rasoul Kahrizi *
Department of Computer Engineering and Information Technology, Razi University, Kermanshah, Iran
mr.kahrizi@chmail.ir
Seyed Jahanshah Kabudian
Department of Computer Engineering and Information Technology, Razi University, Kermanshah, Iran
kabudian@razi.ac.ir

Received: 20/Dec/2018 Revised: 19/Apr/2019 Accepted: 29/May/2019

Abstract

Speech detection systems are known as a type of audio classifier systems which are used to recognize, detect, or mark parts of an audio signal including human speech. Applications of these types of systems include speech enhancement, noise cancellation, identification, reducing the size of audio signals in communication and storage, and many other applications. Here, a novel robust feature named Long-Term Spectral Pseudo-Entropy (LTSPE) is proposed to detect speech and its purpose is to improve performance in combination with other features, increase accuracy and to have acceptable performance. To this end, the proposed method is compared to other new and well-known methods of this context in two different conditions, with uses a speech enhancement algorithm to improve the quality of audio signals and without using speech enhancement. In this research, the MUSAN dataset is used, which includes a large number of audio signals in the form of music, speech, and noise. Also, various known methods of machine learning are used. As well as criteria for measuring accuracy and error in this paper are the criteria for F-Score and Equal-Error Rate (EER), respectively. Experimental results on MUSAN dataset show that if the proposed feature LTSPE is combined with other features, the performance of the detector is improved. Moreover, the proposed feature has higher accuracy and lower error compared to similar ones.

Keywords: Audio Signal Processing; Speech Processing; Speech Activity Detection (SAD); Speech Recognition; Voice Activity Detection (VAD); Robust Feature; LTSPE.

1. Introduction

One of the most critical issues in audio signal processing is processing audio signals in which there is a combination of human speech with other sounds like various types of noises, animals' sound and various sounds of different environments. For example, audio signals recorded from speeches, radio, TV, and satellite or different conversations can be mentioned. These audio signals include various types of speech sound signals.

In some applications like the file size reduction, quality enhancement [2-4], compression, bandwidth usage optimization, detection & identification [5-11], and other applications [12-16], it is needed to detect human speech or remove silence and environmental noises from human speech. Speech detection systems are known as a type of audio signal classifier systems which are used to separate, detect, or mark parts of an audio signal which includes human speech.

2. Literature Review

In this section, some of the methods and features for speech detection are mentioned, which are well-known and applicable in speech processing context, and are used here to compare their performances with the proposed method.

One of the most popular and oldest features is Mel-Frequency Cepstral Coefficients (MFCC) [17].

Long-Term Signal Variability (LTSV) and LTSVG (LTSV Gammatone) features [18] are other features which are used for comparison. The LTSV for each frame of audio signals is equal to the entropy variance of each of the frequency bins in that frame. Also, the initial idea of LTSPE feature is inspired by this algorithm.

Another method is Multi-Band Long-Term Signal Variability (MBLTSV) [19], which is a type of LTSV in which frequency scale is warped [20]. The spectrum is divided into predetermined parts which are known as bands, and LTSV is applied to each band. This process improves MBLTSV significantly. Another feature which is used in this research is Long-Term Spectrum Divergence (LTSD) [21].

Other new features and methods that have been employed in audio signal processing context are also used. One of these methods is the method has been proposed by Sadjadi [22] which includes four features called Harmonicity (a.k.a. harmonics-to-noise ratio), Clarity, linear prediction (LP) error and harmonic product spectrum (HPS). And the other is a method that has been proposed by Drugman [23] which includes three features called Cepstral Peak Prominence (CPP), Summation of the Residual Harmonics (SRH) and SRH*. Readers are encouraged to refer to the references for more details.

* Corresponding Author
ORCiD: 0000-0001-6954-1452
The source code of the LTSPE has been registered as ”Long-Term Spectral Pseudo-Entropy (LTSPE) Feature” and DOI ”10.21227/H2G05K” on IEEE Dataport [1].

3. Long-Term Spectral Pseudo-Entropy (LTSPE)

The purpose here is to introduce a new feature called long-term spectral pseudo-entropy to recognize and detect speech. Due to long-term characteristics and noise resistance, this feature can be recognized as a robust feature. LTSPE method is inspired by LTSV [18]. The main difference of the proposed method with LTSV is that the proposed method calculates entropy along the frequency axis, but in LTSV, entropy is calculated along the time axis. Other differences between the proposed method and the similar methods are the differences in the initialization of parameters and measurement intervals. Also, the reasons for naming the proposed method to pseudo-entropy are the same differences with the usual entropy method. In principle, the LTSPE calculates a modified long-term entropy from the frequency spectrum of audio signals.

After dividing the audio signal into frames with a predetermined size, Fourier transform of each frame is calculated, and the power spectrum of each frame is obtained, then these spectra are normalized. This process can be seen in Equations 1 to 4.

$$X(k,n)=\sum_{l=0}^{N_w-1} w(l)\, x(l+(n-1).N_{sh})\, e^{-j\frac{2\pi kl}{N_w}} \tag{1}$$

Where *w(l)* is window function and *X(k, n)* is a short time Fourier transform (STFT) for the n^{th} frame and k^{th} frequency bin. Moreover, N_w is the number of samples per window, and N_{sh} is the number of samples that are considered for frameshift.

$$S_x(k,n)=\left|X(k,n)\right|^2 \tag{2}$$

Where S_x shows the power spectrum.

$$S_M(f,n)=\frac{1}{M+1}\sum_{k=f-\frac{M}{2}}^{f+\frac{M}{2}} S_x(k,n) \tag{3}$$

$$S_R(f,n)=\frac{S_M(f,n)}{\sum_{k=f-\frac{R}{2}}^{f+\frac{R}{2}} S_M(k,n)} \tag{4}$$

Where S_M is the smoothed power spectrum, and S_R is normalized smoothed power spectrum. *k* shows frequency bin and *n* is the frame index. *M* and *R* are even and positive numbers which specify smoothing and normalization intervals for the power spectrum of each frame.

After all these steps, it is time to calculate entropy for each frame. Eq. 5 shows how entropy is calculated for each frame.

$$\xi(n)=-\sum_{k=1}^{K}\left(S_R(k,n)\log S_R(k,n)\right) \tag{5}$$

Finally, the variance of obtained entropies in a predetermined interval is calculated as in Eq. 6.

$$LTSPE(n)=\frac{1}{R_2+1}\sum_{m=n-R_2/2}^{n+R_2/2}\left(\xi(m)-\overline{\xi(n)}\right)^2 \tag{6}$$

Where R_2 specifies the variance interval, also mean entropy $\overline{\xi(n)}$ is calculated as in Eq. 7.

$$\overline{\xi(n)}=\frac{1}{R_2+1}\sum_{m=n-R_2/2}^{n+R_2/2}\xi(m) \tag{7}$$

As already mentioned, the main idea of LTSPE is inspired by the LTSV method. To better compare the LTSV with LTSPE, the calculation method of the LTSV is shown in equations 8 & 9.

$$LTSV_x(m)=\frac{1}{K}\sum_{k=1}^{K}\left(\xi_k^x(m)-\frac{1}{K}\sum_{k=1}^{K}\xi_k^x(m)\right)^2 \tag{8}$$

$$\xi_k^x(m)=-\sum_{n=m-R+1}^{m}\left(\frac{\left|X(n,\omega_k)\right|^2}{\sum_{l=m-R+1}^{m}\left|X(l,\omega_k)\right|^2} *\log\left(\frac{\left|X(n,\omega_k)\right|^2}{\sum_{l=m-R+1}^{m}\left|X(l,\omega_k)\right|^2}\right)\right) \tag{9}$$

For further details of these equations, and more familiarity with LTSV, refer to [18,19].

4. Evaluation and Results

In order to evaluate performance and to obtain higher accuracy and less error in the speech detection process, the proposed method is compared with some of the robust and well-known features in speech detection context which were introduced in Section 2.

MUSAN corpus [24] is used for evaluation. K-Nearest Neighbors (KNN) and Gaussian Mixture Models (GMM) are used for classification.

In GMM method, 16 components are used to model each of the classes, and the number of repetitions of the EM algorithm is set to 100, and also the variance floor is set to 0.01.

Mahalanobis distance is used in the KNN method, and also one neighbor (k = 1) is considered for the number of neighbors in this classifier. The reason for this choice is the higher accuracy in the experimental evaluation results. Moreover, in the reference [25], it has been stated that in the case where the number of training data goes to infinity, it is guaranteed in the classification with 1-NN (k = 1) that the probability of classification error is less than twice of the probability of Bayes error. And to the other word, the probability of classification error is less than twice the probability of optimal error.

Also, to improve quality and decrease the noise of audio signals, Optimally Modified Log-Spectral Amplitude (OMLSA) speech enhancement algorithm [26] is used to improve the quality of speech signals. Of course, in principle, OM-LSA is a speech estimator for non-stationary noise environments that it is employed in this paper.

For experiments, speech signals are selected from MUSAN corpus. The selected speech signals are used in experiments in two conditions: with OMLSA speech enhancement and without speech enhancement. Also, to train non-speech (silence & noise) class, which is here the second class in classification algorithms, all noise and non-speech signals of the MUSAN corpus are used. Different feature extraction algorithms are employed. The obtained features are used to train speech and non-speech classes by GMM & KNN. Frame size and frameshift are 25 & 12.5 ms, respectively. Parameters *M*, *R*, and R_2 are set to 24, 16, and 60. 13-dimensional Mel-Frequency Cepstral Coefficients are extracted from audio signals (c_0-c_{12} are extracted).

Finally, after creating the training models which are obtained by using KNN and GMM, the error of each method and classification accuracy are evaluated using EER (Equal Error Rate) and F-Score criteria in a 10-fold mode. Equations 10 and 11 show how these criteria are calculated. It is important to note that the values of all parameters and intervals are determined experimentally and in the optimum mode. Also, all the results shown in the tables and figures are obtained by the authors with implementing the methods in the MATLAB application.

$$EER = 1 - \frac{|True\ Accepted\ Frames|}{|Total\ Frames|} \tag{10}$$

Where the *EER* value is between 0 and 1. when the value of *EER* is zero, the best performance and the highest accuracy are achieved, of course, it should be noted, when the false acceptance rate (FAR) is as high as possible equal to the false rejection rate (FRR), this equation shows the EER.

$$F-Score \; = \; \frac{2 * Precision * Recall}{Precision + Recall} \tag{11}$$

Where *F-Score* has a value between 0 and 1, and it is a kind of harmonic average of *Precision* and *Recall* criteria, and when the value of this criterion is 1, best performance and the highest accuracy are earned. Moreover, according to [27], the calculation method of precision and recall is shown in Fig. 1.

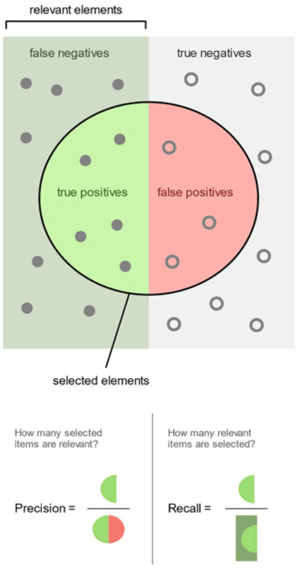


Fig. 1. Method of calculating the precision and recall criteria[1].

Obtained results are presented in Tables 1 and 2 and Figures 2 and 3. As can be seen in results, by considering ERR and F-Score, when the methods in non-combinational mode are used, the best performance is obtained with MBLTSV. However, if other methods and features are used in combinational mode, better options would be obtained like combining LTSPE and MFCC or fusion LTSPE and MBLSTV which give better results compared to when the methods are used non-combinational.

It can be understood from results, that enhancing quality and reducing the noise of speech signals (speech enhancement) before speech detection, has opposite effect unexpectedly and decreases the accuracy of the speech detection system and increases classification error. This result must be accurate and acceptable because some details of the audio signals in the speech enhancement process are eliminated, which reduce the accuracy of the speech activity detection. Also, it should be considered that in this research, the speech enhancement process is entirely separate from speech activity detection.

As it turns out from figures and tables, all in all, the LTSPE has the best results in comparison with similar features such as LTSV, LTSD, and LTSVG. Of course, it should be noted that the similarity of these features is comparable in that the output of these features is only one number per frame.

Table 1. Comparing methods without applying OMLSA (%)

	KNN				GMM			
Features	EER	F-Score			EER	F-Score		
		Overall	Speech	Non-speech		Overall	Speech	Non-speech
Drugman [23]	15.91	63.45	35.97	90.92	21.53	61.90	36.78	87.02
LTSD [21]	15.94	63.49	36.10	90.89	23.55	66.57	48.40	84.74
LTSV [18]	17.79	59.62	29.42	89.82	16.44	66.47	42.53	90.41
LTSVG	15.83	64.09	37.25	90.94	12.56	73.74	54.77	92.71
Proposed LTSPE	14.95	66.04	40.63	91.45	17.16	72.56	55.76	89.36
Sadjadi [22]	18.06	59.32	28.98	89.66	32.72	57.41	36.92	77.91
MBLTSV [19]	3.86	90.89	83.99	97.80	11.72	70.37	47.33	93.40
MBLTSV+Proposed LTSPE	**2.40**	**94.50**	**90.36**	**98.63**	13.29	76.66	61.34	91.97
MFCC [17]	5.41	86.99	77.05	96.93	8.65	83.39	71.90	94.88
MFCC+Proposed LTSPE	3.90	90.78	83.77	97.79	**7.29**	**86.04**	**76.40**	**95.69**
MFCC+LTSV	5.06	87.90	78.67	97.13	9.33	82.72	71.00	94.44

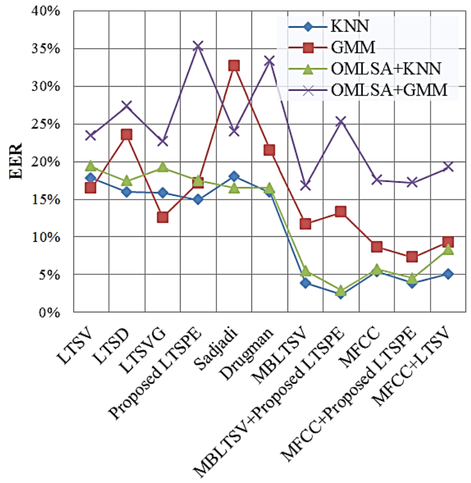


Fig. 2. Comparing classification error in terms of EER in different conditions using different methods (%)

[1] Available: https://en.wikipedia.org/w/index.php?title=Precision_and_recall&oldid=896900450. Accessed: May. 25, 2019.

Table 2. Comparing methods by applying OMLSA (%)

		KNN				GMM		
Features	EER	F-Score			EER	F-Score		
		Overall	Speech	Non-speech		Overall	Speech	Non-speech
Drugman [23]	16.51	62.27	33.97	90.57	33.38	55.43	33.04	77.81
LTSD [21]	17.43	60.56	31.09	90.02	27.32	63.03	44.16	81.91
LTSV [18]	19.40	55.82	22.74	88.91	23.50	58.99	32.19	85.79
LTSVG	19.27	56.43	23.89	88.97	22.68	63.30	40.61	85.99
Proposed LTSPE	17.48	60.48	30.97	89.99	35.36	65.87	56.36	75.37
Sadjadi [22]	16.48	62.52	34.47	90.58	24.03	64.10	43.44	84.76
MBLTSV [19]	5.51	87.11	77.35	96.86	**16.79**	63.38	36.44	90.32
MBLTSV+Proposed LTSPE	**2.91**	**93.42**	**88.49**	**98.34**	25.32	63.74	43.50	83.97
MFCC [17]	5.71	87.26	77.79	96.72	17.53	81.86	69.78	93.93
MFCC+Proposed LTSPE	4.51	89.87	82.19	97.55	17.20	**82.76**	**71.39**	**94.13**
MFCC+LTSV	8.32	80.50	65.39	95.61	19.28	79.78	66.89	92.68

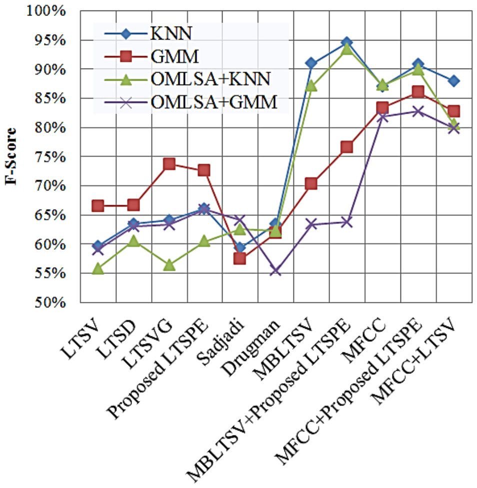


Fig. 3. Comparing classification accuracy in terms of F-Score using different methods (%)

5. Conclusion

The goal of this paper was to introduce a new feature for speech activity detection with the name, long-term spectral pseudo entropy (LTSPE). The LTSPE has been inspired by the LTSV method, and by making some changes to it has been proposed. One of the strengths of the proposed method is that, when combined with other methods, it improves the performance of those methods, even the methods with the highest accuracy, like MBLTSV and MFCC in evaluations. It should be mentioned that comparison in evaluations has been performed only in terms of accuracy and if the size of output data, processing time and calculations of the methods were considered for comparison, results would be different. For example, only four methods have been used in this research where their outputs, feature matrixes, have one column for each audio signal, i.e., four of the mentioned methods give a scalar for each input frame of audio signals, which indicates the smaller size of output data, fewer calculations and faster processing time. These four methods consisted of LTSV, LTSVG, LTSD, and LTSPE, that among them, and on average in all investigated moods, the proposed method LTSPE has had higher accuracy.

References

[1] M. R. Kahrizi, "Long-Term Spectral Pseudo-Entropy (LTSPE) Feature," IEEE Dataport, 2017. [Online]. Available: http://dx.doi.org/10.21227/H2G05K. Accessed: May. 27, 2019.

[2] W. Wang, H. Liu, J. Yang, G. Cao, and C. Hua, "Speech enhancement based on noise classification and deep neural network," Modern Physics Letters B, p. 1950188, 2019.

[3] H. Wang, Z. Ye, and J. Chen, "A Front-End Speech Enhancement System for Robust Automotive Speech Recognition," in 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2018, pp. 1-5: IEEE.

[4] K. Dinesh, R. Prakash, and M. P. Madhan, "Real-time Multi-Source Speech Enhancement for Voice Personal Assistant by using Linear Array Microphone based on Spatial Signal Processing," in 2019 International Conference on Communication and Signal Processing (ICCSP), 2019, pp. 0965-0967: IEEE.

[5] I. Ariav, D. Dov, and I. Cohen, "A deep architecture for audio-visual voice activity detection in the presence of transients," Signal Processing, vol. 142, pp. 69-74, 2018.

[6] M. Pal, D. Paul, and G. Saha, "Synthetic speech detection using fundamental frequency variation and spectral features language," Computer Speech, vol. 48, pp. 31-50, 2018.

[7] B. Mouaz, B. H. Abderrahim, and E. Abdelmajid, "Speech Recognition of Moroccan Dialect Using Hidden Markov Models," Procedia Computer Science, vol. 151, pp. 985-991, 2019.

[8] H. Chen, "Speaker Identification: Time-Frequency Analysis With Deep Learning," ETD Collection for Tennessee State University, 2018.

[9] P. Vecchiotti, G. Pepe, E. Principi, and S. Squartini, "Detection of activity and position of speakers by using Deep Neural Networks and Acoustic Data Augmentation," Expert Systems with Applications, 2019.

[10] F. Tao, "Advances in Audiovisual Speech Processing for Robust Voice Activity Detection and Automatic Speech Recognition," 2018.

[11] A. Ivry, B. Berdugo, and I. Cohen, "Voice Activity Detection for Transient Noisy Environment Based on Diffusion Nets," IEEE Journal of Selected Topics in Signal Processing, 2019.

[12] V. Andrei, H. Cucu, and C. Burileanu, "Overlapped speech detection and competing speaker counting-humans vs. deep learning," IEEE Journal of Selected Topics in Signal Processing, 2019.

[13] C. Vikram, N. Adiga, and S. M. Prasanna, "Detection of Nasalized Voiced Stops in Cleft Palate Speech Using Epoch-Synchronous Features," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019.

[14] N. Mansour, M. Marschall, T. May, A. Westermann, and T. Dau, "A method for conversational signal-to-noise ratio estimation in real-world sound scenarios," The Journal of

the Acoustical Society of America, vol. 145, no. 3, pp. 1873-1873, 2019.

[15] M. Pandharipande, R. Chakraborty, A. Panda, and S. K. Kopparapu, "Robust Front-End Processing For Emotion Recognition In Noisy Speech," in 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2018, pp. 324-328: IEEE.

[16] Y. Malviya, S. Kaul, and K. Goyal, "Music Speech Discrimination," CS229 Final Project, 2016.

[17] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, 1980.

[18] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 3, pp. 600-613, 2011.

[19] A. Tsiartas et al., "Multi-band long-term signal variability features for robust voice activity detection," in Interspeech, 2013, pp. 718-722.

[20] A. Makur and S. K. Mitra, "Warped discrete-Fourier transform: Theory and applications," IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, vol. 48, no. 9, pp. 1086-1093, 2001.

[21] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," Speech Communication, vol. 42, no. 3-4, pp. 271-287, 2004.

[22] S. O. Sadjadi and J. H. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," IEEE Signal Processing Letters, vol. 20, no. 3, pp. 197-200, 2013.

[23] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice activity detection: Merging source and filter-based information," IEEE Signal Processing Letters, vol. 23, no. 2, pp. 252-256, 2016.

[24] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," CoRR, vol. abs/1510.08484, 2015.

[25] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.

[26] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," Signal Processing, vol. 81, no. 11, pp. 2403-2418, 2001.

[27] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

Mohammad Rasoul Kahrizi received the M.Sc. degree in software engineering from the Department of Computer Engineering and Information Technology, Faculty of Engineering, Razi University, Kermanshah, Iran, in 2017. His research interests are in programming, optimization algorithms, data science, signal processing, and machine learning.

Seyed Jahanshah Kabudian is currently an assistant professor at the Department of Computer Engineering and Information Technology in Razi University, Kermanshah, Iran. He received the Ph.D. degree in artificial intelligence from the Amir Kabir University of Technology, Tehran, Iran, in 2010. His research interests lie in signal processing, speech processing, natural language processing, pattern recognition, and machine learning.

DBCACF: A Multidimensional Method for Tourist Recommendation Based on Users’ Demographic, Context and Feedback

Maral Kolahkaj
Department of Computer Engineering, Karaj Branch, Islamic Azad University, Karaj, Iran
maral.kolahkaj@kiau.ac.ir
Ali Harounabadi *
Department of Computer Engineering, Central Tehran branch, Islamic Azad University, Tehran, Iran
a.harounabadi@iauctb.ac.ir
Alireza Nikravanshalmani
Department of Computer Engineering, Karaj Branch, Islamic Azad University, Karaj, Iran
nikravan@kiau.ac.ir
Rahim Chinipardaz
Shahid Chamran University of Ahvaz, Ahvaz, Iran
chinipardaz_r@scu.ac.ir

Received: 28/Jan/2019 Revised: 20/Apr/2019 Accepted: 06/May/2019

Abstract

By the advent of applications in the web 2.0 such as social networks which allow the users to share media, numerous opportunities have been provided for the tourists to recognize and visit attractive and unfamiliar Areas-of-Interest (AOIs); however, finding the appropriate areas based on user’s preferences is very difficult due to several issues such as huge amount of tourist areas, the limitation of the visiting time, etc. In addition, the available methods have yet failed to provide accurate tourist’s recommendations based on geo-tagged media because of several problems such as considering two users with different habits as the same, and ignoring user’s information. Therefore, in this paper, a method called “Demographic-Based Context-Aware Collaborative Filtering” (DBCACF) is proposed to investigate the mentioned problems. DBCACF considers personal and side information in combination with the users' feedbacks to overcome the limitations of collaborative filtering methods in dealing with multi-dimensional data. In addition, a new asymmetric similarity measure is proposed in order to overcome the limitations of symmetric similarity methods. The experimental results on Flickr dataset indicated that the use of personal and side information and the addition of proposed asymmetric scheme to the similarity measure could significantly improve the obtained results compared to other methods which used only user-item ratings and symmetric measures. In particular, our method based on the Cosine similarity measurement has provided a better performance (0.34 for Precision and 0.38 for F-score) as compared to our method based on the Pearson similarity measure over data sparsity and cold-start problems.

