

In the Name of God

Journal of

Information Systems & Telecommunication

Vol. 4, No. 2, April-June 2016, Serial Number 14

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

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

Managing Director: Habibollah Asghari, Assistant Professor, ACECR, Iran

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

Editorial Board

Dr. Abdolali Abdipour, Professor, Amirkabir University of Technology, Iran

Dr. Mahmoud Naghibzadeh, Professor, Ferdowsi University, Iran

Dr. Zabih Ghasemlooy, Professor, Northumbria University, UK

Dr. Mahmoud Moghavvemi, Professor, University of Malaya (UM), Malaysia

Dr. Ali Akbar Jalali, Iran University of Science and Technology, Iran

Dr. Alireza Montazemi, Professor, McMaster University, Canada

Dr. Ramezan Ali Sadeghzadeh, Professor, Khajeh Nasireddin Toosi University of Technology, Iran

Dr. Hamid Reza Sadegh Mohammadi, Associate Professor, ACECR, Iran

Dr. Ahmad Khademzadeh, Associate Professor, CyberSpace Research Institute (CSRI), Iran

Dr. Abbas Ali Lotfi, Associate Professor, ACECR, Iran

Dr. Sha'ban Elahi, Associate Professor, Tarbiat Modares University, Iran

Dr. Ali Mohammad-Djafari, Associate Professor, Le Centre National de la Recherche Scientifique (CNRS), France

Dr. Saeed Ghazi Maghrebi, Assistant Professor, ACECR, Iran

Dr. Rahim Saeidi, Assistant Professor, Aalto University, Finland

Administrative Manager: Shirin Gilaki

Executive Assistant: Behnoosh Karimi

Art Designer: Amir Azadi

Print ISSN: 2322-1437

Online ISSN: 2345-2773

Publication License: 91/13216

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

Email: info@jst.ir

URL: www.jst.ir

Indexed in:

- | | |
|---|-------------------------|
| - Index Copernicus International | www.indexcopernicus.com |
| - Journal of Information Systems and Telecommunication | www.jst.ir |
| - Islamic World Science Citation Center (ISC) | www.isc.gov.ir |
| - Scientific Information Database (SID) | www.sid.ir |
| - Regional Information Center for Science and Technology (RiCeST) | www.ricest.ac.ir |
| - Magiran | 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)

- Abdali Mohammadi Fardin, Razi University, Kermanshah, Iran
- Adibnia Fazlollah, Yazd University, Yazd, Iran
- Ashourian Mohsen, University of Isfahan, Isfahan, Iran
- Dolati Ardeshir, Shahed University, Tehran, Iran
- Ebrahimzadeh Ataollah, Babol Noshirvani University of Technology, Babol, Iran
- Eskandari Marzieh, Alzahra University, Tehran, Iran
- Ghaderi Reza, Shahid Beheshti University, Tehran, Iran
- Ghaffari Ali, Islamic Azad University, Tabriz Branch, Tabriz, Iran
- Ghaffari Valiollah, Persian Gulf University, Boushehr, Iran
- Ghatee Mehdi, Amirkabir University of Technology, Tehran, Iran
- Haji Bagher Naeeni Babak, IRIB University, Tehran, Iran
- Hatefi Seyed Morteza, University of Tehran, Tehran, Iran
- Heydarian Mohsen, Azarbijan Shahid Madani University, Tabriz, Iran
- Kazemipoor Hamed, Islamic Azad University, Parand Branch, Parand, Iran
- Keshavarz Hengameh, University of Sistan & Baluchestan, Zahedan, Iran
- Mahdieh Omid, University of Zanjan, Zanjan, Iran
- Mirroshandel Seyed Abolghasem, University of Guilan, Rasht, Iran
- Moallem Peyman, University of Isfahan, Isfahan, Iran
- Mohammadi Shahriar, Khaje Nasir-edin Toosi University of Technology, Tehran, Iran
- Mohammadzadeh Sajad, University of Birjand, Birjand, Iran
- Rasi Habib, Shiraz University of Technology, Shiraz, Iran
- Rezai Abdlhossein, Academic Center for Education Culture and Research (ACECR), Tehran, Iran
- Taher Atiyeh, Shiraz University, Shiraz, Iran
- Torabi Jahromi Amin, Nanyang Technological University, Singapore
- Vahdatnejad Hamed, University of Birjand, Birjand, Iran
- Yazdani Vahid, ICT faculty, Tehran, Iran

Table of Contents

- Privacy Preserving Big Data Mining: Association Rule Hiding 70
Golnar Assadat Afzali and Shahriar Mohammadi
- COGNISON: A Novel Dynamic Community Detection Algorithm in Social Network 78
Hamideh Sadat Cheraghchi and Ali Zakerolhosseini
- Analysis and Evaluation of Techniques for Myocardial Infarction Based on Genetic Algorithm
and Weight by SVM 85
Hodjatollah Hamidi and Atefeh Daraei
- Optimization of Random Phase Updating Technique for Effective Reduction in PAPR, Using
Discrete Cosine Transform 92
Babak Haji Bagher Naeeni
- Nonlinear State Estimation Using Hybrid Robust Cubature Kalman Filter 98
Behrouz Safarinejadian and Mohsen Taher
- Quality Assessment Based Coded Apertures for Defocus Deblurring 106
Mina Masoudifar and Hamid Reza Pourreza
- Design, Implementation and Evaluation of Multi-terminal Binary Decision Diagram based
Binary Fuzzy Relations 117
Hamid Alavi Toussi and Bahram Sadeghi Bigham
- Unsupervised Segmentation of Retinal Blood Vessels Using the Human Visual System Line
Detection Model 125
Mohsen Zardadi, Nasser Mehrshad and Seyyed Mohammad Razavi

Privacy Preserving Big Data Mining: Association Rule Hiding

Golnar Assadat Afzali*