Keywords: Decision Support Systems; Data Mining; Context-aware Recommendation; Geo-tagged Photo; Asymmetric Similarity.

1. Introduction

By the advent of the web and its relevant applications, new opportunities have been provided for tourists that they can use travel books, personal blogs, and online services such as travel guides, maps and the like [1]; however, the tourists are dealing with problems to find out relevant information about their request by the exponential data growth in the web called "information overload" [2]. Thus, the users have to choose their favorite items from billions of the objects on the web. Obviously, the evaluation of all of these items is impossible by a user [3]. Therefore, the Recommender System (RS) as a type of Decision Support Systems (DSSs) is designed to filter the information overload and present the relevant results. The RS aims to predict and to recommend the ratings and items which the users are interested in visiting [4-7].

Today, in tourism industry, the users prefer a system which is able to recommend the tourism areas based on their own interests. Several RSs have become increasingly popular to present the tourism services such as Tripadvisor, travelblog, and so on. Although the RSs help the users to deal with the information overload, most existing RSs apply the Collaborative Filtering (CF) and Content-Based (CB) methods to compute the preference between a user and an unvisited item [8]. In location RSs the user-location matrix is highly sparse with numerous missing entries, because users have only visited a very small proportion of attractive areas. These methods are suffering from several problems such as low quality recommendations, low accuracy recommendations and unreliability problems [9,10] due to using only two dimensional data (user-item) [11], considering two users with different habits as the same, data sparsity, cold start condition, lack of the benefit data which are utilized in

* Corresponding Author

recommendation process, etc. [12]. The cold-start problem occurs due to an initial lack of ratings for new users who have not rated any item or new items which have not been rated by any user; therefore, it becomes impossible to make reliable recommendations. On the other hand, sparsity problem occurs when the number of users who have rated items is too small compared to the number of items, hence the recommender system cannot generate any recommendations if there is no overlap in ratings with the target user.

Social RSs are a novel generation of recommenders which utilizes the social media and the interaction data among the users to present the recommendations [13]. The location-based social networks (LBSNs) with the possibility of sharing geo-tagged photos are the examples of networks that the tourists can use to share their travel experiences by uploading photos, providing ratings and comments [10,14,15]. By considering the social data on the social networks, new challenges are also emerged, such as what data is useful for the recommendation process and how to handle this data in order to provide the relevant and accurate recommendations [12,16]. In LBSNs, locations are encoded with latitude and longitude, which distinguishes locations form other items, such as books, music and movies in conventional recommender systems [8,10]. Often, the previous methods which were based on analyzing travel reports and geo-tagged photos, mined popularity areas and do not consider the users' appropriate attributes and contexts on the social network [17]. Therefore, the aim of this study is to propose Demographic- Based Context- Aware Collaborative Filtering (DBCACF) method by using a weighted hybrid method. Based on DBCACF, three RSs are combined to design a tourist recommender system including Collaborative Filtering RS, Demographic RS and Context-Aware RS. The main benefit of combination of these RSs is to take advantage of each particular RS while overcoming limitations of individual RSs. For example, user's demographic and context information is used to overcome data sparsity and cold start problems. Also, to overcome the limitations of methods that used two dimensional data (user-item) and to obtain more accurate and relevant recommendations, this study proposes a method to provide suggestion to the users by handling multi-dimensional data. Furthermore, an asymmetric similarity measure is proposed to imply difference between two different users. Though the previous studies have used various techniques in their recommender systems, the novelty of this work is in integrating Collaborative Filtering, Context-Aware and Demographic as a hybrid recommender system. The main contributions of this paper are as follows:

- Utilizing the explicit users' demographic information in tourism RS based on geo-tagged photos.
- Combining demographic information with the contextual information in tourism RS based on geo-tagged photos to address the cold-start problem and to alleviate sparsity problem.
- Presenting a novel asymmetric similarity measure which determines the appropriate user's neighbors and accordingly recommends relevant and accurate suggestions to the user.
- Conducting extensive experiments to evaluate the performance of DBCACF on Flickr data set.

The organization of this paper is as follows: Section 2 explains the related works in the field of RSs and tourism recommendation. Section 3 presents the proposed method. Section 4 discusses the implementation and evaluations of the proposed method. Finally, the conclusion and future works are considered in Section 5.

2. Related Work

In social media websites which are able to share photos and videos, the tourists participate in sharing the geo-tagged photos [15]. These social media play an increasingly important role as information sources for travelers [18]. On the other hand, the users are interested in searching for the attraction places [19]. In this case, the time-based and the location-based data from social media can be used to provide the assorted recommendations [3,20]. A large body of research was conducted to present different RS methods [4,9,21-30]. The important related methods to this study that can be used in the tourism based on the geo-tagged photos are discussed as follows.

2.1 Symmetric Similarity Methods

CF methods are the dominant methods in RS algorithms such as user-based and item-based approaches [31]. These methods improve the recommendation process by considering the users' similar interests and reduce the overspecialization problem [32]. Despite the success of the CF methods, the performance of these methods is strongly influenced by data sparsity and cold start problems due to the numerous items on the web [25]. A large number of studies have presented several similarity measures to obtain the nearest neighbors for the user [33-41]. In memory-based collaborative filtering, the traditional similarity functions can be used, such as Cosine [42], the Euclidean distance, the mean square difference [43], symmetric Kullback-Leibler divergence [15], Pearson correlation [44,45], to calculate the similarity measure between two tourists. These functions calculate the similarity measure as symmetrical term and thus the users have the same effect on each other for receiving recommendations. In the methods which only considered the ratings of common items, the ratings of uncommon items are ignored in calculating the symmetric similarity between users. As mentioned, these methods are unreliable in cold start and data sparsity problems, because of a low number of common items. Therefore, these problems can reduce the quality and accuracy of the recommendations.

2.2 Asymmetric Similarity Methods

The similarity between two users may be asymmetric and the impact of the first user to the second user is different and vice versa in the real world. In order to achieve a more accurate similarity in CF methods, asymmetric weights are assigned for traditional similarities. For example, Pirasteh et al. presented a weighting scheme in which symmetric similarity became asymmetric similarity among the users by considering the fraction of common items in target user items [43,46,47]. In addition, in another study, an asymmetric similarity which used all of the users' ratings was provided [48]. The similarity measure utilized the local similarity obtained from the Pearson correlation between two users' ratings and global similarity extracted from the Bhattacharyya similarity between each pair of the items. Finally, Jaccard similarity between the two users was combined with the global and local similarities in order to provide more significance to the common items. Although these methods have a higher performance in comparison with previous subsection methods, these are still not efficient in dealing with cold start and data sparsity problems due to only use of the user-item ratings.

2.3 Contexts-Aware Methods

The Context-Aware RS considers a diversity of contextual information on the recommendation process such as time, location or social data [49]. These side information can be incorporated in the recommendation process by three approach: pre-filtering, post-filtering, and contextual modeling [50]. Several methods were recommended based on the contextual information such as time and weather [44,51]. In these methods, the pre-filtering approach was used to utilize the contextual information. Despite of the simplicity of the pre-filtering and post-filtering, these approaches eliminate much appropriate data before the recommendation process. Therefore, in the proposed method a contextual modeling approach is utilized in order to incorporate the contextual information in the recommendation process.

2.4 Hybrid Methods

As mentioned, each of the previous recommendation methods utilizes the certain information and they have several advantages and disadvantages to provide the suggestions. Therefore, two or more RS methods can be combined to use multiple resources and utilize different RSs' advantages [24]. For example, CJacMD [52] is regarded as a combination of cosine, Jaccard and mean measure divergence. In CJacMD approach, the users' ratings and the rating habits of individuals were considered to express preferences.

Table 1 displays a comparative review of advantages and disadvantages of the related works.

Table 1. Recommendation methods

Ref.	Evaluation Metric	Data-set	Advantages	Disadvantages
[15]	Precision, Mean Average Precision@50, Blended Ratio	Flickr	Considering time constraint, personalization based on user history	Ignoring context, Applying symmetric measure, Low accuracy, Explicit questions from users for time and cost
[44]	Precision, Mean Average Precision @50, Blended Ratio	Flickr	Considering context	Pre-filtering context, Symmetric similarity, Low accuracy, Ignoring time constraint
[53]	Precision	Flickr	Considering context	Low accuracy, Only evaluate precision, Pre-filtering context, Ignoring time constraint
[51]	Precision, Mean Average Precision @50, Edit Distance, n Discounted Cumulative Gain@5,@10	Flickr	Considering context	No personalization, Low accuracy, Pre-filtering context, Using probability in recommendations Ignoring time constraint
[43]	Root Mean Squared Error	MovieLens DOUBAN	Improving the traditional similarity by weighted schemes	High error value, Ignoring context, Only evaluate RMSE, Unreliable in term of Cold-Start
[47]	Root Mean Squared Error, Mean Absolute Error	MovieLens NetFlix	Improving the traditional similarity by weighted schemes, Calculating the similarity for two users with no common items	High error value, Only evaluate Errors, Ignoring context
[48]	Root Mean Squared Error, Mean Absolute Error, Precision, Recall, F1	MovieLens NetFlix Yahoo music	Calculating the similarity for two users with no common items	Having high complexity in term of time, Ignoring context
[52]	Root Mean Squared Error, Mean Absolute Error	MovieLens 100K MovieLens 1M Each movie	Considering users ratings habit	Only evaluate Errors, Is Not stable in results, Ignoring context
[45]	Root Mean Squared Error, Coverage	Movie, Food	Weighting context	Having high complexity in term of time

The aim of this study is to combine three RS methods for utilizing their advantages, and therefore, obtaining more relevant and accurate recommendations, including: a) collaborative filtering method for utilizing the interests of similar users and reducing the overspecialization problem, b) demographic based method for overcoming the cold start problem, and c) context aware method for dealing with data sparsity and cold start problems. The time and location as contexts information, and age and gender as demographic information are used to manage the users' travel and to save the time and cost for the users. To utilize a demographic RS, data from user profiles is used, and for using a context-aware system, the data from the photos shared by users is used. This information is implicitly extracted and the recommendations are also recommended to the users without their explicit request. Another advantage of the proposed method is to integrate the context and demographic information into the recommendation process in comparison with other travel recommendation methods.

3. Proposed Method

The present study seeks to propose a new personalized tourist RS called "DBCACF" based on geo-tagged photos and suggest the attractive areas for the users without any explicit requests. Therefore, the collection of community contributed geo-tagged photos, the user demographic attributes, contextual and collective information are utilized to suggest the personalized area recommendation. In demographic-based method, no ratings are necessary to create the user profile in spite of collaborative and content-based methods, and thus it is more efficient for dealing with cold-start problems. In addition, DBCACF overcomes the limitations of conventional collaborative filtering methods. As mentioned, the recognition of the user's neighbors who have similar interests with the target user is known as an important principle in the performance of collaborative filtering. Thus, a new similarity measure is proposed to overcome other similarity measure's problems such as data sparsity, relying on low amount of co-rated items and ratings, depending on the users' explicit ratings, and calculating the similarity as symmetric measure. Fig. 1 illustrates the structure of the proposed method.

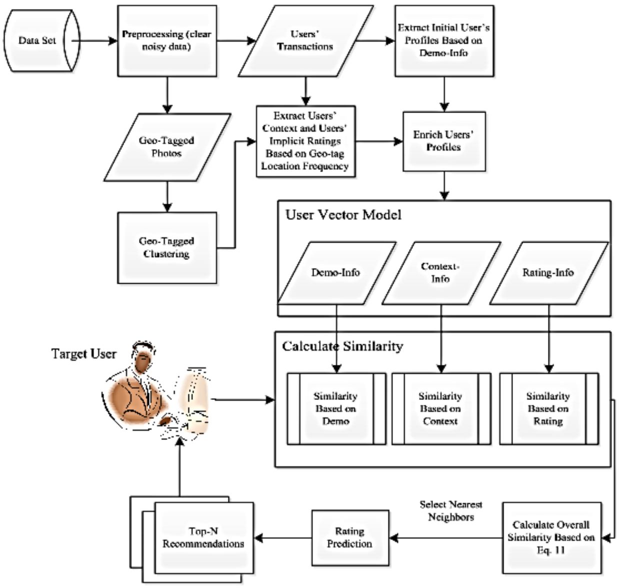

Data Set
Preprocessing (clear noisy data)
Users' Transactions
Extract Initial User's Profiles Based on Demo-Info
Geo-Tagged Photos
Extract Users' Context and Users' Implicit Ratings Based on Geo-tag Location Frequency
Enrich Users' Profiles
Geo-Tagged Clustering
User Vector Model
Demo-Info
Context-Info
Rating-Info
Target User
Calculate Similarity
Similarity Based on Demo
Similarity Based on Context
Similarity Based on Rating
Select Nearest Neighbors
Top-N Recommendations
Rating Prediction
Calculate Overall Similarity Based on Eq. 11

Fig. 1. Structure of proposed method

In short, after pre-processing phase, the users are identified by their ID, and their age and gender are extracted. Then, using geographic information, the photos are clustered to identify the areas of tourism. Using the geo-tagged areas, the user's interests are extracted. These phases are offline phase in the proposed method. In the online phase, the similarity between the new user and other users is calculated and the recommendations are provided to the target user. These steps are described below.

3.1 Profiling and Modeling User's Behavior

After eliminating noisy data, each user is identified based on his/her ID. The information like username, age, and gender, is used as a user vector in order to demonstrate the initial profile of each user.

Next, the geographical data were derived from geo-tagged photos' annotations and each photo is displayed as a photo vector like (photo ID, latitude, and longitude). The estimation of the geo-location of photos is regarded as a challenging task [54,55]. In this paper, the photos were clustered according to their geographical locations and using a density based clustering algorithm called DBSCAN [56], in order to find areas of interest where a photo is taken. DBSCAN clustering requires the least knowledge to determine the input parameters and can discover clusters with each arbitrary shape. In addition, this algorithm can effectively filter noisy data in large datasets.

After obtaining the cluster of each photo, the areas and their visiting time are determined based on the user's historical visiting. In fact, photos with their contextual information such as time and location trace the users who have taken the photos and record the users' temporal-spatial movement. Therefore, in order to take advantage of the user's visiting time and location information, the vector of each user is enriched by (username, age, gender, AOI cluster, and visiting time). The date taken photos are used to determine the time factor and the season visited by tourist is regarded as time context. Next, the user's implicit rating, r, (in this study, user's preference) for each AOI was calculated based on the frequency of visiting the AOI by the user, due to the users' disinclination to provide an explicit rating, $r_{user,AOI} = freq_{user}(AOI)$.

Next, the profile of each user is revised and a hybrid vector model is extracted based on the user's preferences and profile. The final vector model for each user is represented as (username, age, gender, user's preference for each AOI, AOI cluster, visited season). AOI cluster and season represent the contexts information.

3.2 Calculating the Similarity

Social networks allow the users to share their favorite objects. Thus, a considerable number of users are participated for producing the data. Therefore, finding the similar users among all of the users in a social network is one of the most important steps in the proposed method. To this aim, the similarity between each user with other users was calculated by adopting a new similarity measure. The proposed measure combines the demographic and context information with the rating data in the process of calculating the similarity measure. The similarity based on demographic information is to deal with the cold start condition when no preference information is available for a given user. The similarity based on context information is also used to deal with the low number of co-rated or common items in data sparsity problem. Further, the similarity based on rating data is to utilize the gain of the collective wisdom of people. Accordingly, this hybrid similarity dominates the limitations of other similarities. Eq. (1) illustrates the initial proposed similarity measure between two users, *u* and *v*, $(Sim(u,v))$.

$$Sim(u,v) = \beta(Sim_{Demo}(u,v)) + (1-\beta)\big(Sim_{CACF}(u,v)\big) \quad (1)$$

where, Sim_{Demo} represents the similarity base on demographic features between the target user (u) and other users (v), Sim_{CACF} indicates the similarity based on context and collaborative features between u and v, and β is regarded as an adjustment factor for combining two measures. In fact, the Jelinek-Mercer smoothing is used for a linear combination of two similarity measures. The smoothing was implemented to adjust two statistics and avoid the possibility of Zero.

The number of the rated items (in this study, visited AOI) by target user u, I_u, is represented as Eq. (2) in order to overcome the symmetric similarity measure problem and prevent the sensitivity to a low number of co-rated items when a user has rated less items while another user has more rated items, and transform the symmetric similarity to asymmetric similarity.

$$Sim(u,v) = \frac{1}{\gamma|I_u|}[\beta(Sim_{Demo}(u,v)) + (1-\beta)(Sim_{CACF}(u,v))] \quad (2)$$

where γ represents an adjustment factor in different conditions (see Eq. (11)).

The single attribute approach is used to obtain the similarity based on demographic information. In other words, the difference of each user's attribute is first obtained and accordingly the overall similarity of demographic feature is calculated by utilizing the weighted average of differences. The Sim_{Demo} in Eq. (2) is calculated as Eq. (3).

$$Sim_{Demo}(u,v) = \frac{1}{1+\frac{1}{N}\sum_{i=1}^{N} d_i(\alpha_u,\alpha_v)^2}. \quad (3)$$

where *N* represents the number of demographic feature, α_u, α_v indicates the set of the user's demographic features and $d_i(\alpha_u,\alpha_v)$ displays the difference of user u and v on the attribute *i*. The difference is obtained as Eq. (4).

$$d_i(\alpha_u,\alpha_v) = \begin{cases} 1 - OM(\alpha_{iu},\alpha_{iv}) & if \quad \alpha_i = \ nominal\ value \\ \frac{|\alpha_{iu}-\alpha_{iv}|}{max(\alpha)_i - min(\alpha)_i} & if \quad \alpha_i = \ Interval\ value \end{cases}. \quad (4)$$

where OM represents the overlap measure when the attribute type is nominal. The normalized Manhattan distance [57] is used when the attribute type is interval. $max(\alpha)_i and\, min(\alpha)_i$ indicate the maximum and minimum values for the attribute *i*, respectively.

In order to obtain the context-aware collaborative filtering, the contextual modeling approach is utilized as Eq. (5) instead of pre-filtering or post-filtering approaches.

$$Sim_{CACF} = Sim_{CA} \times Sim_{CF}. \quad (5)$$

Based on this equation, the similarity obtained from the user community, Sim_{CF}, has combined with the similarity of contexts, Sim_{CA}. In other words, the context-based similarity is multiplied with Sim_{CF} as a weight. The similarity based on contexts between two users' contexts (C_u, C_v) is calculated based on Jaccard measure (Eq. (6)).

$$Sim_{CA}(C_u, C_v) = \frac{|C_u \cap C_v|}{|C_u \cup C_v|}. \quad (6)$$

According to Eq. (6) and its combination with the similarity of CF, the similar context is not completely ignored and the users can be considered as similar by less weight in the future predictions if the conditions cannot be considered by the two users as similar based on whole matching the contexts.

The similarity-based collaborative filtering is accessed by Eq. (7) as cosine similarity and Eq. (8) as Pearson correlation similarity in order to calculate the similarity between two users' preference vectors (V_{r_u}, V_{r_v}).

$$Sim_{CF}^{COS}(u,v) = \frac{V_{r_u}.V_{r_v}}{\|V_{r_u}\|.\|V_{r_v}\|} = \frac{\sum_{i\in I}(r_{u,i}).(r_{v,i})}{\sqrt{\sum_{i\in I}(r_{u,i})^2.\sum_{i\in I}(r_{v,i})^2}}. \quad (7)$$

$$Sim_{CF}^{PCC}(u,v) = \frac{\sum_{i\in I}(r_{u,i}-\bar{r}_u).(r_{v,i}-\bar{r}_v)}{\sqrt{\sum_{i\in I}(r_{u,i}-\bar{r}_u)^2}.\sqrt{\sum_{i\in I}(r_{v,i}-\bar{r}_v)^2}}. \quad (8)$$

where, I is a set of AOIs where users, u and v, have visited.

If two users share more common contexts, their similarity collaborative weight will be more by multiplying the similarity based on collaborative and context together. In addition, the Sorensen index (SRS in Eq. (9)) is added to Eq. (2) in order to provide more significance for the users with more common items and demographic information (Eq. (10)).

$$SRS = \frac{2|I_u \cap I_v|}{|I_u|+|I_v|}. \quad (9)$$

where $\mathrm{I_u}$, $\mathrm{I_v}$ are the number of visited AOI by target user and other users respectively.

$$Sim_{total}(u,v) = SRS + [\beta(Sim_{Demo}(u,v)) + (1-\beta)(Sim_{CACF}(u,v))]. \quad (10)$$

According to different terms and different similarities between two users, the total DBCACF similarity is defined as Eq. (11).

In this equation, the total similarity between the users is zero when they have no similarity in any respects. The overall similarity between two users is calculated by demographic information or CACF if the similarity based on CACF becomes zero, but the similarity based on demographic is nonzero (no common item or cold-start conditions), or if the similarity based on CACF becomes nonzero, but the similarity based on demographic is zero. In addition, the similarity needs to be an asymmetric measure by $\frac{1}{|I_u|}$ as an asymmetric weight. Finally, the total similarity between two users is calculated by both CACF and demographic information, if both similarities based on CACF and demographic become nonzero. Next, the existing similarity needs to be an asymmetric measure by the weight $\frac{1}{2|I_u|}$, which is added to the Sorensen index.

The obtained similarity measure can calculate the similarity in term of cold-start condition when a user has no co-rated items and suggest the recommendations to such users without relying on the co-rated items. On the other hand, the effect of similarity based on demographic is not completely delimited by increasing the co-rated and common items.

$$Sim_{DBCACF}(u,v) = \begin{cases} 0 & if \quad (Sim_{Demo}(u,v)) = 0, (Sim_{CACF}(u,v)) = 0 \\ \frac{1}{|I_u|}\left[\beta(Sim_{Demo}(u,v)) + (1-\beta)(Sim_{CACF}(u,v))\right] & if \ or \begin{matrix} (Sim_{Demo}(u,v)) = 0, (Sim_{CACF}(u,v)) \neq 0 \\ (Sim_{Demo}(u,v)) \neq 0, (Sim_{CACF}(u,v)) = 0 \end{matrix} \\ SRS + \frac{1}{2|I_u|}\left[\beta(Sim_{Demo}(u,v)) + (1-\beta)(Sim_{CACF}(u,v))\right] & if \quad (Sim_{Demo}(u,v)) \neq 0, (Sim_{CACF}(u,v)) \neq 0 \end{cases} \qquad (11)$$

Next, by considering the obtained similarities, Top-*K* users with the most similarity measure with the target user as nearest neighbors are selected.

3.3 Prediction and Recommendation

In this step, the interesting areas are predicted for the current user based on his preferences, context, and personal features by considering the neighbors' interests. That is, after obtaining the nearest neighbors, the neighbor's items are extracted as candidate items. Next, the score of each candidate item is predicted for the target user by Eq. (12). Finally, the candidate items with the highest importance are suggested to the target user as a recommendation list.

$$pred(u,i) = \frac{\sum_{n \in neighbors(u)} Sim(u,n) \cdot (r_{n,i})}{\sum_{n \in neighbors(u)} Sim(u,n)} \qquad (12)$$

4. Experiments and Evaluations

In this section, our proposed method is evaluated based on the Flickr dataset. Accordingly, the comparison results of the proposed method with other recommendation methods are presented.

4.1 Dataset Description

To demonstrate the effectiveness of our proposed method for recommendations, we conducted several experiments using Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M). The YFCC100M is the largest public multimedia collection ever released, with a total of 100 million media objects, of which approximately 99.2 million are photos and 0.8 million are videos, all uploaded to Flickr [58,59]. First, the metadata related to 16,533 geo-tagged photos were extracted. This data was captured in different cities of Iran between August 01, 1961 and March 28, 2014, and include several information such as the user ID, photo ID, geo-position, date taken and annotation tags. Then, the personal users' information such as, age and gender, was collected Crowley from Flickr social network. Table 2 represents the sample records which are used in this study. Photo ID and User ID were used anonymous for the purpose of privacy.

Table 2. Sample records of tourists and photos

Photo ID	User ID	Gender	Age	Date taken	Longitude	Latitude
1	1	Male	42	9/14/04 1:23 AM	51.460111	35.820523
2	2	-	-	10/14/04 11:38 PM	-	-
3	3	Male	40	5/1/04 4:07 PM	51.332931	35.729513
4	3	Male	40	5/2/04 4:13 PM	51.332931	35.729513

Next, the data with the following conditions was eliminated:

Those having no geographic information such as latitude and longitude,

Those without any user's profile data,

Those collected based on search result with the name of Iran in its metadata such as place, tags, description and title but its geographic information such as latitude and longitude failed to match the geographical context of the Iranian cities (Iran is located between 25-40 degrees of north latitude and between 44-64 degrees of east longitude).

Table 3 displays a descriptive statistic about the related dataset.

Table 3. Statistics on the dataset

Photos		Users	Locations
Raw	Filtered		
16533	10030	19	3873

4.2 Detecting Tourist Areas

In order to find the areas of interest, geographic information from photos was used based on the DBSCAN clustering method. For this study, the parameters of this algorithm are set as Min-Pts= 6 and ε=0.4. To this end, 58 area clusters were obtained and the vector of each photo was enriched as *(photo ID, latitude, longitude, cluster number*).

4.3 Profiling Users

The user vector is shown as (*UID*, *Age*, *Gender*, P_1, P_2, …, P_n, C_1, C_2, …, C_m), where each $Pn \in$ Preferences represents the user preference about each AOI (implicit rating), and each $C_i \in C$ indicates the context which has been visited by a tourist, in other word if a user visits a context, the corresponding value for C_i will be 1, otherwise will be 0.

4.4 Prediction and Recommendation

The recommendation problem is defined as predicting ratings for the interesting objects that have not been seen by target user. In order to predict the tourist's AOIs, 75 % of the data was randomly selected for training and rest of the data was held for testing. The training set was utilized to predict the preferences to the target user by using each method. Accordingly, the data from the test set was used to evaluate the recommendation quality and the accuracy of prediction. Finally, a list of Top-*N* recommendation was matched with the actual list of visited areas by the target user, and the results of different methods were compared for final evaluation.

4.5 Experimental Evaluation

The performance metrics, compared methods, results and discussions to assess the proposed method are as follows.

4.5.1 Evaluation Metrics

Two statistical metrics were utilized to evaluate the prediction accuracy of the proposed DBCACFs and other methods. First, the Mean Absolute Error (MAE) is defined as Eq. (13).

$$\mathrm{MAE} = \frac{1}{|\mathrm{N}|}\sum_{\mathrm{u,i}}\left|\mathrm{r_{u,i}} - \mathrm{Pred_{u,i}}\right|. \quad (13)$$

Second, the Root Mean Squared Error (RMSE) is defined as Eq. (14).

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\mathrm{N}|}\sum_{\mathrm{u,i}}(\mathrm{r_{u,i}} - \mathrm{Pred_{u,i}})^2} \quad (14)$$

where $r_{u,i}$ represents the actual implicit rating of the user *u* provided to item *i*, $Pred_{u,i}$ indicates the predicted rating of the user *u* imputed to item *i* by different methods, and $|N|$ displays the number of the test ratings. A smaller value of MAE or RMSE indicates the better prediction accuracy.

Further, three decision-support metrics including precision, recall and F-score were used to evaluate the prediction quality of the proposed methods. A larger value of precision, recall and F-score represents the better quality of prediction. Precision is regarded as the ability of RS for recommending the relevant suggestions and is interpreted as the fraction of the correct predictions in total number of predictions. Recall is considered as the ability of RS for recommending all of the suggestions which are visited by the target user. In fact, recall is interpreted as the fraction of correct predictions in total number of relevant items; however, F1-Measure can be interpreted as a weighted average of the precision and recall, where an F- Measure reaches the best and worst value at 1 and 0, respectively.

4.5.2 Compared Methods

CFcos is a Collaborative Filtering method (user-based kNN [61]). Cosine similarity is implemented to obtain similar users with target user in term of their ratings [34].

CFpcc is another Collaborative Filtering method (user-based kNN [61]) in which Pearson similarity is used to obtain similar users with target user in term of their ratings [44].

CACFpcc is regarded as Context-Aware Collaborative Filtering method. Pearson similarity is used to obtain the similar users' ratings with target user's ratings. While Jaccard similarity is utilized to access the similar contexts of two users [45].

CACFcos is described as another Context-Aware Collaborative Filtering method. Cosine similarity is used to obtain similar users' ratings with target user's ratings in this method. Like CACFpcc Jaccard similarity is implemented to access the similar contexts of two users.

ACOS is an asymmetric Collaborative Filtering method which used the Cosine, the asymmetric Jaccard and the Sorensen Index in order to obtain similar users' ratings with target user's ratings [47].

APCC is regarded as another asymmetric Collaborative Filtering method which utilized the Pearson, the asymmetric Jaccard and the Sorensen Index to obtain similar users' ratings with target user's ratings [46].

4.5.3 Results and Discussions

It is a nontrivial problem to choose best values of parameters in Eq. (11) that will produce meaningful and insightful results. Therefore, after thorough experimentation of different values for parameter β (from 0 to 1 by step 0.1), it was concluded that this parameter produced interesting and informative results for $\beta = 0.3$.

Figs. 2-3 and Table 4 illustrate the performance of the proposed DBCACF and other methods in terms of Precision, Recall, F-score, MAE and RMSE.

As illustrated in Fig. 2, the increasing the number of neighbors results in increasing the precision values for all methods.

It can be seen that Collaborative Filtering methods have the lowest precision values in compared with other methods, and provide almost equal results for all neighborhood size. The main reason for these results is that CF methods cannot deal well with the sparse data.

However, by using an asymmetric similarity scheme, the recommendation results for ACOS and APCC methods improve significantly for neighborhood size from 60 to later. It should be noted that, two users with different rating vectors have a different impact on each other and this is exactly observable in the results. It is worth noting, due to the sparsity problem, these methods need more number of neighbors to provide better recommendation results.

As another observation, the CACF methods also yield better results as compared to Collaborative Filtering methods. One reason for this improvement is that, these methods incorporate additional information into the recommendation process other than the user-item data and this information can improve the recommendation results in the sparse space. However, these methods use symmetric similarity scheme to select nearest neighbors for a target user and therefore they have still a lower precision values as compared with ACOS and APCC methods.

On another hand, the precision value for the proposed method is higher than 0.34 which considerably increased in comparison with other methods. When we deal with the cold start or data sparsity problems, the users have a few ratings in their visiting history. In these cases, using users' demographic information can provide similar neighbors in term of age and gender and it is possible that they have similar interests with the target user. Based on these data, when the first user has lower ratings and the second one has more ratings, the similarity measure is different by considering our asymmetric similarity measure. That is, the second user is a valuable neighbor for the first one. Therefore, the recommendation has a higher quality in the proposed method; while the first user resembles less and is a valueless neighbor for the second one. Therefore, in the proposed method, the user is less influenced for providing recommendations based on the experimental results. In summary, the main reasons for the improvement in the proposed methods are that, incorporating additional information such as contexts and demographics in addition to user-item data into the recommendation process by using a new hybrid method and providing a new asymmetric similarity scheme to select the nearest neighbors for the target user over data sparsity and cold-start conditions.

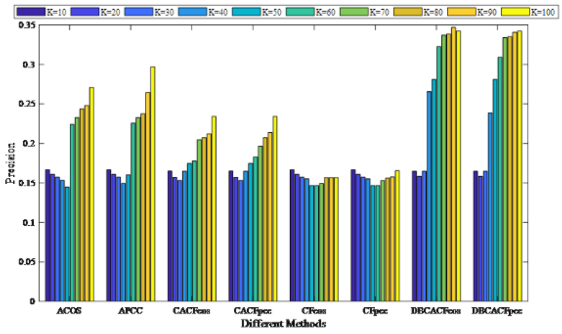


Fig. 2. Precision Measure of different methods

To further evaluate the quality of the proposed method as the recall and precision are regarded as the important metrics in this study, the effect of both metrics was evaluated in the proposed method by using the F1-measure (Fig. 3). Based on the results presented in Fig. 3 for the F-score measure, it can be concluded that the quality presented in the precision results is reliable (F-score results are almost similar to the precision values).

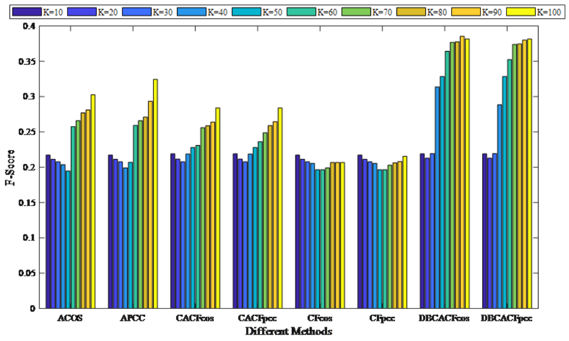


Fig. 3. F1 Measure of different methods

To better illustrate the advantages and characteristics of the proposed method, the average quality and accuracy criteria are demonstrated in Table 4.

Based on the results reported in Table 4, the proposed methods can provide better estimation of scores, and their prediction error is less than other methods. It can be observed that the proposed methods can recognize the similar neighbors for all neighborhoods, and provide the appropriate prediction error; while other methods have provided a higher error rate in prediction results. On the other hand, by considering the sparsity of this dataset, the proposed method can perform better than others on different numbers of neighbors.

Table 4. Average quality and accuracy metrics

Methods	Quality			Accuracy	
	Precision	Recall	F-Score	MAE	RMSE
CFcos	0.1556	0.4223	0.2057	0.6096	0.6966
CFpcc	0.1562	0.4235	0.2064	0.6171	0.7061
ACOS	0.1958	0.4295	0.2382	0.5586	0.6647
APCC	0.2011	0.4307	0.2424	0.5572	0.6636
CACFcos	0.1839	0.3998	0.2368	0.5403	0.6625
CACFpcc	0.1823	0.3997	0.2353	0.5381	0.6612
DBCACFcos	0.2692	0.4271	0.3152	0.5294	0.6577
DBCACFpcc	0.2644	0.4254	0.3108	0.5310	0.6586

We argue that, based on the experimental results, DBCACFcos and DBCACFpcc could yield better results as compared to ACOS, APCC, CACFs and CFs methods in term of precision, F-score, MAE and RMSE metrics; while, the recall measure for the APCC was a little higher than other methods and the ACOS was at the next rank. After that the DBCACFcos was placed at the later one. In general, better results were obtained when the contexts were considered in the recommendation process, compared to collaborative filtering methods. In fact, the results demonstrated the contexts of each user are actually regarded as useful information in tourist recommendations. In addition, the proposed methods outperformed all other recommendations methods when contextual and demographic information were utilized to provide the area recommendations. It should be note that, we deal with sparse rating data because a considerable number of users can visit awhile areas and several users are envisaged the cold start problem due to numerous areas of interest, time and cost constraints. In these conditions, utilizing demographics and contexts information integrated with the user feedbacks can provide better quality and accurate results. Furthermore, using the proposed asymmetric similarity measure could find the users which are more similar to the target user when it deals with a few numbers of co-rated and common items.

In general, our method based on the Cosine similarity measurement has provided a better performance in compared with our method based on the Pearson similarity measure over data sparsity and cold-start problems.

5. Conclusions and Future Work

As mentioned, previous methods are suffering from impersonalized recommendations, low quality recommendations, low accuracy recommendations and unreliability problems due to using only user-item data, considering two different users as the same, data sparsity, cold start condition, etc. In this study, the information related to demographics, contexts, and collective wisdom of people were utilized to provide area recommendations. Based on this information a hybrid user profile was created. In addition, a new hybrid similarity measure was proposed based on an asymmetric scheme which was calculated between each pair of users in order to overcome the limitations of other methods to present the user nearest neighbors and recommend the personalized tourist's area.

Based on the experimental results, DBCACFcos and DBCACFpcc could yield better results, compared to ACOS, APCC, CACFs and CFs methods in term of Precision, F-score, MAE and RMSE values over cold-start and data sparsity conditions.

Despite the aforementioned advantages, the proposed method has several weaknesses. First, the sequence of areas was not intended to recommend the suggestions on the purpose of planning travel. Second, no time and budget constraints were considered due to the need of explicit questions from the users for each recommendation and unavailability this information in the used dataset. Finally, time and location contexts received the same values in recommendation process.

Future work can concentrate on the following directions. First, using the sequences between locations visited by the users can be emphasized as an appropriate factor for better recommendations. The sequence of visited locations by employing a sequential pattern mining algorithm can help planning travel. Second, other user's features, such as occupation, companion and the like, can be used to investigate whether these features are valuable for tourist recommendations based on community contributed geo-tagged photo collection. Third, selecting and weighting the most important contexts can be included in the recommendation model among all the available contexts. Finally, extracting hidden factors that affect the recommendation process, such as outlier, is a good topic for future research.

References

[1] D. Godoy and A. Corbellini, "Folksonomy-Based Recommender Systems: A State-of-the-Art Review," Int. J. Intell. Syst., vol. 31, no. 4, pp. 314-346, 2016.

[2] P. Yu, L. Lin, and Y. Yao, "A Novel Framework to Process the Quantity and Quality of User Behavior Data in Recommender Systems," in Proceedings of the 17th International Conference on Web-Age Information Management, WAIM 2016, Nanchang, China, Part I, 2016, pp. 231-243.

[3] Z. K. Zhang, T. Zhou, and Y. C. Zhang, "Tag-Aware Recommender Systems: A State-of-the-Art Survey," Journal of Computer Science and Technology, vol. 26, no. 5, pp. 767-777, 2011.

[4] Z. L. Zhao, C. D. Wang, Y. Y. Wan, and J. H. Lai, "Recommendation in feature space sphere," Electronic Commerce Research and Applications, vol. 26, no. Supplement C, pp. 109-118, 2017.

[5] S. Khusro, Z. Ali, and I. Ullah, "Recommender Systems: Issues, Challenges, and Research Opportunities," in Proceedings of the Information Science and Applications (ICISA), Singapore, 2016, pp. 1179-1189.

[6] M. Kolahkaj, A. Harounabadi, and M. Sadeghzade, "A Recommender System for Web Mining using Neural Network and Fuzzy Algorithm," International Journal of Computer Applications, vol. 78, no. 8, pp. 20-24, 2013.

[7] M. Kolahkaj and M. Khalilian, "A recommender system by using classification based on frequent pattern mining and J48 algorithm," in 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), 2015, pp. 780-786.

[8] R. Safa, S. Mirroshandel, S. Javadi, and M. Azizi, "Venue Recommendation Based on Paper's Title and Co-authors Network," Journal of Information Systems and Telecommunication, vol. 1, no. 6, pp. 209-217, 2017.

[9] X. Ma, H. Lu, Z. Gan, and J. Zeng, "An explicit trust and distrust clustering based collaborative filtering recommendation approach," Electronic Commerce Research and Applications, vol. 25, no. Supplement C, pp. 29-39, 2017.

[10] R. Gao, J. Li, X. Li, C. Song, and Y. Zhou, "A personalized point-of-interest recommendation model via fusion of geo-social information," Neurocomputing, vol. 273, pp. 159-170, 2018.

[11] Y. M. Afify, I. F. Moawad, N. L. Badr, and M. F. Tolba, "A personalized recommender system for SaaS services," Concurrency and Computation: Practice and Experience, vol. 29, no. 4, p. e3877, 2017.

[12] N. M. Villegas, C. Sánchez, J. Díaz-Cely, and G. Tamura, "Characterizing context-aware recommender systems: A systematic literature review," Knowledge-Based Systems, vol. 140, pp. 173-200, 2018.

[13] I. Cenamor, T. de la Rosa, S. Núñez, and D. Borrajo, "Planning for tourism routes using social networks," Expert Systems with Applications, vol. 69, pp. 1-9, 2017.

[14] G. Cai, K. Lee, and I. Lee, "Itinerary recommender system with semantic trajectory pattern mining from geo-tagged photos," Expert Systems with Applications, vol. 94, pp. 32-40, 2018.

[15] I. Memon, L. Chen, A. Majid, M. Lv, I. Hussain, and G. Chen, "Travel Recommendation Using Geo-tagged Photos in Social Media for Tourist," Wirel. Pers. Commun., vol. 80, no. 4, pp. 1347-1362, 2015.

[16] D. Wang, S. Deng, and G. Xu, "Sequence-based context-aware music recommendation," Information Retrieval Journal, vol. 21, no. 2, pp. 230-252, 2018.

[17] D. Bachmann et al., "(CF)2 architecture: contextual collaborative filtering," Information Retrieval Journal, 2018.

[18] Z. Xiang and U. Gretzel, "Role of social media in online travel information search," Tourism Management, vol. 31, no. 2, pp. 179-188, 2010.

[19] V. Subramaniyaswamy, V. Vijayakumar, R. Logesh, and V. Indragandhi, "Intelligent Travel Recommendation System by Mining Attributes from Community Contributed Photos," Procedia Computer Science, vol. 50, pp. 447-455, 2015.

[20] R. S. Aquino, H. A. Schänzel, and K. F. Hyde, "Unearthing the geotourism experience: Geotourist perspectives at Mount Pinatubo, Philippines," Tourist Studies, vol. 18, no. 1, pp. 41-62, 2018.

[21] L. Ravi and S. Vairavasundaram, "A Collaborative Location Based Travel Recommendation System through Enhanced Rating Prediction for the Group of Users," Intell. Neuroscience, vol. 2016, p. 7, 2016.

[22] M. Pazzani and D. Billsus, "Learning and Revising User Profiles: The Identification ofInteresting Web Sites," Mach. Learn., vol. 27, no. 3, pp. 313-331, 1997.

[23] D. H. Park, H. K. Kim, I. Y. Choi, and J. K. Kim, "A literature review and classification of recommender systems research," Expert Systems with Applications, vol. 39, no. 11, pp. 10059-10072, 2012.

[24] M. Nilashi, O. bin Ibrahim, N. Ithnin, and N. H. Sarmin, "A multi-criteria collaborative filtering recommender system for the tourism domain using Expectation Maximization (EM) and PCA–ANFIS," Electronic Commerce Research and Applications, vol. 14, no. 6, pp. 542-562, 2015.

[25] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734-749, 2005.

[26] M. Balabanovi and Y. Shoham, "Fab: content-based, collaborative recommendation," Commun. ACM, vol. 40, no. 3, pp. 66-72, 1997.

[27] W. Wu et al., "Improving performance of tensor-based context-aware recommenders using Bias Tensor Factorization with context feature auto-encoding," Knowledge-Based Systems, vol. 128, pp. 71-77, 2017.

[28] D. H. Lee and P. Brusilovsky, "Improving personalized recommendations using community membership information," Information Processing & Management, vol. 53, no. 5, pp. 1201-1214, 2017.

[29] T. Ha and S. Lee, "Item-network-based collaborative filtering: A personalized recommendation method based on a user's item network," Information Processing & Management, vol. 53, no. 5, pp. 1171-1184, 2017.

[30] A. Sesagiri Raamkumar, S. Foo, and N. Pang, "Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems," Information Processing & Management, vol. 53, no. 3, pp. 577-594, 2017.

[31] K. H. L. Tso-Sutter, L. B. Marinho, and L. Schmidt-Thieme, "Tag-aware recommender systems by fusion of collaborative filtering algorithms," in Proceedings of the ACM symposium on Applied computing, Fortaleza, Ceara, Brazil, 2008, pp. 1995-1999.

[32] A. M. Nagrale and A. P. Pande, "User Preferences-based Recommendation System for Services Using Map Reduce Approach for Big Data Applications," International Journal of Innovations & Advancement in Computer Science, vol. 4, pp. 528-532, 2015.

[33] M. Y. H. Al-Shamri, "Effect of Collaborative Recommender System Parameters," Adv. in Artif. Intell., vol. 2016, pp. 1-10, 2016.

[34] Z. Zhang, X. Zheng, and D. D. Zeng, "A framework for diversifying recommendation lists by user interest expansion," Knowledge-Based Systems, vol. 105, pp. 83-95, 2016.

[35] J. Bobadilla, F. Ortega, A. Hernando, and J. Alcalá, "Improving collaborative filtering recommender system results and performance using genetic algorithms," Knowledge-Based Systems, vol. 24, no. 8, pp. 1310-1316, 2011.

[36] Q. Cheng et al., "The new similarity measure based on user preference models for collaborative filtering," in Proceedings of the IEEE International Conference on Information and Automation, 2015, pp. 577-582.

[37] G. Guo, J. Zhang, and N. Yorke-Smith, "A Novel Evidence-Based Bayesian Similarity Measure for Recommender Systems," ACM Trans. Web, vol. 10, no. 2, pp. 1-30, 2016.

[38] M. Y. H. Al-Shamri, "User profiling approaches for demographic recommender systems," Knowledge-Based Systems, vol. 100, pp. 175-187, 2016.

[39] S. Zammali, K. Arour, and A. Bouzeghoub, "A Context Features Selecting and Weighting Methods for Context-Aware Recommendation," in Proceedings of the IEEE 39th Annual Computer Software and Applications Conference, 2015, vol. 2, pp. 575-584.

[40] H. Liu, Z. Hu, A. Mian, H. Tian, and X. Zhu, "A new user similarity model to improve the accuracy of collaborative filtering," Knowledge-Based Systems, vol. 56, pp. 156-166, 2014.

[41] Z. L. Zhao, C. D. Wang, and J. H. Lai, "AUI&GIV: Recommendation with Asymmetric User Influence and Global Importance Value," PLoS One, vol. 11, no. 2, pp. 1-21, 2016.

[42] S. H. Cha, "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions," International Journal of Mathematical Models and Methods in Applied Sciences, vol. 1, no. 4, pp. 300-307, 2007.

[43] P. Pirasteh, D. Hwang, and J. E. Jung, "Weighted Similarity Schemes for High Scalability in User-Based Collaborative Filtering," Mob. Netw. Appl., vol. 20, no. 4, pp. 497-507, 2015.

[44] A. Majid, L. Chen, G. Chen, H. T. Mirza, I. Hussain, and J. Woodward, "A context-aware personalized travel recommendation system based on geotagged social media data mining," Int. J. Geogr. Inf. Sci., vol. 27, no. 4, pp. 662-684, 2013.

[45] Y. Zheng, R. Burke, and B. Mobasher, "Recommendation with Differential Context Weighting," in Proceedings of the 21th International Conference on User Modeling, Adaptation, and Personalization, UMAP 2013, Rome, Italy, 2013, pp. 152-164.

[46] P. Pirasteh, J. J. Jung, and D. Hwang, "An Asymmetric Weighting Schema for Collaborative Filtering," in In: Proceedings of the New Trends in Computational Collective Intelligence, 2015: Springer, pp. 77-82.

[47] P. Pirasteh, D. Hwang, and J. J. Jung, "Exploiting matrix factorization to asymmetric user similarities in recommendation systems," Knowledge-Based Systems, vol. 83, pp. 51-57, 2015.

[48] B. K. Patra, R. Launonen, V. Ollikainen, and S. Nandi, "A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data," Know.-Based Syst., vol. 82, no. C, pp. 163-177, 2015.

[49] S. Vairavasundaram, V. Varadharajan, I. Vairavasundaram, and L. Ravi, "Data mining-based tag recommendation system: an overview," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 5, no. 3, pp. 87-112, 2015.

[50] G. Adomavicius and A. Tuzhilin, "Context-Aware Recommender Systems," in Recommender Systems Handbook1st ed.: Springer US, 2011, pp. 217-253.

[51] A. Majid, L. Chen, H. T. Mirza, I. Hussain, and G. Chen, "A system for mining interesting tourist locations and travel sequences from public geo-tagged photos," Data & Knowledge Engineering, vol. 95, pp. 66-86, 2015.

[52] Suryakant and T. Mahara, "A New Similarity Measure Based on Mean Measure of Divergence for Collaborative Filtering in Sparse Environment," Procedia Computer Science, vol. 89, pp. 450-456, 2016.

[53] Z. Xu, "Trip similarity computation for context-aware travel recommendation exploiting geotagged photos," in Proceedings of the 30th IEEE International Conference on Data Engineering Workshops, 2014, pp. 330-334.

[54] S. Kisilevich, F. Mansmann, and D. Keim, "P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos," in Proceedings of the 1st ACM International Conference and Exhibition on Computing for Geospatial Research & Application, Washington, D.C., USA, 2010, pp. 1-4.

[55] L. J. Li, R. K. Jha, B. Thomee, D. A. Shamma, L. Cao, and Y. Wang, "Where the Photos Were Taken: Location Prediction by Learning from Flickr Photos," in Proceeding of the Large-Scale Visual Geo-Localization, 2016, pp. 41-58.

[56] M. Ester, H. P. Kriegel, r. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996, pp. 226-231.

[57] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," Third ed. Boston: Morgan Kaufmann, 2012, pp. 39-82.

[58] B. Thomee et al., "YFCC100M: the new data in multimedia research," Commun. ACM, vol. 59, no. 2, pp. 64-73, 2016.

[59] Flickr. (2018, 6 jan. 2017). [Online]. Available: http://www.Flickr.com.

[60] WebscopeYahooLabs. (15 Feb 2017). [Online]. Available: https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67.

[61] A. Bellogín and P. Sánchez, "Collaborative filtering based on subsequence matching: A new approach," Information Sciences, vol. 418, pp. 432-446, 2017.

Maral Kolahkaj received the B.S. degree in Computer Engineering from Azad University, Dezfoul Branch, Iran in 2009, and M.S. degree in Software Systems from Azad University, Science & Research Branch, Iran, in 2013. Currently she is Ph.D. Candidate in Azad University, Karaj Branch, Iran. Her research interests include Information retrieval, Recommendation systems, Human-computer interaction, Context awareness, Web mining and Data mining.

Ali Harounabadi is an assistant professor of Computer Engineering at Azad University, Central Tehran Branch, Iran. His research is focused on Recommender Systems, Web mining and Methodologies in Software Engineering.

Alireza Nikravanshalmani is a full-time Assistant-Professor of Computer Engineering in the Faculty of engineering at Azad University, Karaj Branch, Iran. He received his Ph.D. degree in Computer Science from Kingston University-London, England in 2011. He is conducting research activities in the areas of Image and Data Analysis Software Systems.

Rahim Chinipardaz received the B.S. degree in Statistics and Computer Science from Shahid Chamran University, Ahwaz, Iran in 1984, and M.S. degree in Statistics from Tarbiat Modares University, Tehran, Iran, in 1989. He received his Ph.D. degree from Newcastle University of England in 1996. Now, he works as full-professor in the Faculty of Mathematics, Statistics and Computer Science at Shahid Chamran University, Ahwaz, Iran. His area research interests include Statistics, Probability, Time series analysis and Discriminant Analysis.

A Novel Method for Image Encryption Using Modified Logistic Map

Ardalan Ghasemzadeh*
Computer Engineering & Information Technology Department, Urmia University of Technology, Urmia, Iran
a.ghasemzadeh@uut.ac.ir
Omid R.B. Speily
Computer Engineering & Information Technology Department, Urmia University of Technology, Urmia, Iran
speily@uut.ac.ir

Received: 19/Nov/2018 Revised: 19/Mar/2019 Accepted: 07/May/2019

Abstract

With the development of the internet and social networks, the multimedia data, particularly digital images, has been of increasing interest to scientists. Due to their advantages including high speed, high security, and complexity, chaotic functions have been broadly employed in image encryption. The present paper proposed a modified logistic map function which resulted in higher scattering in the obtained results. Confusion and diffusion functions, as the two main actions in cryptography, are not necessarily performed in order, i.e. each of these two functions can be applied on the image in either order, provided that the sum of total functions does not exceed 10. So, to calculate the sum of functions, confusion has the factor of 1 and diffusion has the factor of 2. To simulate this method, a binary stack was used. Application of binary stack and pseudo-random numbers obtained from the modified chaotic function increased the complexity of the proposed encryption algorithm. The security key length, entropy value, NPCR and UACI values, and correlation coefficient represented in the analytical results revealed the capability and validity of the proposed method. Analyzing the obtained results and comparing the algorithm to other investigated methods clearly verified high efficiency of the proposed method.

Keywords: Encryption; Decryption; Logistic Map; Confusion; Diffusion.

1. Introduction

With widespread use of the Internet, especially after the advent of online communities such as social networks, millions of bytes of information are being transmitted every day [1]. This information is transmitted through text, audio, image and video between different users [2]. Encryption is a traditional technique for the secure transmission of this information. However, the traditional text encryption techniques cannot protect images efficiently due to the big difference between images and texts. Image Encryption is a serious challenge in governmental (military, medical) digital services, multimedia systems, and Internet-based communications. On the other hand, the recent advances in information and communication technology and e-commerce have provided potential markets for distributing digital content such as image over the Internet [3]. A big challenge is how to protect the intellectual property of multimedia content, namely image, in multimedia networks. Accordingly, development of efficient methods for the storage and transfer of digital data has become an attractive topic for researchers. It is highly

Necessary for these data to be transformed to a template preventing the access of invalid users to them.

Therefore, ensuring the security of image messages is a dramatically important topic today [4]. In the last decade, different encryption algorithms have been introduced based on various principles in the literature, such pixel adaptive diffusion [5], fractional wavelet transform [6], image filtering [7], elliptic curve [8], and reversible cellular automata [9]. 2D cellular automatic machines [10, 11], collusion function-based methods [12], domain phase-based algorithms [13], and chaos theory [14-16] are more popular in image encryption. Due to the properties of chaotic systems such as acceptable speed, security, and high complexity, chaos-based encryption algorithms can be useful in many applications.

An ideal encryption method is expected to address some of the fundamental requirements of encryption including chaos, distribution, and randomness. Due to their random behavior and high sensitivity to primary parameters and conditions, chaotic systems provide great potential to resist the attack of invalid users. Natiq et al. proposed a new hyperchaotic map based on Sine map and two-dimensional Henon map. They indicated that their proposed method could encrypt digital images with high complexity performance and low implementation cost [17]. Huang and Ye used an image encryption algorithm based on irregular wave representation [18]. Although the method's NPCR and UACI were higher than 0.9 and 0.33, respectively, it was time-consuming. In [19], a synchronous permutation-diffusion technique was used for image encryption. In that paper, in order to reduce the sending process time, permutation and diffusion steps for any pixel are performed in the same time. An image encryption scheme based on chaotic tent map is proposed by Li et al. [20]. They show that, image encryption systems based on such map show some better performances. A new cryptographic method was proposed based on Henon chaotic map in [21]. Color image encryption using random

* Corresponding Author

transforms, phase retrieval, chaotic maps, and diffusion was presented in [22]. A combination of several encryption tools is used in that paper.

Chaotic functions are generally employed in the encryption of text and image files. Basic logistic map chaos function has some disadvantages such as short alternation period, undesirable uniformity, and low independence from the generated data [23].

In this paper, a modified basic logistic map function is introduced to overcome the disadvantages such as undesirable uniformity and low independence. Then the modified logistic map function is employed in the encryption of images to efficiently apply diffusion and confusion functions on images using a stack structure and favorably performed encryption.

The rest of this paper is organized as follows. Section 2 introduces the theoretical foundations of the proposed algorithm. Section 3 investigates the implementation of the proposed algorithm and its effect on some sample images. The experimental results along with their analysis are presented in section 4. Conclusions are represented in section 5.

2. Theoretical Foundations of the Proposed Algorithm

General procedure of image encryption algorithms using chaotic functions is as follows:

i. First, the input image is transformed into a 2D arrangement of constituent pixel values.

ii. A chaotic map function is chosen to generate pseudo-random numbers. In this step, parameters and primary values of the chaotic function are chosen in order for the signal to have chaotic behaviour in the considered interval.

iii. All parameters and values required for the encryption of the image in the destination are included in the encryption key.

iv. The procedure of confusion is applied. In this step, matrix values of the image are relocated to random positions using random numbers generated by the chaotic function. This relocation takes place according to box, row-column, singular, or other methods. Method selection affects the operational speed of the algorithm.

v. The procedure of diffusion is applied. In this step, every single pixel in the image matrix is changed. In fact, the value of each pixel is added to the random value obtained from chaotic function and other parameters (depending on the used algorithm considering speed and security) were balanced within the valid interval of pixel values.

Depending on the type of the algorithm, each of the two above steps could be repeated until the algorithm reached the required resistance to different attacks. The order of repeat was included in the encryption key. Decryption of the image was done at the destination with the encryption key with the reverse order of the above steps.

2.1 Algorithm Description

In the proposed algorithm, the modified linear logistic chaotic map was used. Logistic map function is defined by Eq. (1) [24]:

$$x_{n+1} = r.x_n(1 - x_n) \quad n = 1, 2, 3, \ldots \tag{1}$$

Fig. (1) shows the behavior of this function over time.

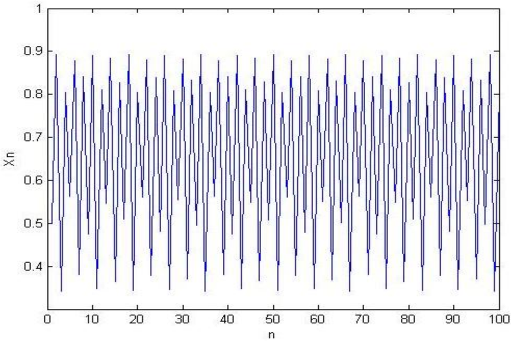


Fig. 1. Behavior of logistic map function

The main problem of logistic map function is that the scattering of numbers generated in the random interval is low, especially in the beginning and end of the interval. By introducing some changes in the basic logistic map function, as shown below, more chaotic and random behavior was observed:

```
Logistic(x0,r)
Begin
   pow←1
   t←(r * x0 * (1 - x0))
   if ((t * 100) Mod 2) = 0 then
  pow ← 2
   p ← t + Math.Pow(-1, pow) * (t /100)
   if (0 < p and p < 1) then
      return p
   else
return Logistic(t, r)
end.
```

Algorithm 1. Modified Logistic Map algorithm

In this algorithm, t stands for the value obtained from basic logistic map function. If t is even it increases by $t/100$, otherwise $t/100$ is subtracted from its value.

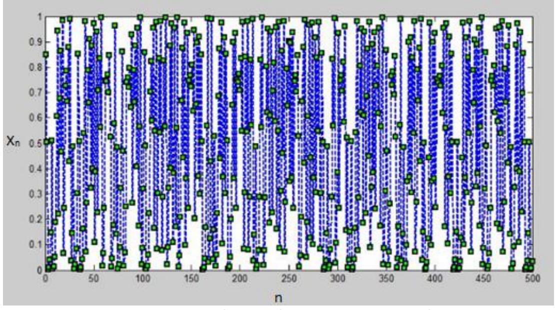


Fig. 2. Behavior of modified logistic map function

Only values between 0 and 1 were used as random values. The behavior of the modified function for the primary value of $x_0 = 0.3$ and parameter of $r = 3.988$ is shown in Fig. 2. The values obtained from the algorithm were arranged in a linear array and sorted in an ascending order. After sorting, the primary array index was used as the values for encryption. The scattering of basic and modified logistic map in the interval of [0,1] for 500 repeats is given in Table 1.

Table 1. Results for logistic map and modified logistic map

Primary values	$x_0 = 0.5 \quad r = 3.988 \quad n = 500$
Main Algorithm	79,59,36,27,35,40,40,27,43,114
Modified Algorithm	86 ,49,34,43,21,52,36,41,53,85
Primary values	$x_0 = 0.8 \quad r = 3.978 \quad n = 500$
Main Algorithm	80,45,33,38,42,36,33,30,56,107
Modified Algorithm	33 ,44,42,40,25,33,39,86,53,105

As it is clear in the results, the scattering of the modified mapping is better for primary similar values.

2.2 Stack Structure

The proposed algorithm is not sensitive to the order of confusion and diffusion applied on the image provided that the sum of the applied procedure does not exceed 10. To calculate the sum of procedures, confusion procedure has the factor of 1 and diffusion procedure has the factor of 2. For the simulation of this method, the binary stack was used so as to make it possible to perform the procedures in the reverse order in the decryption step and obtain the original image. During encryption, the stack status is included in the encryption key and the stack structure is restored at the destination according to the values of encryption key.

2.3 Encryption Key Definition

In the introduced algorithm, key consists of two parts and has a length of 128 bits. The first part keeps parameters and primary values of the chaotic function and the second part is used for mapping the stack status. At the beginning and for each image, the part associated with the parameters and primary values in the encryption key is generated randomly. Furthermoer, according to these 0-1 strings, primary value and parameter of chaotic function is generated.

3. Implementation of the Proposed Algorithm

The proposed algorithm is implemented based on the following steps:

a. First, the values of three elements of color image, namely red, green, and blue, are separated and put into separate matrices, i.e. if the dimensions of the input image are in the form of $L = [h * W]$ then matrices will be in the form of $L_2 = [h * 3W]$.

b. The encryption key is generated as a sequence of random binary numbers, as shown in Fig. 3.

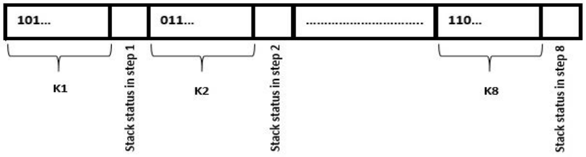


Fig. 3. key structure

According to the generated key, parameter R and primary value X_0 are obtained by the following equations:

$$K = K_1, K_2, \dots, K_8 \quad (2)$$

$$K' = (K_1 \oplus K_2), \dots, (K_7 \oplus K_8) \quad (3)$$

$$K'' = (K'_1 \oplus K'_2), (K'_3 \oplus K'_4) \quad (4)$$

$$Sum = K'_1 + K'_2 + K'_3 + K'_4 \quad (5)$$

$$K''' = K''_1 \oplus K''_2 \quad (6)$$

$$X_0 = [(float)(K''' + Sum)\%256]/256 \quad (7)$$

$$r = 3.97 + [(float)\ Sum\ \%\ 299]/10000 \quad (8)$$

According to Fig. 3, each K_i includes 14 bit-containers of encryption key. Applying these equations, parameters and primary values are obtained randomly within the following intervals:

$$x \in [0, 0.9999] \qquad r \in [3.97, 3.9999]$$

In order to apply confusion, the two following steps are taken:

a. Linear array I_1 with length h is created and filled with decimal values obtained from chaotic function. The sorted values of I_1 are put into array I_2. Then, values of I_2 are searched in I_1 and their index is input into another array with the name of Index, whose length is equal to I_1 and I_2. This way, column indices of the image are randomly placed in the Index array. Then, each column of L_2 matrix is relocated according to the Index array, as well illustrated in Fig. 4.

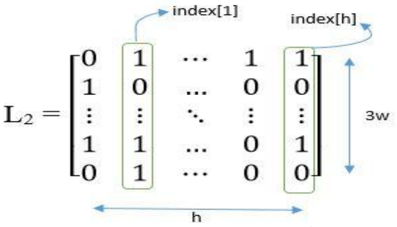


Fig. 4. First step in applying confusion

b. Now, according to Fig. 5, the procedure of step c is applied to rows in order to relocate 3w dimension.

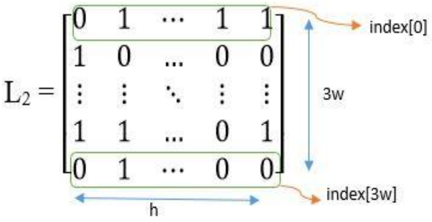


Fig. 5. Second step in applying confusion

Moreover, the following steps are taken for applying diffusion algorithm.

c. First, in linear chaotic matrix with the length of $h * w_2$, random numbers from the interval $[0, (h * w) - 1]$ are located. The parameter X_0 is divided by 3, i.e. $X'_0 = X_0/3$.

d. In this step, the parameter sum is added to each element of a pixel. Here, the sum includes the value of red element associated with the previous pixel and the random number generated for the current pixel. The red color element of the previous pixel is added so that if a small variation takes place in the decryption of the image, the original image cannot be restored even if the correct key is available. Then, the value of each color element of the pixel is added with sum and balanced into the interval [0,255]. This step is applied on every pixel of the image.

e. Finally, all three matrix elements of the image are integrated.

At last, having the parameter and primary value as well as the employed logistic function, random numbers used during encryption can be restored to decrypt the image. That is to say, using the above-mentioned steps in the reverse order, the original image is obtained. In the Fig. 6, the proposed method is shown in the flowchart format.

3.1 Results of Applying Algorithm on Some Sample Images

The proposed algorithm was applied on some standard images including Figs. 7, 8, and 9 as primary, encrypted, and decrypted images, respectively. Moreover, the color histogram of each image is shown within three main color channels. The color histograms clearly shows the performance of the proposed algorithm for Baboon and Lena images.

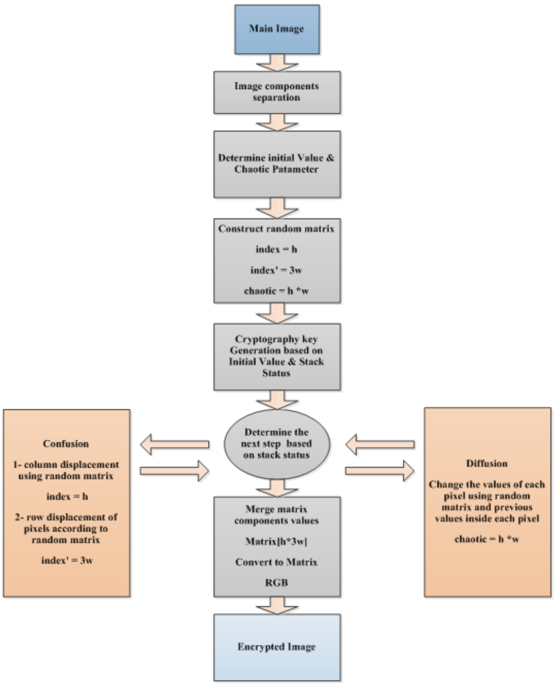

Main Image
Image components separation
Determine initial Value & Chaotic Patameter
Construct random matrix
index = h
index' = 3w
chaotic = h *w
Cryptography key Generation based on Initial Value & Stack Status
Determine the next step based on stack status
Confusion
1- column displacement using random matrix
index = h
2- row displacement of pixels according to random matrix
index' = 3w
Diffusion
Change the values of each pixel using random matrix and previous values inside each pixel
chaotic = h *w
Merge matrix components values
Matrix[h*3w]
Convert to Matrix
RGB
Encrypted Image

Fig. 6. The flowchart of the proposed method

4. Results and Discussions

An optimal encryption algorithm mush have enough security and efficiency against different types of decryption attacks, statistical attacks, comprehensive operation of key space, and brute-force. In what follows, the performance of the proposed algorithm in these tests is described.

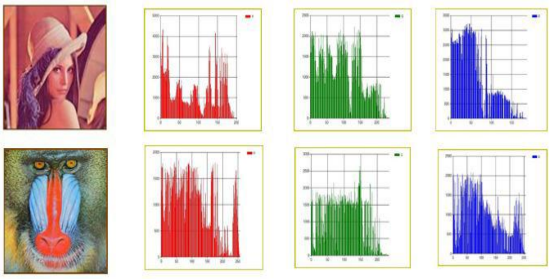


Fig. 7. Primary images and color histograms

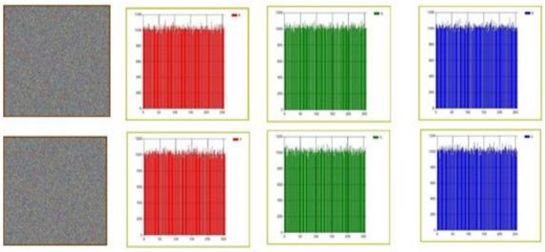


Fig. 8. Ciphered images and color histograms

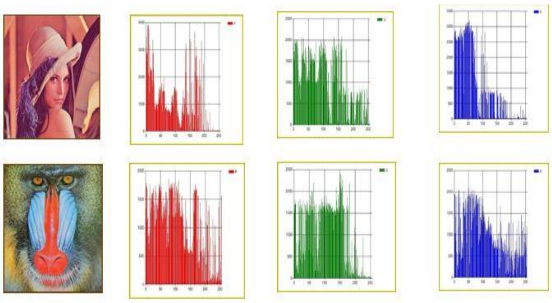


Fig. 9. Decrypted images and color histograms

4.1 Key Space Analysis

In an encryption algorithm, key space has to be big enough to be resistant against brute-force attacks. NIST organization has predicted the minimum required length of encryption key to be 80 bits [25]. The key space of the proposed method is 2^128 which is much bigger than 2^80, therefore it is safe against comprehensive key search attacks.

4.2 Histogram Analysis

Histogram shows the number of pixels at each gray level of an image. If the distribution of gray levels is not uniform in an image, the original image can be restored merely with attack by having the encrypted image and not needing the key. Therefore, a good encryption algorithm has to act in order for the histogram of the encrypted image to have random and uniform appearance and the attackers cannot get any information from this aspect of the image.

In Figs. 7, 8, 9, the histograms of two standard images of Lena and Baboon are shown in three states including original, encrypted, and decrypted, respectively. Histogram of the image in encrypted state was perfectly uniform and different from the histogram of the original image. That is to say that the attackers cannot obtain any information on the original image by analyzing the histograms of encrypted images.

4.3 Correlation Coefficient Analysis

The correlation between two adjacent pixels is known as the correlation coefficient which is one of the most important features in image encryption area [16]. To investigate the correlation between two adjacent pixels in an image, first 4069 couples of adjacent pixels were chosen totally randomly. Later, the correlation coefficient of each couple was calculated using Eq. (9) [26].

$$r_{xy} = \frac{\mathrm{Cov}(x,y)}{\sqrt{D(x)}\,\sqrt{D(y)}} \tag{9}$$

where x and y are the gray levels of two adjacent pixels. Eqs. (10)-(12) define parameters used in Eq. (9).

$$E(x) = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{10}$$

$$\mathrm{Cov}(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - E(x))(y_i - E(y)) \tag{11}$$

$$D(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - E(x))^2 \tag{12}$$

Below, the results of this test for adjacent pixels in diagonal state for two images of Lena and Baboon are represented. Figs. 10 and 11 display the diagonal correlation for the two images.

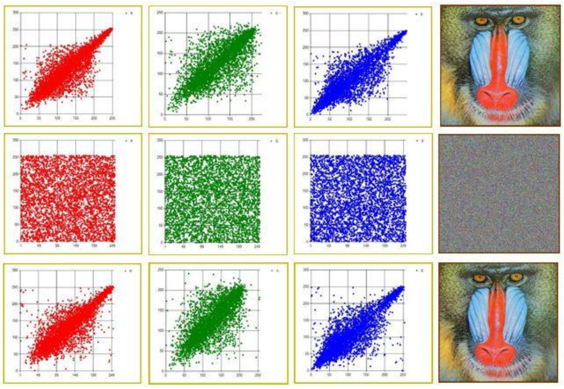


Fig. 10. Diagonal correlation test for original, encrypted, and decrypted images

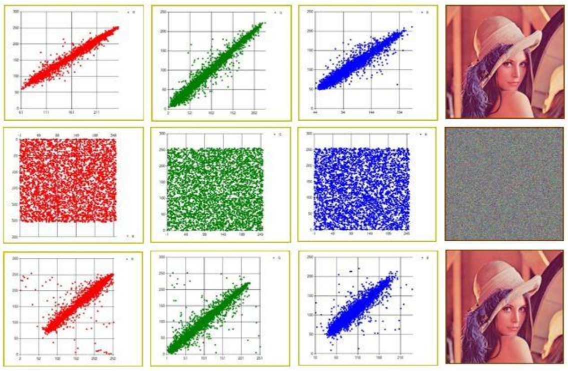


Fig. 11. Diagonal correlation test for original, encrypted, and decrypted images

As shown in Table 2, all three diagonal, horizontal, and vertical correlations are calculated for the image of Lena.

Table 2. Correlation of Lena image

Correlation type	Plain image	Cipher image
Horizontal correlation	0.97028249	0.0079959
Vertical correlation	0.9628058	0.0096502
Diagonal correlation	0.985534	0.0109409

According to the results and figures, it is clear thatthe correlation of pixels is significantly reduced in the encrypted state .

4.4 Analysis of the Sensitivity of the Algorithm

In addition to the key sensitivity, plaintext sensitivity is also an important rule to evaluate the efficiency of a designed image encryption algorithm. In other words, the algorithm should be very sensitive to the plain-image even just for one-bit change [18]. To analyze the sensitivity of the algorithm, the original image was encrypted at first. Then, one pixel of the original image was changed in a completely random way. The obtained image was encrypted one more time and finally the two encrypted images were compared based on the following equations. The effect of changing one pixel in the original image on the encrypted image was investigated with two measurement criteria, namely NPCR and UACI [27].

NPCR is the average number of pixels in the encrypted image varied due to the change of one pixel in the original image. For two encrypted images C1 and C2 whose original images were different only in one pixel, a 2D arrangement of $D(i,j)$ was first calculated according to the following equation:

$$D(i,j) = \begin{cases} 0 & ؛\ C1(i,j) = C2(i,j) \\ 1 & ؛\ C1(i,j) \neq C2(i,j) \end{cases} \tag{13}$$

Where $C1(i,j)$ and $C2(i,j)$ stand for the value of gray level of pixels in encrypted images of $C1$ and $C2$, respectively. Based on [27], NPCR is the absolute number of pixels which changes value in differential attacks, can be evaluated by Equation (14).

$$\mathrm{NPCR} = \frac{\sum_{i,j} D(i,j)}{M \times N} * 100 \tag{14}$$

Where M and N are the dimensions of the original image.

UACI is the average of lightness intensity difference between two images which is calculated as follows [27]:

$$\mathrm{UACI} = \frac{1}{M \times N}\left[\frac{\sum_{i,j}|C1(i,j) - C2(i,j)|}{2^L - 1}\right] * 100 \tag{15}$$

Where L is the number of bits used for displaying the image, which was 8 in the present study.

To design an acceptable encryption method, the NPCR should be greater than 0.99 and the UACI is about 0.33 [28]. Average NPCR and UACI for this algorithm were 0.993454 and 0.334267, respectively.

4.5 Analysis of Entropy

Entropy or irregularity is a criterion for describing the randomness of information source which was first introduced by Shanon in 1949. Moreover, it is calculated as follows and the theoretical value is 8 for a message in gray level [15].

$$\mathrm{H(s)} = -\sum_{\mathrm{i=0}}^{\mathrm{N-1}} \mathrm{P(S_i).\log}\left(\frac{1}{\mathrm{P(S_i)}}\right) \qquad (16)$$

where N is the number of gray levels used in the image which was $2^8 = 256$ here and $\mathrm{P(S_i)}$ shows the probability of occurrence of the i-th gray level in the image. Table 3 shows the results of entropy test for original and encrypted images of Lena.

Table 3. Entropy test results for Lena image

	Red element	Green element	Blue element
Plain image	7.2530845	7.5951533	6.968516
Cipher image	7.999302	7.999353	7.999269

4.6 Performance Analysis

To provide more evaluations, the proposed method was compared with several methods including hyper chaotic map [7], irregular wave representation [15], synchronous permutation-diffusion technique [19], chaotic tent map [20], chaos-based fast image encryption algorithm [28], hybrid genetic algorithm and chaotic function model [29], chaos-based symmetric image encryption using a bit-level permutation [30], DNA sequence operation and hyper-chaotic system [31], quantum logistic map [32], and coupled two-dimensional piecewise chaotic map [33] , Total Chaotic Shuffling Scheme [34].

First, this comparison was made based on the correlation of adjacent pixels, shown in.Table 4.

The results show that the correlation in the proposed method is smaller than that in many other methods.

Table 5 summarizes the sensitivity of the algorithm to the original image and compares it with other algorithms.

Table 4. Correlation comparison between the proposed method and other related methods

	Horizontal	Vertical	Diagonal
Proposed algorithm	0.0009995	0.006650	0.00109
Ref. [18]	0.0172	0.0277	0.0039
Ref. [19]	0.0008	0.0021	0.0005
Ref. [20]	0.0016	0.0025	0.0003
Ref. [28]	0.0009	-0.0022	0.0149
Ref. [29]	-0.0054	0.0093	-0.00009
Ref. [30]	0.002	0.0009	0.0016
Ref. [31]	-0.0002	0.0038	0.0009
Ref. [32]	0.0065	0.0055	0.0082
Ref. [34]	0.00235	0.001235	0.00036

As can be seen in Table 5, after encryption of both images, the values of NPCR and UACI exceed 99% and 33%, respectively, outperforming other comparable methods and proving that the proposed method has the capability to resist differential attacks.

Table 5. The performance of the proposed scheme and other comparable methods based on NPCR and UACI

Algorithm	UACI	NPCR
Proposed algorithm	0.334267	0.993454
Ref. [10]	0.3327	0.9941
Ref. [19]	0.335989	0.996304
Ref. [28]	0.335615	0.996427
Ref. [29]	0.331084	0.971394
Ref. [30]	0.334815	0.996473
Ref. [31]	0.335989	0.996304
Ref. [34]	0.334627	0.996086

In Table 6, the entropies of the original and encrypted images at three states are shown and compared with other algorithms.

The results show that the entropy of the proposed method is very close to the ideal entropy value, higher than that of many other existing algorithms.

Table 6. Entropy Comparison

Algorithm	Entropy
Proposed Algorithm	7.999308
Ref. [10]	7.9965
Ref. [19]	7.9994
Ref.[20]	7.9998
Ref. [28]	7.9994
Ref. [29]	7.9978
Ref.[30]	7.9993
Ref.[31]	7.9975
Ref.[33]	7.9992
Ref.[34]	7.9993

4.7 Resistance to Noise Analysis

The resistance of an encryption system to noise in real-world communication technologies is one of the most important issues. When an image is transferred through a communication channel, it can be exposed to destructive noise. A good encryption algorithm needs to have the potential of preventing severe destruction of the decrypted image on receiver's side when the encrypted images are subjected to noise by being transferred through a communication channel [25,35].

The results obtained by tests on the images show that the proposed algorithm has the required resistance to salt and pepper noise and Gaussian noise. Figs. 12 and 13 present the results obtained by applying salt and pepper noise on sample images. Fig. 12 depicts the results obtained by testing the image after applying 15% salt and pepper noise on the encrypted Baboon image which can be observed after decryption on the right side of the figure.

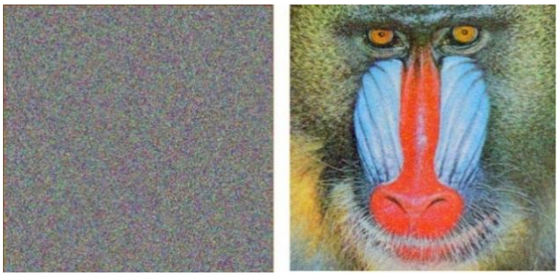


Fig. 12. Applying Salt and pepper noise with 15% density on encrypted image

Fig. 13 shows the application of 40% noise on an image of an airplane.

Noise can have a great destructive effect on the key such that decryption becomes impossible even by small changes in the key. Transferring key through a safe channel, independent from the channel through which the image is transferred, can prevent this problem.

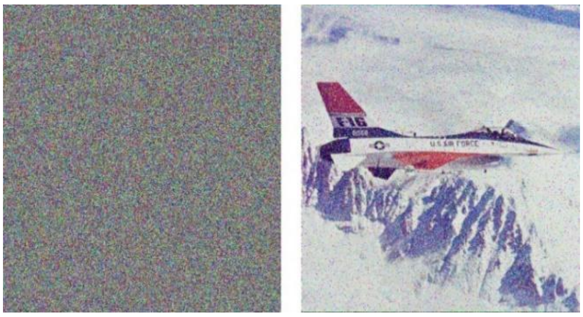


Fig. 13. Applying Salt and pepper noise with 40% density on encrypted image

5. Conclusion

Different patterns have proposed for image encryption among which those based on chaos theory have gained special popularity. The present paper proposed a novel algorithm based on chaos theory and modified logistic map. Pseudo-random numbers obtained from the modified function had higher distribution than those obtained from the basic logistic mapping function.

For encryption, confusion and diffusion were applied on the pixels of the original image in a random order. Furthermore, the random order was managed based on a binary stack and a 128-bit key on both sender and receiver sides. The key consisted of two parts where the first part kept the parameter and primary value of chaotic function and the second part was used for mapping stock states. First, for each image, the part related to the parameter and primary value was randomly generated in the encryption key. Then, based on this 0-1 strings, primary value and the parameter of the chaotic function were created.

The proposed algorithm was tested on sample and standard images. Moreover, the parameters required for the analysis of the proposed algorithm were discussed and compared with other algorithms, verifying the efficiency and security of the proposed lagorithm to different attacks.

References

[1] Kardan, O. R. B. Speily. Increasing Information Reposting Behavior in Online Learning Community. Educational Technology & Society, 21(4), 100–110. 2018.

[2] Speily, O. R. B.. De-lurking in Online Communities Using Repost Behavior Prediction Method. Information Systems & Telecommunication, 192. 2017.

[3] Tarokh, M. J., Arian, H. S., & Speily, O. R. B.. Discovering Influential Users in Social Media to Enhance Effective Advertisement. Advances in Computer Science: An International Journal, 4(5), 23–28. 2015.

[4] H. Natiq,M. Said, & A. Kilicman. "A new hyperchaotic map and its application for image", https://doi.org/10.1140/epjp/i2018-11834-2. 2018.

[5] Z. Hua, S. Yi & Y. Zhou, "Medical image encryption using high-speed scrambling and pixel adaptive diffusion", Signal Processing. https://doi.org/10.1016/j.sigpro.2017.10.004, 2017.

[6] B. Gaurav, Q. M. Jonathan Wu & B. Raman. "Discrete fractional wavelet transforms and its application to multiple encryption." Inf. Sci. 223-297-316, 2013.

[7] Z. Hua & Y. Zhou. "Design of image cipher using block-based scrambling and image filtering". Information Sciences. https://doi.org/10.1016/j.ins.2017.02.036, 2017.

[8] L. D. Singh & K. M. Singh. 'Image Encryption using Elliptic Curve Cryptography", Procedia - Procedia Computer Science, 54, 472–481. https://doi.org/10.1016/j.procs.2015.06.054, 2015.

[9] X. Wang & D. Luan. "A novel image encryption algorithm using chaos and reversible cellular automata". Communications in Nonlinear Science and Numerical Simulation, 18(11), 3075–3085. https://doi.org/10.1016/j.cnsns.2013.04.008, 2013.

[10] O. Lafe, "Data Compression and Encryption using Cellular Automata Transform", Engineering Applications of Articial Intelligence, Vol. 10, No. 6, pp. 581–591, 1998.

[11] R. J. Chen and J. L. Lai, "Image Security System using Recursive Cellular Automata Substitution," Pattern Recognition , Vol. 40, pp. 1621–1631, 2007.

[12] A. Jolfaei and A. Mirghadri, "Survey: Image Encryption Using Salsa20", International Journal of Computer Science Issues, Vol. 7, Issue 5, pp. 213-220, September 2010.

[13] S. E. Borujeni and M. Eshghi, "Chaotic Image Encryption System using Phase-Magnitude Transformation and Pixel Substitution", J. Telecommun. Syst. DOI:10.1007/s11235-011-9458-8. 2011.

[14] C. Zhu, "A Novel Image Encryption Scheme based on Improved Hyperchaotic Sequences," Journal of Optics Communications, Vol. 285, pp 29–37, 2012.

[15] B. Norouzi, S. Mirzakuchaki, S. M. Seyedzadeh, and M. R. Mosavi, "A Simple,Sensitive and Secure Image Encryption Algorithm based on Hyper-Chaotic System with Only One Round Diffusion Process," The Journal of Multimedia Tools and Applications, DOI 10.1007/s11042-012-1292-9, 2012.

[16] X. Wang and L. Teng, "An Image Blocks Encryption Algorithm based on Spatiotemporal Chaos", J Nonlinear Dyn., Vol. 67, pp. 365–371, 2012.

[17] H. Natiq, M. R. M. Said & A. Kilicman. "A new hyperchaotic map and its application for image". https://doi.org/10.1140/epjp/i2018-11834-2, 2018.

[18] X. Huang, G. Ye, "An image encryption algorithm based on irregular wave representation". https://doi.org/10.1007/s11042-017-4455-x. 2017.

[19] R. Enayatifar, A. H. Abdullah, I. F. Isnin, A. Altameem & M. Lee, "Image encryption using a synchronous permutation diffusion technique". Optics and Lasers in Engineering, 90(June 2016), 146–154. https://doi.org/10.1016/j.optlaseng.2016.10.006. 2017.

[20] C. Li, G. Luo & K. Qin. "An image encryption scheme based on chaotic tent map". Nonlinear Dynamics. https://doi.org/10.1007/s11071-016-3030-8.2016.

[21] K. Mishra, R. Saharan, & B. Rathor, "A new cryptographic method for image encryption", 32, 2885–2892. https://doi.org/10.3233/JIFS-169231. 2017.

[22] M. H. Annaby, M. A. Rushdi, & E. A. Nehary. “Color image encryption using random transforms, phase retrieval, chaotic maps, and diffusion”. Optics and Lasers in Engineering, 103, 9–23. https://doi.org/10.1016/j.optlaseng.2017.11.005. 2018.

[23] S. M. Ismail, L. A. Said, A. G. Radwan, A. H. Madian & M. F. Abu-elyazeed. “Generalized double-humped logistic map-based medical image encryption”. Journal of Advanced Research, 10, 85–98. https://doi.org/10.1016/j.jare.2018.01.009. 2018.

[24] L. Zhang, X. Liao, X. Wang, “An image encryption approach based on chaotic maps”, Chaos Solitons & Fractals, Vol. 24, pp. 759-765, 2005.

[25] Barker, E., Roginsky, A., & Barker, E. (n.d.). “Transitioning the Use of Cryptographic Algorithms and Key Lengths”, NIST Special Publication 800-131A Transitioning the Use of Cryptographic Algorithms and Key Lengths. July 2018.

[26] A. Kulsoom, D. Xiao, A. Rehman, S. A. Abbas, “An efficient and noise resistive selective image encryption scheme for gray images based on chaotic maps and DNA complementary rules”, Multimed Tools Appl., DOI 10.1007/s11042-014-2221-x. 2014.

[27] Wu, Yue, Joseph P. Noonan, and Sos Agaian. "NPCR and UACI randomness tests for image encryption." Cyber journals: multidisciplinary journals in science and technology, Journal of Selected Areas in Telecommunications (JSAT) 1.2: 31-38, (2011).

[28] Y. Wang, K. Wong, X. Liao & G. Chen. “A new chaos-based fast image encryption algorithm”, 11, 514–522. https://doi.org/10.1016/j.asoc.2009.12.011. 2011.

[29] A. Hanan,R. Enayatifar & M. Lee., “A hybrid genetic algorithm and chaotic function model for image encryption. AEUE - International Journal of Electronics and Communications, 66(10), 806–816. https://doi.org/10.1016/j.aeue.2012.01.015. 2012.

[30] Z. Zhu, W. Zhang, K. Wong & H. Yu. “A chaos-based symmetric image encryption scheme using a bit-level permutation”. Information Sciences, 181(6), 1171–1186. https://doi.org/10.1016/j.ins.2010.11.009. 2011.

[31] G. Zhang & Q. Liu. “A novel image encryption method based on total shuffling scheme”. OPTICS, 284(12), 2775–2780. https://doi.org/10.1016/j.optcom.2011.02.039. 2011.

[32] A. Akhshani, A. Akhavan, S. C. Lim, Z. Hassan, “An image encryption scheme based on quantum logistic map”, Communications in Nonlinear Science and Numerical Simulation, Vol. 17, pp. 4653-4661, 2012.

[33] S. M. Seyedzadeh, S. Mirzakuchaki, “A fast color image encryption algorithm based on coupled two-dimensional piecewise chaotic map, “Signal Processing, Vol. 92, pp. 1202-1215, 2012.

[34] E. Vaferi, R. Sabbaghi-Nadooshan, “A New Encryption Algorithm for Color Images based on Total Chaotic Shuffling Scheme”, Optik - International Journal for Light and Electron Optics, Volume 126, Issue 20, Pages 2474–2480, October 2015.

[35] M. Mohammadizadeh, B. Pourabbas, K. Foroutani, M. Fallahian, “Conductive polythiophene nanoparticles deposition on transparent PET substrates: effect of modification with hybrid organic-inorganic coating”, International Journal of Engineering (IJE), TRANSACTIONS C: Aspects Vol. 28, No. 4, (April 2015) 567-572.

Ardalan Ghasemzadeh received the B.Sc. degree in software engineering from the Kharazmi University, Tehran, Iran, in 2001 and the M.Sc. degree in Artificial Intelligence and Robotics from Shiraz University, Shiraz, Iran, in 2003. He is currently Ph.D. Candidate in Department of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran. Since 2012, he served as a lecturer at Computer Engineering & Information Technology, Urmia University of Technology, Urmia, Iran. His research interests are Image & Audio processing, machine learning, deep learning & cryptography. His email address is: a.ghasemzadeh@uut.ac.ir.

Omid Reza Bolouki Speily received the B.Sc. degree in Computer Engineering from Urmia University, the M.Sc. & Ph.D. degrees in Information Technology from the AmirKabir University of Technology. He worked as a researcher at the Iran Telecommunication Research Center (ITRC). Since 2009 he joined the Urmia University of Technology as a faculty member of Information Technology & Computer Engineering Department. His research interest includes the dynamics complex networks, graph theory, intelligent system, e-Services. His email address is: speily@uut.ac.ir.

An SRN Based Approach for Performance Evaluation of Network Layer in Mobile Ad hoc Networks

Meisam Yadollahzadeh Tabari*
Department of Computer Engineering, Babol Branch, Islamic Azad University, Babol, Iran
m_tabari@baboliau.ac.ir
Ali.A Pouyan
Department of Computer & IT Engineering, Shahrood University of Technology, Shahrood, Iran
apouyan@shahroodut.ac.ir

Received: 02/Aug/2018 Revised: 28/Feb/2018 Accepted: 19/May/2019

Abstract

The application of mobile ad hoc networks in emergency and critical cases need a precise and formal performance evaluation of these networks. Traditional simulators like NS-2 and OPNET usually need considerable time for producing high-level performance metrics. Also, there is no theoretical background for mentioned simulators. In this research, we propose a framework for performance evaluation of mobile ad hoc networks. The presented framework points to the network layer of MANETs using Stochastic Reward Nets modeling tool as a variation of Generalized Stochastic Petri Nets. Based on the decomposition technique, it encompasses two separate models: one for analysis of data flowing process and the other for the modeling routing process. For verifying the presented model, an equivalence-based method is applied. The proposed model has been quantified by deriving two performances metrics as the Packet Delivery Ratio and End-to-end Delay. The results show the obtained values from the presented model well matched to the values generated from the NS-2 simulator with considerable shorter execution time.

Keywords: Mobile ad Hoc Network; Stochastic Reward Net; Performance Evaluation; Modeling.

1. Introduction

Mobile ad-hoc network (MANET) is a dynamic and scalable type of networks, which is free from the constraints of infrastructure. Mobile ad hoc networks are becoming very attractive and useful in many kinds of communication and networking applications. This is due to their efficiency, the simplicity of installation and use, low relative cost, and the flexibility provided by their dynamic infrastructure [1]. High performance is a fundamental goal in designing communication systems such as MANETs. Therefore, the performance evaluation of ad hoc networks is needed to compare various network architectures and protocols for their performance and to study both the effect of varying network parameters and the interaction between them [2,3]. It should be noted that most researches accomplished in the area of MANETs performance evaluation, utilized broadband of Discrete Event Simulators (DES) such as NS2 [4], OPNET [5], and GLOMOSIM [6]. The principal drawback of DES models is the time and resources needed to run such models for large realistic systems, especially when highly accurate results (i.e., narrow confidence intervals) are desired. In some cases, the results also differ from one simulator to another [7].

In addition to a large amount of computation time, it is challenging to study high-level typical specifications such as deadlock and concurrency in MANETs using DES. This is because the network simulators implement the network at a low level of abstraction and specifications at a higher level cannot be supported well. Also, the most critical challenge that simulators face is the lack of any scientific and dependable modeling tool for depicting the correctness of the model [8]. Due to the advantages of quick construction, numerical analysis, and the ability of high-level performance evaluation, analytical modeling techniques, such as stochastic Petri nets, process algebra, and Markovian modeling have been used for the performance analysis of communication systems.

However, there are a few practical issues associated with analytical modeling tools that need addressing. First of all, modeling a real system requires a detailed concept of model formalism and the description of the system. Modeling tools usually emphasis on the stochastic behaviors of the investigated system. The probability of a packet reaching the destination, the likelihood of possible routes and the probability of a link failure are essential to be investigated. State space explosion problem is another major problem for modeling most of the complicated systems. An analytical model that is designed for a system should have two specific characteristics. First, it should be detailed enough that describes all the features and behaviors of the investigated system. Second, the model should not contain too many places, which increase the possibility of trapping in the state explosion problem.

Petri nets [10], and its variations like Stochastic Petri Net, Generalized Stochastic Petri Net, Colored Petri Net and etc.) as a known modeling tool that our work is based on, currently are widely used for modeling computer networks. In the definition, it includes places which can

* Corresponding Author

point to the states of a system and tokens in each of them. The latter shows in which state a system is. Transitions along a pair of states represent system' dynamics, which cause the events. It allows characters like synchronization, concurrency, conflict, and mutual exclusion, which are features of communication protocols.

Compared to mathematical and Markov chains models, stochastic Petri nets models can easily be modified to cope with changes in a modeled system [21]. Also, the effectiveness of stochastic Petri nets has been demonstrated for modeling complex communications protocols. E.g., this formalism can deal with more than one data stream at a time, which offers excellent clarity for modeling distributed and parallel systems [11].

There are two distinct approaches for modeling a communication system with Petri net. In the first approach, each real network workstation bounds to a place in the Petri net model. Data flow between workstations is represented via a transition in Petri net. However, the main issue in this modeling approach is that the size (i.e., a high number of nodes), which usually leads to the state explosion problem that affects the performance of the proposed model

For evaluating the performance of a network with the Petri net model, it is recommended to apply a behavioral approach, rather than considering each network node as a Petri net place. In this approach, the number of nodes in the network does not affect the structure (i.e., size) of the proposed model. The model considers the functionality of each node and the interaction between them. In this paper, a representation of a behavioral framework for performance analysis of a mobile ad hoc network is proposed, which is based on the concept of decomposition using Stochastic Reward nets (SRN) with the emphasis on the behaviors of a node in the network layer

2. Related Works

There are not too many investigations, performed on performance evaluation of mobile ad hoc networks using analytical modeling. Most of the literature use analytical modeling tools like the Markovian model, and a few of them use Petri net. Here, we performed a brief review of some significant research conducted in this area.

In [12] the authors modeled all stations in an IEEE 802.11 based WLAN in one SPN model. The complete model was solved using simulation because it was too large for direct analytical analysis, due to state space explosion. Although the authors introduced two compact analytical models, they did not include some aspects of the IEEE 802.11 DCF protocol.

An approximate stochastic Petri net model for ad hoc network was presented in [13]. The proposed model tried to take advantage of the symmetry between nodes by describing the behavior of one node under a workload that is generated by the whole network. The SPN model consists of two subnets; incoming and outgoing subnets. The incoming subnet represents the processing of packets received from other nodes, whereas the outgoing subnet models the transmission of packets generated in the current node. The model is so simple and seems not to be able to represent detailed characteristics of mobile ad hoc network. Xiong et al. [14] modeled and simulated ad hoc routing protocol using colored Petri Nets (CPNs). They used topology approximation mechanism to solve the problem of topology changes, which is an essential character of ad-hoc network. Ciardo et al. [15] modeled a scalable, high speed interconnect, which is a continuous hexagonal mesh-like wired network, with stochastic Petri Nets. They presented both exact and tractable approximate SPN model and compared it with simulation results based on CPNs. In [9] the authors perform a performance study of the distributed coordination function of 802.11 networks. They also illustrate the different classes of Petri Nets used for modeling network protocols and their robustness in modeling based formal methods. Their proposed model uses on Object-oriented Petri Nets for modeling IEEE 802.11b which considers Back-Off procedure and time synchronization. Then, performance analyses are evaluated by simulation for a dense wireless network and compared with other measurements approaches. Their model assessed using different metrics such as collision rates, the transition time. It does not mention how the performance metrics are derived from the model.

The authors in references [2,21] performed the most serious work for performance evaluation of mobile ad hoc network using Petri net. They proposed a general framework for performance analysis of MANET using SRN; a variation of GSPN. The model consists of two separate models for Data Link and network layer in addition to three mathematical models for spotting nodes behaviors in MANET. In their proposed SRN model for the network layer, there is no sign of routing protocol which is the most drawback of this job. Also, the model is only investigated on a single performance metric as Goodput.

In [7, 16] the authors present a node based Petri-net simulation model of a wireless mobile ad hoc network. They claim that the model covers all the fundamental aspects of the behavior of a network and uses a novel scheme of orientation-dependent (or sector-dependent) internode communication, with random states of links. Their proposed scheme enables the representation of reliability aspects of wireless communication, such as fading effects, interferences, the presence of obstacles and weather conditions. The simulation model was implemented in terms of a class of extended Petri nets to explicitly represent parallelism of events and processes in the WLAN as a distributed system. Because the presented model was a node-based model, it suffers from the state explosion problem along increasing in the number of nodes.

Also, in some investigations like [8,17], a class of Petri net is used the name as Fuzzy-Petri net, in which a type of fuzzy inference accomplished for firing each transition. In [8] the authors proposed a secure routing protocol for mobile ad hoc network based on fuzzy Petri modeling tool. In their presented approach, each network node assigned to a place in the Petri net model. The link between any pair of nodes modeled by a transition. For each transition, four distinct fuzzy parameters are assigned to encounter the security levels of links and influenced nodes. The firing of each transition produces a security coefficient propagated along with the link from source to destination. For a network with a large number of nodes, this usually leads to the state explosion problem.

3. Stochastic Reward Nets

Basically, (SRN) is a kind of GSPN formalism [11]. GSPN is also innovated based on SPNs. SPN [10] is a class of Petri net in which, a firing delay is associated with each transition. Then, the transitions will fire after a probabilistic delay determined by a random variable. Stochastic Petri Nets were mainly proposed for quantitative analysis (performance evaluation) of the complex discrete event systems. A stochastic Petri net is defined with a five-tuple; SPN = (*P, T, F, M0, Λ*), where

P : is a set of states, called places.

T : is a set of transitions.

F : where $F \subset (P \times T) \cup (T \times P)$ is a set of flow relations called "arcs" between places and transitions (and between transitions and places).

M0 : is the initial marking.

Λ : is the array of firing rates, λ associated with the transitions. The firing rate, a random variable, can also be a function $\lambda(M)$ of the current marking.

In Generalized Stochastic Petri net (GSPN), each transition has an associated firing time, which can be zero (immediate transition) or exponentially distributed with a parameter dependent on the marking (timed transition). Automatically, an immediate transition has priority over timed transitions. Timed transitions have exponential distribution firing time function with rate λ. Having a transition with a constant time *T*, the firing rate is set to *1/T*. Furthermore; GSPN uses the advantage of inhibitor arcs. This arc joints a transition to a place with a tiny circle sticking to place. For firing this transition, no token should be held in directed place. This simple ability causes a produced model more powerful with a fewer number of places and transitions. SRNs as a superset of GSPNs increases the modeling power of the GSPN by adding guard functions, marking dependent arc multiplicities, general transition priorities, and reward rates at the net level. A guard function is a Boolean function associated with a transition. Whenever the transition satisfies all the input and inhibitor conditions in a marking *M*, the guard is evaluated. The transition is considered enabled only if the guard function evaluates to true. Marking dependent arc multiplicities allow either the number of tokens required for the transition to be enabled, or the number of tokens removed from the input place, or the number of tokens placed in an output place to be a function of the current marking of the PN. Such arcs are called variable cardinality arcs. As the simplest way, in each SRN model, a set of performance metrics can be derived using a combination of specifications related to transitions and places. For transitions, characteristics like, Throughput: which is the number of times the transition fires per time unit and Probability of firing is used. Average marking and Steady-state distribution of tokens are also used for spotting characteristics of places. The combination of evaluation tools directly depends on the modeled system and may vary from one system to another. Each graphical SRN model also has structurally algebraic properties known as P-Invariant and T-Invariant. Those could usually represent high-level behaviors of a system which is modeled by Petri net [10, 11]

4. Model Description

The topology of MANET is highly dynamic because of frequent mobility of nodes. Thus, there are many inter-dependent parameters, mechanisms, and phenomena which should be considered in the presented model. Two requirements should be fulfilled in advance to present an approach for the modeling and analysis of large-scale ad hoc network systems. First, the model should be elaborated in detail to express all network characteristics that have a significant impact on its performance. This leads to a model with a large number of Places and Transitions which may trap in the state explosion problem. Second, it should be simple enough to be scalable, analyzable and also recognizable. It is clear that these two requirements are contradictory. Therefore, the approximate model presented based on the idea of decomposition [28]. In this technique, the model decomposes into two or more sub-models that are solved iteratively. The approximated model should work with a large number of nodes and basically describes the behavior of one node under a given workload that is generated by the whole ad hoc network. The approximated model covers all necessary activities from incoming and outgoing subnets representing different behavior of nodes from the perspective of a single node in the network layer. Despite general network parameters, routing protocol has a significant impact on the generated model.

Regarding AODV as the selected routing protocol, there are two distinct flows in a network. A flow for finding a path from the source node to the destination is known as the routing process and a kind of flow for delivering the data packet to a destination known as the data flow process. These two processes differ in a way that the latter uses a ready route that is produced by the first model. This inspired us to decompose the overall model into two separate models for data flow and routing process. Those are presented in figures 1&3. In both models, the places are the mappings of nodes' states in a network. Timed transitions stand for the required time for performing a task in the network. As an example, the time needed for delivering a packet from a node to its neighbor is represented by a timed transition.

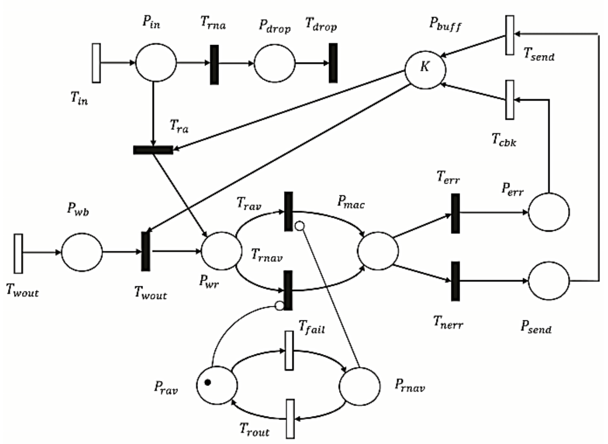


Fig. 1. SRN model for network data flow inspired by [2]

Immediate transitions are also used for declaring a conflict or choice in network behavior. The probabilities associated with these conflict transitions, a well-defined probability function should be assigned to each transition which is deduced from network behaviors.

A firing rate of transition or its probability can be obtained from network standards specifications or mathematical calculations. Sub-sections 4-1 and 4-2 give precise and detailed information for both models with the values associated with parameters.

4.1 SRN Model for Modeling Network Data Flow

As it is illustrated in Figure 1, the model which is designed for data flow process encompasses two sections, inspired by [2] the two mentioned sections considered for spotting incoming and out-coming data flows.

Out-coming flow directs to the actions performed for generating a packet in a node and delivering it to the destination. This flow started with the transition T_{out} in figure 1. The firing rate of this transition equals λ which is a network parameter, and it is the same as the packet generation rate in a network. Packets are usually generated using CBR traffic, and its rate can be interpreted as λ. A generated packet needs to buffer space for starting its routing process.

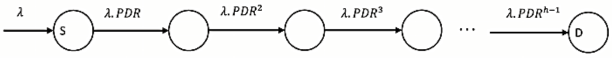


Fig. 2. The input λ data traffic rate that is produced from the source node along the path toward the destination

This is shown using place P_{wb} indicating packets which are waiting for occupying a buffer in the current node for flowing their data.

If there is at least one token from K supposed buffer space in network layer standard, this transition would be fired and flows the token to the place P_{wr} . Due to modeling issues, we investigate a network with established routes and ignored the first routing process which is needed at the beginning of a network lifetime. For an established route, there is a probability of route failure which is due to nodes movements or other issues. This is represented by T_{fail} which its firing rate is deduced from [18]. The authors in this research proposed a Petri net model for path connection availability for multihop ad hoc network. They investigated the effects of transmission range, network size, data transmission rate, and routing protocol. After occurrence a failure, a new routing process is needed which is shown using transition T_{route}. The required time for this transition is derived from the routing process model introduced in section 4 (B).

Along specifying route toward the destination and completing the routing process, the generated packet should be sent to the next node through the Data Link layer.

A token, in this case, injected to the place P_{MAC} which indicates that the MAC layer is ready for transmitting the data packet to the next ongoing node. IEEE 802.11 DCF as a known Data Link protocol is responsible for the nodes coordination in this layer. Despite, the advanced technique which is used in this protocol, a collision may occur for sending packets through a common channel in a neighbor area. The probability associated for this probability which is shown using the choice (T_{err}-T_{nerr}) is derived from an investigation performed in [21]. Also, the required time for transitions T_{send} and T_{CBK} which stand as the time needed for sending a data packet and sending Call Back error is also derived from this research. For incoming section, in AODV routing protocol, if a node receives a data packet which has no unexpired route for, it drops the Packet immediately. Otherwise, it delivers the packet to the MAC layer for sending the received packet to the next hop node in route toward the destination. The two possible outcomes are represented with a choice structure using transitions T_{ra} and T_{rna}. As stated, in our presented model it is supposed that all routes are initialized. Then, T_{rna} shows the event that a route is disturbed because of nodes movements or other conditions which may occur in a route failure.

The probability associated with T_{rna} is derived from our previous study in [3]. Then T_{ra} would be as 1- T_{rna}. As another parameter, the firing rate of transition T_{in} should be defined. This transition points to the data traffic which is forwarded toward a node for sending it to the next consequent node in an established route. The firing rate of this transition has direct relevance to the number of source nodes (M_S), their produced data traffic rate (λ) and the Packet Delivery Ratio (*PDR*) of each node in a route.

The obtained value is averaged entirely over the whole number of nodes in the network. The input λ data traffic rate that is produced from the source node is multiplied over the *PDR* value of each node along the path from the source node to the destination. Figure 2 illustrates the process. The overall expression for rate(T_{in}) expressed via equation1 using the mathematical calculations reported in [2]. The *PDR* value used in this equation calculated via equation 2. h points to the value of hop-count that its value produced from the instruction reported in [27]. $Thr(T_x)$ points to the throughput value of transition T_x.

$$rate(T_{in}) = \frac{M_S.\lambda.(PDR+PDR^2....+PDR^{h-1})}{M} \quad (1)$$

$$PDR = \frac{Thr(T_{send})}{Thr(T_{out})+(Thr(T_{in})\times Pr(T_{ra}))} \quad (2)$$

4.2 SRN Model for the Modeling Routing Process

For this section, the behavior of AODV routing protocol is modeled for finding a route to the destination. In AODV routing protocol, the routing process initiates through the propagation of RREQ messages to the neighbor's nodes with a TTL number indicating the dissemination range of such RREQ propagation. This is decremented by one for each transmission of RREQ from one node to another. RREQ passes through nodes, in order to reach the destination. This process terminated by reaching a node has a valid route toward destination. The whole process is modeled in Figure 2. As it can be seen in this figure T_{RREQ}, initiates the routing process. After that, it puts the TTL number of tokens to the place P_{TTL}. Firing of transition T_{TTL} also, decrement TTL count by one and sends the token to the place *Pwr2* for determining path availability through routing table of the next nodes which are in one-hop communication of the current node.

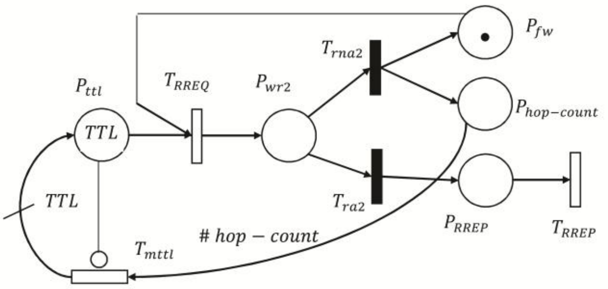


Fig. 3. SRN model for routing process

According to [23] Supposing Random Way Point (RWP) as our selected mobility model, and totally *M* number of nodes communicating in a network, the number of neighbor (*N*) in a neighbor area would be as equation 3.

$$N = N_{r_a}^{RWP} = \frac{r_a^2}{3}((4 - 2P_p + {P_p}^2) - \frac{4}{\pi}{P_p}^2 r_a - 3(1 - P_p){r_a}^2) \times \frac{\pi}{A} \times M \quad (3)$$

In this equation, r_a points to the transmission range of each node, A is network area, and P_p denotes to the probability that a node in a specific time be in its pause time duration and calculated as $P_p = \frac{\tau_P}{\tau_P + E(T)}$.

In that r_a points to the transmission range of each node, A is network area, and P_p denotes to the probability that a node in a specific time be in its pause time duration and calculated as $P_p = \frac{\tau_P}{\tau_P + E(T)}$.

Where $E(T)$ is also points to the expected time between two waypoints and estimated using the method introduced in [24] as equation 4. a stands for the network area dimension. V_{max} and V_{min} points to the nodes minimum and maximum speed, respectively. τ_P is also the ordinal pause time; the nodes stay in a position before moving to a new location. If all of the neighbor nodes act legitimately, there is a possibility that one or more of them have a route toward the destination and send back legitimate RREP. This is also represented by T_{ra} and T_{rna}. If one of the neighbor nodes has a valid route to the destination, the transition T_{ra} would be fired. This probability has direct relevance to the routing table size of each node over the whole number of nodes. Note that each entry in a routing table points to the next hop node in a path along to the destination. Average routing table size is derived from an investigation performed in [24]. Form this explanation $\Pr(T_{rna2})$ expressed as equation 5.

$$Pr(Trna2) = \left(1 - \frac{(RT\_Size)}{M}\right)^N \quad (5)$$

$$Pr(T_{rna2}) = 1 - Pr(T_{rna}).$$

Place $P_{hop\text{-}cunt}$ preserved for counting average the number of hops toward destination. Also, a guard function specified to transition T_{ttl} indicating that it would not fire nevertheless any number of tokens remained in places P_{wr2}, $P_{legitimate}$, P_{fw} and P_{RREP} . The time associated with transition T_{RREq} equals to the time needed for sending RREQ packet via a common broadcast channel. IEEE 802.11 DCF usually uses 2-handshake (BA) method for sending control packet toward the channel. In this approach, a node simply sends its low size data or control packet after a DIFS interval of sensing the channel free. Along receiving a packet from sender, the receiver node sends back an ACK frame to announce its successful reception after a SIFS interval of sensing the channel free. Then,

$$Time(T_{RREQ}) = \tau_{DIFS} + \tau_{SIFS} + \tau_{ACK} + \tau_{control\,packet} \quad (6)$$

τ_x equals to the time needed for sending packet *x* via broadcasting in common channel. Also, the time needed for firing T_{RREP} equals to the time needed for sending back RREP packet toward the node that missed the route. This is formulated as equation 7. $\#(P_{hop-count})$ stands for the average number of tokens in place $P_{hop-count}$.

$$Time(T_{RREP}) = (\tau_{DIFS} + \tau_{SIFS} + \tau_{ACK} + \tau_{control\,packet}) \times \#(P_{hop-count}) \quad (7)$$

The standard times τ_x which are used in equations 3,4 derived from the Data Link layer standard [21]. The time required for firing T_{route} in network data flow model which depends on rothe uting process model derived using equation 8. In this equation TTL is the assigned value for Time To Live, $\#(P_{ttl})$ is the average number of tokens in place P_{ttl}, $Thr(T_{ttl})$ stands for firing throughput for the transition T_{ttl}. T_{RREP} is also the taken time for the sending a single RREP message.

$$Time(T_{route}) = \frac{TTL - \#(P_{ttl})}{Thr(T_{ttl})} + T_{RREP}. \quad (8)$$

Table 1. Parametrs value related to equtions 3, 4

Parameter	value	Parameter	value
T_{DIFS}	$50\ \mu s$	T_{CTS}	$\frac{(112\ + phH)\ bit}{1\ Mbps}$
T_{SIFS}	$10\ \mu s$	T_{ACK}	$\frac{(112\ + phH)\ bit}{1\ Mbps}$
T_{RTS}	$\frac{(160\ + phH*)\ bit}{1\ Mbps}$	T_{Data}	$\frac{(Data\ + phH)\ bit}{2\ Mbps}$

∗ Physical Header =192 bit

5. Performance Evaluation & Model Validation

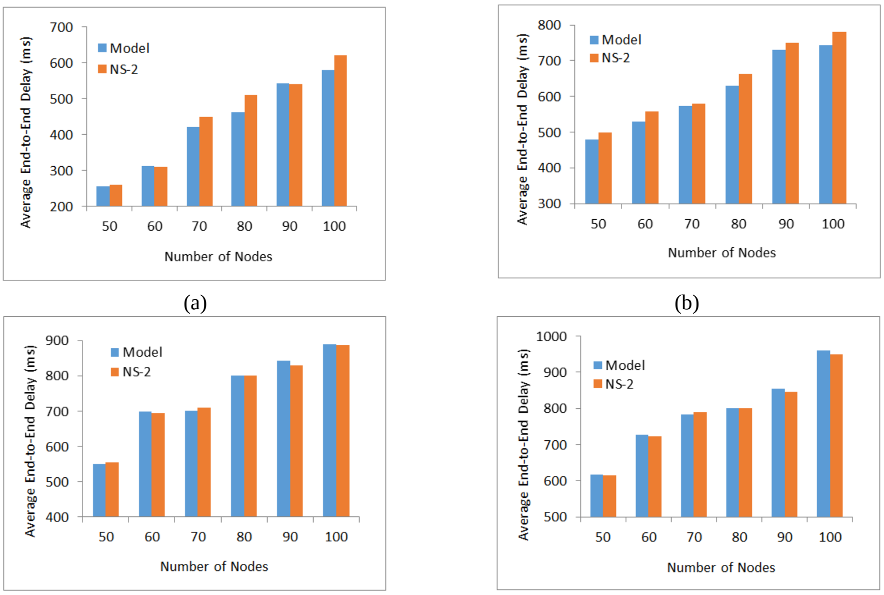
For measuring performance metrics and showing the correctness of our model, an equivalence-based method is used. For doing that, two performance metrics are quantified in both NS-2 SPNP PIPE simulator. SPNP [19] is an analytical environment specifically for modeling SRN. It includes all required components with all methods for analyzing an SRN model. *PDR* and *End-to-End delay* are the two metrics which are used. By definition *End-to-End delay* refers to the time needed for sending a packet from source node to the destination. It includes three different times. 1) The initial time required for the routing process, 2) The time required for delivering the packet from source node to the next 1-hop communication in the path. 3) The time required for delivering a packet from an intermediate node to the next one as it reaches the destination. According to *h* hop count from the source node to the destination, the latter time will be multiplied to *h*. Then, the overall expression for the End-to-End delay metric is as equation 9. In that, the expression $Time(T_{route})$ expressed using equation 8 and the two latter times, is represented using *little law*.

$$End\ to\ End\ Delay = [Time(T_{route})] + \left[\frac{\#(P_{in})\times Pr(T_{ra})}{Thr(T_{send})}\right] + \left[\frac{\#(P_{wb})}{Thr(T_{send})} \times h - 1\right] \qquad (9)$$

Also, the **PDR** is defined as the percentage of the number of packets sent by a constant bit rate from a sources node per the number of received packets by sink/destination [25,26]. This performance evaluation parameter measures the delivery reliability, effectiveness, and efficiency of the routing protocol. The expression for this metric is expressed in equation 2 (previous section).

5.1 Implementation

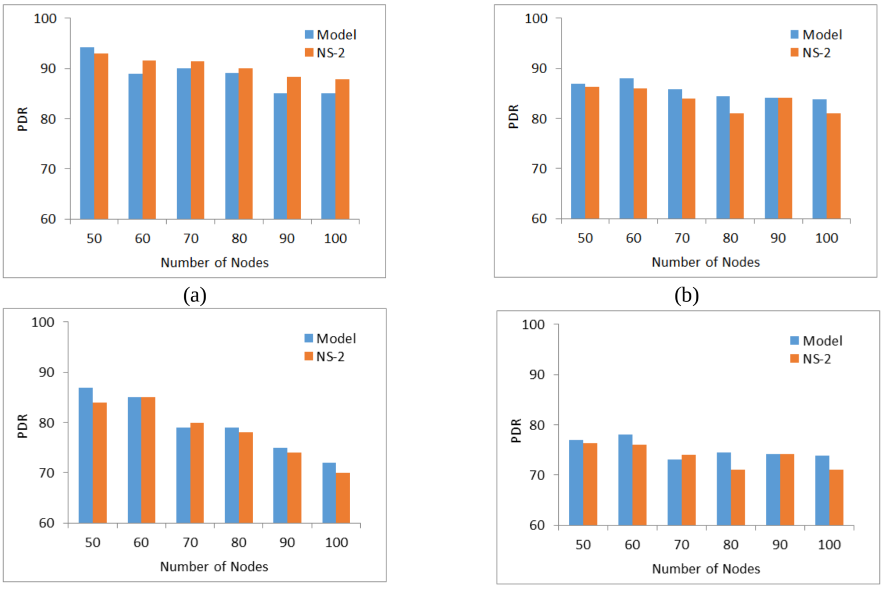
In the implementation phase, in a specific NS-2 scenario, mobile nodes were chosen for moving in an area as 1000 × 1000 dimension using the random waypoints mobility model. They move with a speed that is uniformly selected from 0 to 20 in a network with 2Mb bandwidth. The pause time (τ_P) is set to 120 (s). Each node has the transmission range of 250 m and carrier sensing range of 550 m. Network CBR is used to generate data packets. Four different packet generation rates (λ) investigated as 300 kbps, 600 kbps, 900 kbps, and 1.2 Mbps. Each metric is also investigated for 50, 60, 70, 80, 90 and 100 numbers of nodes. The Routing protocol is set to AODV as our assumption. Network CBR is used to generate data packets. For each traffic generation rate (λ), the CBR amount is adjusted to match this value. The required time for computing equations 6,7 which is related to the firing rate of transitions T_{RREQ} and T_{RREP} is also derived from the Data Link Layer standard. The assigned times for the transitions T_{DIFS}, T_{SIFS}, T_{RTS}, T_{CTS}, T_{ACK}, T_{DATA} which are used in equations 6,7 are elaborated in Table 1. Also, the K value as a parameter of the SRN model which direct to the number of free buffer in a node is set to 64 which is a standard value. For NS-2 each of metrics value averaged from 10 runs of each scenario. This is because some considerable fluctuations have been seen in each run of a specific scenario in this environment. Figures 4 and 5 show obtained values for the performance metrics End-to-End delay and *PDR* respectively. The obtained values plotted versus number of nodes for both NS-2 and presented SRN model.



(a) (b)

(c) (d)

Fig. 4. End-to-End delay versus number of nodes using NS-2 and model.
λ= a) 300 Kbps b) 600 Kbps c) 900 Kbps d) 1.2 Mbps



(a) (b)

(c) (d)

Fig. 5. PDR versus number of nodes using NS-2 and SRN model.
λ= a) 300 Kbps b) 600 Kbps c) 900 Kbps d) 1.2 Mbps

As it is clear, for all metrics the values obtained from the proposed SRN model is well matched to the value obtained from NS-2. This can prove the correctness of the presented model and analysis.

As shown in figure 4, increasing in the number of nodes leads to higher waiting time for the nodes in order to transmit their data packets. Through preparing the results it is observed that increasing in the number of nodes, raises the probability of failure for forwarding a packet through the Data Link layer. Also rising the number of nodes makes more firing rate for the transition T_{fail} which in turn cause more firing count for the transition T_{route} in the model. Unlike, this phenomenon leads to a shorter firing time for the transition T_{route}. This is because increasing in the number of nodes leads to the more number of nodes in a neighboring area which raises the chance for a requester node to get RREP from one of its neighbors. This probability is shown using the choice structure (T_{ra2} – T_{rna2}) in the routing process model. Totally, the results show that the mentioned outcomes lead to a more End-to-End delay by increasing the number of nodes. Packet generation rate (λ) has a similar effect on this metric. This increase in the packet generation rate in the Data Link Layer which leads to the more firing probability for the transition T_{err}. This also increases the firing rate for the transitions T_{in} and T_{out} which accumulate more tokens in place T_{MAC}. As our investigation, it does not effect on firing time of transition T_{route}.

As it is clear in Figure 5 it can be seen a decreasing trend for the *PDR* value, in terms of increasing in the number of nodes. For this metric, also the two distinct values obtained from SRN and NS-2 is jointly agreed together. Like previous, raising the number of nodes and packet generation rate, raises the probability of failure for forwarding a packet through the Data Link layer. This makes more firing rate for the transition T_{fail} and raises the probability value for transition T_{rerr}. This gives lower firing throughput for the transition T_{send} which reduces *PDR* value totally.

As our major claim about the spanned time for deriving performance metrics from the SRN model in comparison to the NS-2 network simulator, a brief discussion presented here about how long those are taken. For the SRN model, the solution time (the time needed to generate the Markov chains model and computing the required performance metrics) for both presented models is highly related to the number of generated states. The number of states of the *data flowing process model* is highly depended on firing rates of the transitions T_{in}, T_{out}, and T_{fail}. In its worst case, it doesn't hit 30 (s). On the other hand, increasing in the number of nodes, growths the solving time in NS-2 simulator, exponentially. In some cases, it takes about one hour for generating a result from NS-2 network simulator.

6. Conclusions

In this research, an SRN model presented for performance evaluation of network layer in mobile ad hoc networks. The model encompasses two separate models. One model stands for data flow process and another for the routing process which is based on AODV routing protocol. For verifying the correctness of the presented model, an equivalence-based method used. For this matter, two known performance metrics as End-to-End Delay and *PDR* are derived from the model and compared to the values that are obtained from NS-2. The results showed that the values obtained from the presented SRN model well matched to the values driven from NS-2 results. This can show the correctness of the presented model. Spanned time was quite negligible for presented SRN model compared to the time needed for NS-2. The presented SRN model can be used and expanded more as a general framework for any analysis and extracting MANET behaviors in the future.

References

[1] Tabari MY, Hassanpour H, Pouyan A, Saleki S. Proposing a light weight semi-distributed IDS for mobile ad-hoc network based on nodes' mode. In6th International Symposium on Telecommunications (IST) 2012 Nov 6 (pp. 948-953). IEEE.

[2] Younes O, Thomas N. Modelling and performance analysis of multi-hop ad hoc networks. Simulation Modelling Practice and Theory. 2013 Nov 1;38:69-97.

[3] Pouyan A, Yadollahzadeh Tabari M. Estimating reliability in mobile ad-hoc networks based on Monte Carlo simulation. International Journal of Engineering. 2014 Jan 21;7(5):739-46.

[4] "The Network Simulator ns2," available at http://www.isi.edu/nsnam/ns/.

[5] "OPNET Modeler," available at http://www.opnet.com.

[6] Zeng X, Bagrodia R, Gerla M. GloMoSim: a library for parallel simulation of large-scale wireless networks. InProceedings. Twelfth Workshop on Parallel and Distributed Simulation PADS'98 (Cat. No. 98TB100233) 1998 May 26 (pp. 154-161). IEEE.

[7] Kostin A, Oz G, Haci H. Performance study of a wireless mobile ad hoc network with orientation - dependent internode communication scheme. International Journal of Communication Systems. 2014 Feb;27(2):322-40.

[8] Pouyan AA, Yadollahzadeh Tabari M. FPN - SAODV: using fuzzy petri nets for securing AODV routing protocol in mobile Ad hoc network. International Journal of Communication Systems. 2017 Jan 10;30(1):e2935.

[9] Masri A, Bourdeaud'Huy T, Toguyeni A. Performance analysis of IEEE 802.11 b wireless networks with object oriented Petri nets. Electronic Notes in Theoretical Computer Science. 2009 Jul 13;242(2):73-85.

[10] Marsan MA, Balbo G, Conte G, Donatelli S, Franceschinis G. Modelling with generalized stochastic Petri nets. John Wiley & Sons, Inc.; 1994 Oct 1.

[11] Tang Y, Chen L, He KT, Jing N. SRN: an extended Petri-net-based workflow model for Web service composition. InProceedings. IEEE International Conference on Web Services, 2004. 2004 Jul 6 (pp. 591-599). IEEE.

[12] German R, Heindl A. Performance evaluation of IEEE 802.11 wireless LANs with stochastic Petri nets. InProceedings 8th International Workshop on Petri Nets and Performance Models (Cat. No. PR00331) 1999 (pp. 44-53). IEEE.

[13] Zhang C, Zhou M. A stochastic Petri net-approach to modeling and analysis of ad hoc network. InInternational Conference on Information Technology: Research and Education, 2003. Proceedings. ITRE2003. 2003 Aug 11 (pp. 152-156). IEEE.

[14] Xiong C, Murata T, Tsai J. Modeling and simulation of routing protocol for mobile ad hoc networks using colored petri nets. InProceedings of the conference on Application and theory of petri nets: formal methods in software engineering and defence systems-Volume 12 2002 Jun 1 (pp. 145-153). Australian Computer Society, Inc.

[15] Ciardo G, Cherkasova L, Kotov V, Rokicki T. Modeling a scalable high-speed interconnect with stochastic Petri nets. InProceedings 6th International Workshop on Petri Nets and Performance Models 1995 Oct 3 (pp. 83-92). IEEE.

[16] Kostin A, Oz G, Haci H. Performance study of a wireless mobile ad hoc network with orientation-dependent inter-node communication links. In2009 24th International Symposium on Computer and Information Sciences 2009 Sep 14 (pp. 316-321). IEEE.

[17] Chiang TC, Tai CF, Hou TW. A knowledge-based inference multicast protocol using adaptive fuzzy Petri nets. Expert Systems with Applications. 2009 May 1;36(4):8115-23.

[18] Younes O, Thomas N. A path connection availability model for MANETs with random waypoint mobility. InComputer Performance Engineering 2012 Jul 30 (pp. 111-126). Springer, Berlin, Heidelberg.

[19] Ciardo G, Muppala J, Trivedi K. SPNP: stochastic Petri net package. InProceedings of the Third International Workshop on Petri Nets and Performance Models, PNPM89 1989 Dec 11 (pp. 142-151). IEEE.

[20] Younes O, Thomas N. An SRN model of the IEEE 802.11 DCF MAC protocol in multi-hop ad hoc networks with hidden nodes. The Computer Journal. 2011 Jun;54(6):875-93.

[21] Yadollahzadeh Tabari M, Pouyan AA. Misbehavior analysis of IEEE 802.11 MAC layer in mobile ad hoc network using stochastic reward nets. International Journal of Communication Systems. 2017 Nov 10;30(16):e3385.

[22] Soltani MD, Purwita AA, Zeng Z, Haas H, Safari M. Modeling the random orientation of mobile devices: Measurement, analysis and LiFi use case. IEEE Transactions on Communications. 2019 Mar;67(3):2157-72.

[23] Bettstetter C. On the connectivity of ad hoc networks. The computer journal. 2004 Jan 1;47(4):432-47.

[24] Xu J. On the fundamental tradeoffs between routing table size and network diameter in peer-to-peer networks. InIEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428) 2003 Mar 30 (Vol. 3, pp. 2177-2187). IEEE..

[25] Hu X, Jiao L, Li Z. Modelling and performance analysis of IEEE 802.11 DCF using coloured Petri nets. The Computer Journal. 2016 Oct; 59(10):1563-80.

[26] Erbas F, Kyamakya K, Jobmann K. Modelling and performance analysis of a novel position-based reliable unicast and multicast routing method using coloured Petri nets. In2003 IEEE 58th Vehicular Technology Conference. VTC 2003-Fall (IEEE Cat. No. 03CH37484) 2003 Oct 6 (Vol. 5, pp. 3099-3104). IEEE.

[27] Hieu TD, Choi SG. Simulation modeling and analysis of the hop count distribution in cognitive radio ad-hoc networks with shadow fading. Simulation Modelling Practice and Theory. 2016 Dec 1;69:43-54.

[28] Ciardo G, Trivedi KS. A decomposition approach for stochastic Petri net models. InProceedings of the Fourth International Workshop on Petri Nets and Performance Models PNPM91 1991 Dec 2 (pp. 74-83). IEEE.

Meisam Yadollahzadeh Tabari: Received his Ph.D. degree in Artificial intelligence from Shahrood University of technology in 2016. He also received his M.Sc. and B.Sc. degree from Arak and Iran University of science, respectively. Now he is Faculty member in the department of computer engineering of Islamic Azad university-Babol branch. His research interests include Performance evaluation, Analytical modeling and Data mining.

Ali.A pouyan: Dr. Ali A. Pouyan is currently with the department of computer engineering in Shahrood University of technology, Iran. He received his Ph.D. in computer engineering from Swinburne University of Technology, Melbourne, Australia. He completed his post-doctoral program at Industrial Research Institute Swinburne (IRIS) in control software. Dr. Pouyan published several research articles in international journal and conference papers. He also collaborated in several national industrial projects including telemetry of vital human signals, GSM platform, and designing and implementation of a software package for location problem. He is currently engaged in several research projects, including Petri net modeling, wireless networks, smart home, machine learning, and signal processing.

Toward an Enhanced Dynamic VM Consolidation Approach for Cloud Datacenters Using Continuous Time Markov Chain

Monireh Hosseini Sayadnavard
Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
mo.sayadnavard@gmail.com
Abolfazl Toroghi Haghighat*
Department of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
haghighat@qiau.ac.ir
Amir Masoud Rahmani
Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
rahmani@srbiau.ac.ir

Received: 08/Jan/2019 Revised: 18/Mar/2019 Accepted: 23/Apr/2019

Abstract

Dynamic Virtual Machine (VM) consolidation is an effective manner to reduce energy consumption and balance the resource load of physical machines (PMs) in cloud data centers that guarantees efficient power consumption while maintaining quality of service requirements. Reducing the number of active PMs using VM live migration leads to prevent inefficient usage of resources. However, high frequency of VM consolidation has the negative effect on the system reliability and we need to deal with the trade-off between energy consumption and system reliability. In recent years many research work has been done to optimize energy management using power management techniques. Although these methods are very efficient from the point of view of energy management, but they ignore the negative impact on the system reliability. In this paper a novel approach is proposed to achieve a reliable VM consolidation method. In this way, a Markov chain model is designed to determine the reliability of PMs and then it has been prioritized PMs based on their CPU utilization level and reliability status. Two algorithms are presented to determining source and destination servers. The efficiency of our proposed approach is validated by conducting extensive simulations. The results of the evaluation clearly show that the proposed approach significantly improve energy consumption while avoiding the inefficient VM migrations.

Keywords: Cloud Computing; VM Consolidation; Energy Efficiency; Reliability; Markov Chain.

1. Introduction

Cloud computing as one of the most interesting developments in technology is an on-demand computing model that provides services to users through Internet [1]. The ever increasing demand for computing resources has resulted in establishment of large-scale data centers, which require enormous amount of power and hence consume a lot of energy. Statistics of the worldwide data center electricity consumption show non-linear growth during the last decade and a similar trend is expected for the upcoming years [2]. The amount of computing resources and the inefficient use of these resources could lead to huge energy wastage. An effective way to improve the resource utilization and energy efficiency in cloud data centers is VM consolidation that has been widely studied in recent years [3,4]. During consolidation, VMs are periodically reallocated using live migration according to their current resource demand to minimize the number of active physical servers and the idle PMs are switched to low power modes to reduce the energy consumption [5]. Since most modern applications experience dynamic patterns of resource consumption because of highly variable workloads, VM consolidation in clouds is a complicated operation. Unconstrained VM consolidation may lead to degraded performance when an application is faced with increasing demand and resource usage. If the resources requirements of an application are not met, the response time will increase. While the QOS guarantee defined in the Service Level Agreements (SLAs) between the Cloud provider and their users is essential. Hence, the Cloud providers must consider the trade-off between performance and energy consumption in order to fulfil QOS requirements [6].

On the other hand, high VM consolidation has the negative effect on the reliability of the system [4,7,8]. most existing research on consolidation has focused on the performance-energy trade-off. While there are some works that consider the relationship of system reliability and energy efficiency in Cloud environment [8,9] and still there is a distinct need for more research on the mentioned challenge.

Server consolidation may increase the probability of server failure and compromise the reliability of the system by increasing the load on some servers and shutting down some of them. Hence, we need to consolidate servers in flexible manner with considering both energy efficiency and reliability to cover different operating conditions and scenarios.

In this paper we present a novel approach to dynamic VM consolidation by considering both reliability and

* Corresponding Author

energy efficiency simultaneously. We try to manage energy consumption along with considering the reliability of each PM in every phase of consolidation to reach equilibrium between these two metrics. we have calculated the reliability under the probability of failures occurrence in a heterogeneous environment. A computational model for PMs reliability prediction is presented and based on the results of this phase and the CPU utilization level of each PM, they are divided into different categories. Then, we can make proper decision to VMs migration to realization of our purpose. Specifically, the major contributions of this paper can be summarized as below:

- Designing a Markov chain model for predicting and analysing PMs reliability for VM consolidation implementation to optimize the relationship between energy efficiency and reliability.
- Propose a new policy to target PMs selection in dynamic VMs consolidation process to improve energy efficiency while considering the reliability factor of PMs.

The remainder of the paper is organized as follows: The current and past research on VM consolidation are reviewed in Section 2. The system model and problem statement are explained in Section 3. Our Markov Chain based reliability model and the proposed approach for dynamic VM consolidation are described in detail in Section 4. The experimental setup and results are shown and discussed in Section 5. Finally, conclusions are presented in the last section.

2. Related Works

There are several research works that addresses the VM consolidation. In this section we review relevant approaches in the literature related to the similar issues.

Many studies formulated the VM consolidation as a well-known NP-hard bin packing problem [6, 10-13]. Various heuristics like greedy algorithms are utilized to approximate the optimal solution of this NP-hard problem. These include worst fit and best fit in [10], first fit decreasing(FFD) and best fit decreasing (BFD) [11]. The authors in [12] have divided VMs consolidation into the four following phases: host overload detection, selection of VMs should be migrated, VM placement, and running PMs shrinking. Due to the complexity of VMs consolidation, the VMs consolidation issues in [12] were separated into several sub-problems, and then they have proposed novel adaptive heuristics for each sub-problem. They proposed MBFD algorithm to VM placement by considering power consumption and SLA violation. In this algorithm, the VMs are first sorted in decreasing order based on their utilizations. Then, these VMs are allocated to the hosts having minimum increase of energy consumption.

In [13], a VM consolidation framework is proposed to minimize the performance-energy trade-off. The VM placement problem is resolved using semi-online multidimensional bin packing. The authors in [14] have considered rack, cooling structure and network topology when consolidating VMs. In this paper the MBFD algorithm is improved and then three structure-aware VM placement methods are proposed to consolidate VMs in the servers to minimize the number of active racks that results in turning off idle routing and cooling equipment in order to reduce the energy consumption. In [15] a burstiness-aware server consolidation algorithm, QUEUE, is proposed. First, the burstiness of workload is captured using a two-state Markov chain, then some extra resources on each PM is reserved to avoid live migrations. Shen et al. in [16] proposed a mechanism that predicts the VM resource utilization patterns and consolidates complementary VMs with spatial/ temporal awareness into one PM to reduce the number of PMs, maximize resource utilization and reduce the number of VM migrations. Complementary VMs are the VMs whose total demand of each resource dimension in the spatial space nearly reaches their host PM’s capacity during VM lifetime period in the temporal space.

In [17] DVFS-aware consolidation procedure is presented to eliminate the inconsistencies between consolidation and DVFS techniques. this paper also has proposed PSFWT as a fuzzy DVFS-aware multi criteria and objective resource allocation solution for VM placement in Cloud data centers that simultaneously optimizes important objectives including energy consumption, SLA violation, and number of VM migrations. different criteria of the system including CPU, RAM, and network bandwidth in decision making process is considered.

Beloglazov and Buyya [5] investigated the problem of overloaded hosts detection using a Markov chain model. A specified QoS goal is defined to maximizing the mean time between VM migrations for any known stationary workload. The unknown nonstationary workloads are also handled using a multi size Sliding Window workload estimation. In [18] a heuristics based multi-phase approach for server consolidation is proposed which effectively reduces residual resource fragmentation along with reducing the number of active PMs. Residual Resource Fragmentation refers to the state where sufficient amount of residual resources is available but are fragmented and distributed across multiple active PMs. In [19] a VMs placement algorithm is proposed that considers computation resources, Quality of Service (QoS) metrics, virtual machine status and I/O data with priority based probability queuing model. Data location during Virtual machines placement is considered to avoid unnecessary migration to gain high performance for applications. The authors in [20] studied the influence of four aspects on energy consumption and QoS, namely, the dynamic workload, CPU utilization, times of VM migrations, and opportunity of VM migration from nine related factors. They created a Bayesian Network based estimation model (BNEM) for dynamic VM migration using these factors that each node represents one aspect of VM migration. Khani et al. [21] proposed a distributed mechanism for dynamic consolidation of virtual machines using a non-cooperative game for reducing power consumption in data centers with heterogeneous PMs. In [22] a prediction-based consolidation approach is proposed that considers both an estimation of future requested resources using Kernel Density Estimation technique, and future migration traffic to decrease the number of migrations.

There are also various metaheuristic algorithms that have been proposed to solve the VM consolidation problem in cloud computing environments. These algorithms rely on a probabilistic approach to find near optimal solutions to the problems. In [23], ant colony optimization method (ACO) is used to pack the VMs into the least number of physical machines while preserving Quality of Service requirements. A multi-objective function is defined that considers both the number of dormant PMs and the number of migrations. The GABA approach [24] is a genetic algorithm (GA) based algorithm that dynamically finds the optimum reconfiguration for a set of VMs according to the predicted future demand of the running workload. The algorithm decreases the number of PM significantly and converges within reasonable time. In [25], a VM consolidation approach is proposed based on the particle swarm optimization (PSO) algorithm, which considered reducing energy consumption and improving resource utilization in the data center as the optimization objective. In [26], a nonlinear model is introduced to quantify PM power consumption and then VM placement is formulated as a bi-objective optimization problem, which is solved using an ACO based algorithm.

Deng et al. in [8] presented a Reliability-Aware server Consolidation stratEgy (RACE) to address a multi-objective problem with considering hardware reliability and energy efficiency. A utility model has been formulated that uses three parameters U_{SLA}, U_r, and U_e to determine the best VM-to-PM mapping. U_{SLA} ensures that there are enough resources to support the SLA, U_r value shows the impacts of turning servers on and off and temperature variation on reliability and lifetime, and U_e estimates the amount of power usagereduction. Finally, the mapping that has the maximum value of the sum of these three parameters is chosen to provide an optimized solution of the problem.

There are also some works that considered energy efficiency and reliability in cloud computing at the same time that a review of them has been provided in [7]. But most of these works have not specifically addressed the issue of consolidation and focus on resource allocation in a reliable and energy efficient manner. However, our proposed approach provides a reliability model of PMs to use in consolidation process with the aim of saving unnecessary wastage of energy that will be required to restart all the running process that were interrupted during the failure.

3. System Model And Problem Definition

We consider a system consist of a single data center with heterogeneous resources as the scope of our work is restricted to migrations within a data center. Let PM = {pm$_1$, pm$_2$, …, pm$_i$,…, pm$_m$ } be the set of active PMs in the current state of the data center and VM$_i$={vm$_1$, vm$_2$, …, vm$_j$, … , vm$_n$} be the set of deployed VMs that in PM$_i$. Each PM is characterized by the CPU performance defined in Millions of Instructions Per Second (MIPS), amount of RAM, network bandwidth, and disk storage. But the disk storage space in any PM is usually large and dynamic variations in disk space requirements are usually not observed. Hence, it can be safely neglected. The other three resources CPU, Memory and Network Bandwidth are considered for the consolidation process.

At any given time, a cloud data center usually serves many simultaneous users. Users submit their requests for provisioning *n* heterogeneous VMs, which are allocated to the PMs and characterized by requirements of resources. The length of each request is specified in millions of instructions (MI). It is assumed that each of the *n* VMs is already placed in some PM in the data center. The problem is to minimize the number of PMs used, by maximizing the resource utilization in each PM using live migration of VMs so that the freed PMs can be set to a power saving state.

3.1 Reliability Model

Reliability is defined as an evaluation parameter to measure the system's ability to functioning correctly under certain conditions over a specified interval of time. In cloud computing system there are two general aspects of reliability in server consolidation approaches [4], service reliability, and hardware reliability. In this study, the second category is considered. We formulated the reliability of PMs by using the reliability of two important components of a PM, hardware, and hypervisor (VMM) that is explained in the next section and it can be expressed as [27]:

$$R_{PM_i}(t) = R_{HW_i}(t)R_{VMM_i}(t) = e^{-\left(\lambda_{HW_i}+\lambda_{VMM_i}\right)t} \qquad (1)$$

4. Proposed Approach

The proposed approach to consists of the following two main components.

- Prediction module: observes energy consumption caused by VMs and PMs and collect historical data of past failures that can be utilized in Markov chain based prediction model. The module is executed on each PM locally.
- Decision making unit: manages VM placement on PMs in the data center. According to the received PMs messages and states analysis, this unit determines each PM belongs to which category. Then, VM selection and target PM selection algorithms are carried out and appropriate decisions are made to solve the consolidation problem.

4.1 Markov Chain Based Prediction Model

Markov chain model is the most fundamental and general state-based stochastic method that concerns about a sequence of random variables, which correspond to the states of a system, in such a way that the state at one time epoch depends only on the one in the previous time epoch [28].

Markov chains are usually classified into two categories: Discrete Time Markov Chains (DTMC) and Continuous Time Markov Chains (CTMC). CTMC, semi-Markov process and Stochastic Petri Net (SPN) have been used widely for evaluating the performance [29], reliability/ availability [30], and performability [31] of computer systems. in this paper, we choose CTMC model to develop a prediction mechanism to analysis PMs reliability. Since, exponential random variable is the only continues random variable with Markov property and hardware and software fault are commonly modelled as exponential distribution, we assume that the time to transit from a system state to another due to failures and recovery follows an exponential distribution. Fig.1 shows the CTMC model state transition diagram for the probabilistic reliability behavior of each PM in data centre. Although in many works only two active and failed states are considered for a host, but there are some factors that result in performance degradation.

In this study we consider that hypervisor (VMM) is affected by software aging. One of the common ways to deal with this problem is software rejuvenation as a proactive fault management technique to prevent or postpone failures in VMMs and VMs. Migrate-VM rejuvenation [32] is an effective technique for VMM rejuvenation. In this technique, before triggering the VMM rejuvenation, running VMs are migrated to another host and then VMM rejuvenation starts. If we choose a PM with aged VMM as a VMs migration destination in consolidation process, leads to increase the number of migrations and waste energy. Therefore, in order to model the reliability of PMs, in addition to hardware failures, VMM failures are also considered which can be extended to other software failures.

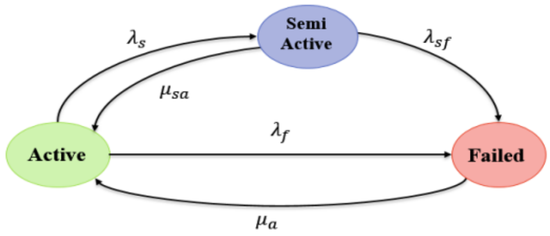


Fig. 1. State transition diagram

As depicted in fig.1, the model consists of three states including Active, Semi active and Failed. Let X(t) with discrete state space S = {A, SA, F} represents the state of PM at time t.

$$X(t) = \{X_A(t), X_{SA}(t), X_F(t)\} \tag{2}$$

if X(t)= X_A(t), PM is in proper condition and active state. In other two states that are inactive states we consider hypervisor failure and hardware failures. Hardware failures are critical and lead PMs to the failed state. Semi active state is about VMM rejuvenation process during which the PM is not available.

We define μ and λ as the repair rate and the failure rate of a PM, respectively. With this assumptions, the transient process X(t) can be modelled mathematically as a homogeneous CTMC on the state space S. for each time t>0, the probability of a PM in state *i* is given by X_i(t)=Pr{X(t) = *i*} , *i* ∈ S. The Markov process is defined by generator Q whose is given by:

$$Q = \begin{bmatrix} -\lambda_s - \lambda_f & \lambda_s & \lambda_f \\ \mu_{sa} & -\mu_{sa} - \lambda_{sf} & \lambda_{sf} \\ \mu_a & 0 & -\mu_a \end{bmatrix} \tag{3}$$

First Passage Time: Let τ_j be the expected value of random time to reach state *j* (for the first time), given that it started in state *i*. These are sometimes referred to as mean first-passage time. The first passage time into state N is defined to be

$$T = min\{n \geq 0 : X(t) = N\} \tag{4}$$

Where {1, …, N} represent the state space. the expected value E(T) is defined as

$$\tau_j = E\big((T|X_0 = i)\big) \tag{5}$$

According to a theorem defined in [33], the expected first passage time satisfy the following relation,

$$r_i\tau_i = 1 + \sum_{j=1}^{N-1} r_{i.j}\tau_j \qquad . \ 1 \leq i \leq N-1 \qquad 1 \tag{6}$$

Where $r_i = \sum_{j=1}^{N} r_{i.j}$ and $r_{i.j}$ is the entry of rate matrix R,

$$R = \begin{bmatrix} 0 & \lambda_s & \lambda_f \\ \mu_{sa} & 0 & \lambda_{sf} \\ \mu_a & 0 & 0 \end{bmatrix} \tag{7}$$

historical data and past failures can be utilized to estimate the λ and μ. Then, the generator matrix is constructed based on the estimated rates and the CTMC transition diagram. In the next step, transient state analysis performs to predict the state of PM and also compute the defined metrics values including expected time for the first occurrence of failure. The difference between the predicted and actual values can be used to train and modify the transition rates. Then, the obtained results are sent to decision making unit and used to classify the PMs for consolidation process.

4.2 VM Consolidation Process

After performing the prediction phase, PMs status is sent to the decision making unit. To determine whether the host is overloaded, we apply LR method proposed by beloglazov et al. [34]. this method utilizes local regression to fit a trend polynomial to the last *k* observations of the CPU utilization. In LR method for each new observation a new trend line is found. This trend line is used to estimate the next observation. Then the algorithm decides that the host is considered overloaded and some VMs should be migrated from it. Underloaded PMs can be found by comparing the CPU utilization with a low threshold. Other PMs are considered as well-utilized. According to the obtained results from previous steps, each PM will be in one of the six sets WR, OR, UR, WU, OU and UU. These sets represent the well-utilized and reliable, overloaded and reliable, under-loaded and reliable, well-utilized and unreliable, overloaded and unreliable, and under-loaded and unreliable PMs, respectively. Unreliable state is related to semi active and fail states in Markov model. Then these sets are divided into critical, optimal and sub optimal categories.

To select the migration source, the categories whose PMs are in critical situation are candidate. PMs in OU set have the highest priority. PMs of critical, optimal, and sub optimal categories are sorted based on MFPT in ascending order. The pseudocode of the PMs categorization algorithm is given as algorithm 1. At the end of this phase, the potential source PMs are determined. It should be noted that if all the PMs are in WR set, no migration will be done.

4.2.1 VM Selection and Placement

When finding a set of PMs in critical category, some VMs in the hosts are migrated to guarantee QoS for the users. Therefore, a VM selection policy is needed in the dynamic VM consolidation. Here, VMs are selected based on the Minimum Migration Time (MMT) policy [34]. MMT migrates a VM *v* that requires the minimum time to complete a migration relatively to the other VMs allocated to the PM. The migration time is estimated as the amount of RAM utilized by the VM divided by the spare network bandwidth available for the PM *j*.

After the VMs to be migrated are acquired, we need a policy to select appropriate target for migrations. When a PM is selected as the destination of VMs migration, its state likely change due to increasing workload and resource usage. Therefore, the proposed algorithm in this phase tries to find a proper host with sufficient residual capacity and considering energy consumption and reliability. In this way PMs in WR list of the optimal category is explored at first.

If the algorithm fails to find adequate PM, the search process continues in underutilized and reliable PM within the sub optimal category. In our proposed policy, a VM will be migrated to a PM with the highest score that is estimated according to Eq. (8). To specify each PM score, energy cost and reliability are considered. Then, scores are determined using weight assignment to each criterion. α is an adjustable weight to obtain different trade-off points, since each cloud provider will pursue various objectives and business requirements.

According to the proposed scoring method, VMs assigned to the PMs that their mean first passage time to an unreliable state is longer than the others to prevent additional migrations. The idea behind this is that such PMs will probably stay in reliable state for a longer period of time. Therefore, the migrated VMs can complete their works on the same server without any interruption or wasting time because of forced migration.

$$Score_i = \alpha \times \left(R_{PM_i} + MFP\_time_{PM_i}\right) + (1 - \alpha) \times \left(EC_i^{curr} - EC_i^{after}\right) \quad (8)$$

Where R_{pm} is the reliability of target server and $MFP\_time_{PM_i}$ is the expected value of random time to reach the unreliable state starting from the reliable state. EC_i^{after} and EC_i^{curr} are the energy cost after and before the VM_i placement, respectively. In order to allocate VMs to PMs, VM migration list will be sorted according to their CPU capacity requirements in decreasing order. Then the score of each PM in WR list is computed (see Algorithm 2).

After performing the aforementioned steps, we can safely shut down remaining underloaded PMs in UR and UU sets of suboptimal category. First we attempt to migrate all VMs on the PMs of UU set because of unreliability. If the proper destination PMs was found to hosting the migrated VMs, then source PM is switched to sleep mode. Finally, UR set is explored to reduce the number of active PMs as much as possible.

5. Performance Evaluation

This section describes our experimental results for the proposed approach. In this study we have chosen CloudSim toolkit [35] as the simulation platform that is a modern simulation framework for cloud computing environments. The experiments simulate a data center comprised of 800 heterogeneous PMs, half of which are HP ProLiant ML110 G4 (Intel Xeon 3040 2 Cores 1860 MHz, 4 GB) servers, and the other half are HP ProLiant ML110 G5 (Intel Xeon 3075 2 Cores 2260 MHz, 4 GB). VMs are supposed to correspond to Amazon EC2 instance types with the only exception that all the VMs are single-core, because of the fact that the workload data used for the simulations come from single-core VMs.

There are four types of VMs in the experiments: High-CPU Medium Instance, Extra Large Instance, Small Instance, and Micro Instance. After creating PM and VM instances on the CloudSim platform, the VMs are deployed to random PMs based on their resource requirements. After each round of VMs consolidation, VMs resource demands changes according to workload data. We assume HighTR and LowTR thresholds equal to 0.8 and 0.4, respectively. The parameter α is set to 0.5 in our experiments. To estimate the reliability of PMs, reliability of hardware is computed based on the decrease of mean time to failure(MTTF) models presented in [8] with considering CPU and disk reliability degradation.

In order to make the results of simulation more realistic, it is important to conduct experiments using workload traces from a real system. We have used data that provided as a part of the CoMon project, a monitoring infrastructure for Planet Lab [36]. In this project, the data on the CPU utilization is obtained every five minutes by more than a thousand virtual machines from servers located at more than 500 places around the world. We have chosen 10 different days from the workload traces gathered during March and April 2011, randomly.

In order to reasonably evaluate the efficiency of our proposed approach, we adopt several metrics that were presented by beloglazov et al. [34]. There are many metrics to measure the efficiency and superiority of various algorithms for VM consolidation problem. The main targets of VM consolidation in the cloud data center is to reduce energy consumption and SLA violations. So, we have chosen the related metrics to these objectives. One metric is the energy consumption consumed by the data center and the metrics used for quantifying SLA violations are based on the model provided in the CloudSim simulator (SLATH, PDM and SLAV).

The SLATAH is defined as Eq. (9), measures the percentage of time during which active hosts have experienced CPU utilization of 100%.

$$SLATAH = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i^s}{T_i^a} \tag{9}$$

Algorithm 1: Categorization

```
Input: PMs states
Output: Categories
1:  foreach host in PM set do
2:     if PM_state = OU │OR│WU add to Critical_cat
3:         elseif PM_state = WR add to Optimal_cat
4:         elseif PM_state= UR │UU add to SubOptimal_cat
5:     end if
6:  end for
7:  Sort lists in Critical_cat based on mfpt in ascending order.
8:  Sort list in Optimal_cat based on mfpt in ascending order.
9:  Sort lists SubOptimal_cat based on mfpt in ascending order.
10: return categories
```

Algorithm 2: Target PM selection Algorithm

```
Input: MigrationList, WR_list, UR_list
Output: MigrationSchedule
1: Sort Migration_list by resource requirements in descending order
2: foreach VM_j in MigrationList do
3:   best_score = Min
4:   target_PM = Null
5:    foreach PM_i in WR_list do
6:        Calculate Score_i using eq. (8)
7:        if Score_i  > best_score then
8:            best_score = Score_i
9:            target_PM = PM_i
10:       endif
11:    end for
12: if target_PM= Null then
13:    foreach PM_i in UR_list do
13:        repeat steps 7-10
14:       calculate new PM_i_load
15:       if PM_i_load> LowTR then
16:           Add PM_i to WR_list
17:       endif
18:    endfor
19:  endif
20:   MigrationSchedule.put(VM_j, target_PM)
21:  end for
22: return MigrationSchedule
```

Where n is the total number of physical machines, T_i^s is the total time of SLAV caused by the CPU resource overload of PM_i , T_i^a is the running time of PM_i. Another metric, PDM is calculated as follow:

$$PDM = \frac{1}{m}\sum_{j=1}^{m}\frac{C_j^d}{C_j^r} \tag{10}$$

Where m is the number of VMs, C_j^d is the unsatisfied CPU required capacity caused by the migration of VM_j, and C_j^r is the CPU capacity requested by VM_j. SLAV is a combined metric of two aforementioned metrics that evaluates a single-day QoS of the data center and is defined as:

$$SLAV = SLATAH \times PDM \tag{11}$$

Table 1. Simulation results of different algorithms

Method	EC(KWh)	SLAV	ESV(%)	VM Migrations
LR-MMT	160.21	0.49355	78.3394	32095
LR-MC	147.35	0.77112	111.2479	27350
LR-RS	146.01	0.78238	115.9127	26367
R-VMC	122.47	0.14171	19.6698	9457

ESV as described in EQ. (12), is a metric consist of energy consumption of a data center per day(EC) and the level of SLA violations. A lower estimation of ESV indicates that energy saving is higher than the SLA violations.

$$ESV = EC \times SLAV \tag{12}$$

5.1 Simulation Results

In this section the result of our experiments are discussed. Since we use LR method to host overload detection, three traditional combination methods, LR-MMT, LR-MC, and LR-RS [34], are selected to compare and evaluate our proposed approach. These methods apply PABFD algorithm [34] to target selection for migrated VMs. The safety parameter is set to 1.2 in experiments. The CTMC model parameter default values are found in the literature [27,32,37].

Comparison between other methods and our proposed algorithm (R-VMC) is shown in Table 1. The obtained results indicate that energy consumption is reduced by R-VMC algorithm compared to LR-MMT, LR-MC, and LR-RS, due to decreasing number of migrations and switching the underload and unreliable PMs to sleep mode which leads to energy saving. In terms of SLAV, R-VMC has optimal SLAV compared to others and LR-RS has the highest SLAV. According to the results, R-VMC's SLAV is only 18% of LR-RS's SLAV. These results reveal that R-VMC is better than other algorithms in guaranteeing QoS. The ESV index in Table 1 indicates that the comprehensive performance of R-VMC is considerably higher than others. The ESV of R-VMC is only 25% of LR-MMT which has the closest value to R-VMC, 17.6% of LR-MC and 16% of LR-RS. Eventually, the methods are compared in terms of the number of VM migration based on the experimental results. R-VMC has the lowest number of VM migrations because it avoids additional migration by selecting proper and reliable destination PMs. Fig. 2 shows the energy consumption of our proposed algorithm and the other algorithms. As can be seen, R-VMC is better than other algorithms in terms of energy consumption.

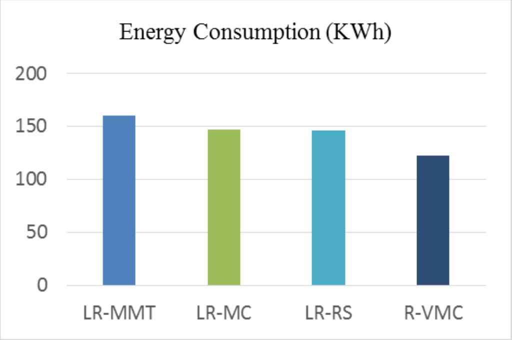


Fig. 2. The energy consumption of algorithms

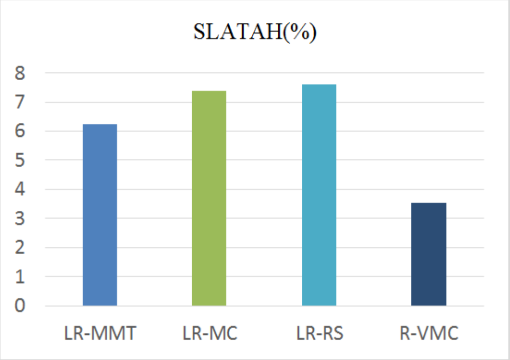


Fig. 3. Comparison of SLATAH

The reason is that R-VMC can avoid inefficient and extra migration due to selecting more reliable target PMs. So, minimizing the number of active physical servers and reducing the VM migrations leads to decreasing energy consumption.

Fig. 3 compares the results of SLATAH with other algorithms. It is completely obvious that R-VMC

outperforms the other methods and reduces PMs overload risk. The reason is that proposed approach can proactively migrate VMs from a host before the

Host become overloaded. On the other hand, R-VMC considers reliability which effectively leads to proper target PM selection. So, the QOS of running PMs is maintained.

Fig. 4 compares the R-VMC and the other algorithms in terms of PDM. As depicted in this figure, R-VMC has better performance. Prevention of extra migrations effects on this parameter directly and migrating VMs to the safer PMs with considering failures and VMM rejuvenation reduce the number of VM migration. Indeed, according to obtained results in experiments, we can conclude that one of our objectives about decreasing VM migration, has been achieved.

Fig. 5 shows the number of migrations of proposed algorithm and other algorithms. According to the results, the R-VMC has a smaller number of migrations and outperforms LR-MMT, LR-MC and LR-RS significantly.

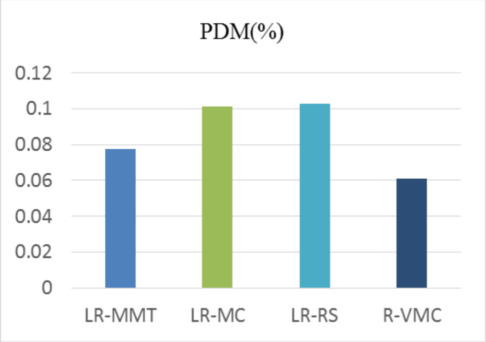


Fig. 4. Comparison of PDM

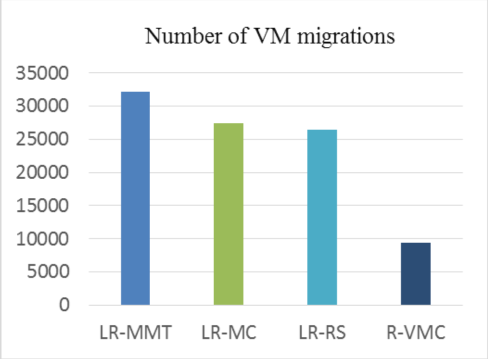


Fig. 5. Comparison of the number of VM migrations

The reason is that our proposed approach properly selects reliable servers as destination of migrations and prevents unnecessary migrations by avoiding unprofitable and aggressive reconfigurations. When we consider reliability, the probability of VMs migration, because of failure occurrence, is reduced. Therefore, while consuming a lower amount of energy, R-VMC has a fewer number of migrations.

6. Conclusion

In this paper, we have proposed a novel dynamic VM consolidation method in cloud data centers considering the reliability of each PM along with reducing the number of active PMs simultaneously. Most of the existing works on VM consolidation have been focused only on reducing the number of active PMs using VM live migration to prevent inefficient usage of resources. But on the other hand, high frequency of VM consolidation has the negative effect on the system reliability. Also, frequent turning on or off resources or putting them in sleep mode tends to make them more susceptible to failure and result in increasing the overall response time, service delays, and the SLA violation. Therefore, in this paper we tried to consolidate servers in flexible manner with considering both energy efficiency and reliability.

First, we have introduced a Markov model for reliability estimation of PMs. Then PMs are categorized based on the obtained results from the model and CPU overload detection algorithm (LR). Finally, we consider utilization along with the reliability in consolidation steps to select source and target PMs that leads to proper decision making and reduce migrations number, energy consumption, and consequently SLA violation. To evaluate the proposed VM consolidation method, CloudSim was chosen as the simulation platform and the simulation results have shown the effectiveness of R-VMC compared to other algorithms in terms of SLATAH, PDM, SLAV, EC, ESV and the number of VM migrations.

References

[1] M. Armbrust et al., "A view of cloud computing," Communications of the ACM, vol. 53, no. 4, pp. 50-58, 2010.

[2] M. Mills, "The cloud begins with coal. big data, big networks, big infrastructure, and big power. an overview of the electricity used by the digital ecosystem," ed, 2013.

[3] R. W. Ahmad, A. Gani, S. H. A. Hamid, M. Shiraz, A. Yousafzai, and F. Xia, "A survey on virtual machine migration and server consolidation frameworks for cloud data centers," Journal of Network and Computer Applications, vol. 52, pp. 11-25, 2015.

[4] A. Varasteh and M. Goudarzi, "Server consolidation techniques in virtualized data centers: A survey," IEEE Systems Journal, vol. 11, no. 2, pp. 772-783, 2017.

[5] A. Beloglazov and R. Buyya, "Managing overloaded hosts for dynamic consolidation of virtual machines in cloud data centers under quality of service constraints," IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 7, pp. 1366-1379, 2013.

[6] A. Beloglazov, "Energy-efficient management of virtual machines in data centers for cloud computing," 2013.

[7] Y. Sharma, B. Javadi, W. Si, and D. Sun, "Reliability and energy efficiency in cloud computing systems: Survey and taxonomy," Journal of Network and Computer Applications, vol. 74, pp. 66-85, 2016.

[8] W. Deng, F. Liu, H. Jin, X. Liao, and H. Liu, "Reliability-aware server consolidation for balancing energy- lifetime tradeoff in virtualized cloud datacenters," International Journal of Communication Systems, vol. 27, no. 4, pp. 623-642, 2014.

[9] A. Varasteh, F. Tashtarian, and M. Goudarzi, "On Reliability-Aware Server Consolidation in Cloud Datacenters," arXiv preprint arXiv:1709.00411, 2017.

[10] L. Grit, D. Irwin, A. Yumerefendi, and J. Chase, "Virtual machine hosting for networked clusters: Building the foundations for autonomic orchestration," in Proceedings of the 2nd International Workshop on Virtualization Technology in Distributed Computing, 2006, p. 7: IEEE Computer Society.

[11] B. Speitkamp and M. Bichler, "A mathematical programming approach for server consolidation problems in virtualized data centers," IEEE Transactions on services computing, vol. 3, no. 4, pp. 266-278, 2010.

[12] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," Future generation computer systems, vol. 28, no. 5, pp. 755-768, 2012.

[13] A. Beloglazov and R. Buyya, "Energy efficient resource management in virtualized cloud data centers," in Proceedings of the 2010 10th IEEE/ACM international conference on cluster, cloud and grid computing, 2010, pp. 826-831: IEEE Computer Society.

[14] S. Esfandiarpoor, A. Pahlavan, and M. Goudarzi, "Structure-aware online virtual machine consolidation for datacenter energy improvement in cloud computing," Computers & Electrical Engineering, vol. 42, pp. 74-89, 2015.

[15] S. Zhang, Z. Qian, Z. Luo, J. Wu, and S. Lu, "Burstiness-aware resource reservation for server consolidation in computing clouds," IEEE Transactions on Parallel and Distributed Systems, vol. 27, no. 4, pp. 964-977, 2016.

[16] H. Shen and L. Chen, "Compvm: A complementary vm allocation mechanism for cloud systems," IEEE/ACM Transactions on Networking (TON), vol. 26, no. 3, pp. 1348-1361, 2018.

[17] E. Arianyan, H. Taheri, and V. Khoshdel, "Novel fuzzy multi objective DVFS-aware consolidation heuristics for energy and SLA efficient resource management in cloud data centers," Journal of Network and Computer Applications, vol. 78, pp. 43-61, 2017.

[18] K. S. Rao and P. S. Thilagam, "Heuristics based server consolidation with residual resource defragmentation in cloud data centers," Future Generation Computer Systems, vol. 50, pp. 87-98, 2015.

[19] A. Ponraj, "Optimistic virtual machine placement in cloud data centers using queuing approach," Future Generation Computer Systems, vol. 93, pp. 338-344, 2019.

[20] Z. Li, C. Yan, X. Yu, and N. Yu, "Bayesian network-based virtual machines consolidation method," Future Generation Computer Systems, vol. 69, pp. 75-87, 2017.

[21] H. Khani, A. Latifi, N. Yazdani, and S. Mohammadi, "Distributed consolidation of virtual machines for power efficiency in heterogeneous cloud data centers," Computers & Electrical Engineering, vol. 47, pp. 173-185, 2015.

[22] T. Mahdhi and H. Mezni, "A prediction-Based VM consolidation approach in IaaS Cloud Data Centers," Journal of Systems and Software, vol. 146, pp. 263-285, 2018.

[23] F. Farahnakian et al., "Using ant colony system to consolidate VMs for green cloud computing," IEEE Transactions on Services Computing, vol. 8, no. 2, pp. 187-198, 2015.

[24] H. Mi, H. Wang, G. Yin, Y. Zhou, D. Shi, and L. Yuan, "Online self-reconfiguration with performance guarantee for energy-efficient large-scale cloud computing data centers," in Services Computing (SCC), 2010 IEEE International Conference on, 2010, pp. 514-521: IEEE.

[25] H. Li, G. Zhu, C. Cui, H. Tang, Y. Dou, and C. He, "Energy-efficient migration and consolidation algorithm of virtual machines in data centers for cloud computing," Computing, vol. 98, no. 3, pp. 303-317, 2016.

[26] H. Zhao, J. Wang, F. Liu, Q. Wang, W. Zhang, and Q. Zheng, "Power-aware and performance-guaranteed virtual machine placement in the cloud," IEEE Transactions on Parallel and Distributed Systems, vol. 29, no. 6, pp. 1385-1400, 2018.

[27] B. Wei, C. Lin, and X. Kong, "Dependability modeling and analysis for the virtual data center of cloud computing," in High Performance Computing and Communications (HPCC), 2011 IEEE 13th International Conference on, 2011, pp. 784-789: IEEE.

[28] N. B. Fuqua, "The applicability of markov analysis methods to reliability, maintainability, and safety," Selected Topic in Assurance Related Technologies (START), vol. 2, no. 10, pp. 1-8, 2003.

[29] K. S. Trivedi, Probability & statistics with reliability, queuing and computer science applications. John Wiley & Sons, 2008.

[30] A. Goyal, S. S. Lavenberg, and K. S. Trivedi, "Probabilistic modeling of computer system availability," Annals of Operations Research, vol. 8, no. 1, pp. 285-306, 1987.

[31] J. F. Meyer, "Closed-form solutions of performability," IEEE Transactions on Computers, no. 7, pp. 648-657, 1982.

[32] F. Machida, D. S. Kim, and K. S. Trivedi, "Modeling and analysis of software rejuvenation in a server virtualized system with live VM migration," Performance Evaluation, vol. 70, no. 3, pp. 212-230, 2013.

[33] J. K. Ghosh, "Introduction to Modeling and Analysis of Stochastic Systems, by VG Kulkarni," International Statistical Review, vol. 80, no. 3, pp. 487-487, 2012.

[34] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," Concurrency and Computation: Practice and Experience, vol. 24, no. 13, pp. 1397-1420, 2012.

[35] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. De Rose, and R. Buyya, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," Software: Practice and experience, vol. 41, no. 1, pp. 23-50, 2011.

[36] K. Park and V. S. Pai, "CoMon: a mostly-scalable monitoring system for PlanetLab," ACM SIGOPS Operating Systems Review, vol. 40, no. 1, pp. 65-74, 2006.

[37] R. d. S. Matos, P. R. Maciel, F. Machida, D. S. Kim, and K. S. Trivedi, "Sensitivity analysis of server virtualized system availability," IEEE Transactions on Reliability, vol. 61, no. 4, pp. 994-1006, 2012.

Monireh Hosseini Sayadnavard received the B.S. degree in computer engineering from Islamic Azad University, Lahijan, Iran, in 2005, the M.S. degree from Islamic Azad University, Qazvin, Iran, in 2010. She is now a Ph.D. candidate at Islamic Azad University (Tehran Science and Research branch). Her research interests include cloud computing, distributed systems and evolutionary computing.

Abolfazl Toroghi Haghighat received his B.S. and M.S. degrees in Electrical Engineering from Sharif University of Technology, Tehran, Iran, in 1993 and 1996, respectively. His Ph.D. degree from Amirkabir University of Technology (AUT), Tehran, Iran, in 2003. He is an Assistant Professor of Computer Engineering and Information Technology at Islamic Azad University of Qazvin. His research interests are in wireless networks, pattern recognition, fault-tolerant computing, and distributed systems.

Amir Masoud Rahmani received his B.S. in computer engineering from Amir Kabir University, Tehran, in 1996, the M.S. in computer engineering from Sharif University of technology, Tehran, in 1998 and the Ph.D. degree in computer engineering from IAU University, Tehran, in 2005. He is a professor in the Department of Computer Engineering at the IAU University. He is the author/co-author of more than 150 publications in technical journals and conferences. His research interests are in the areas of distributed systems, ad hoc and sensor wireless networks, scheduling algorithms and evolutionary computing.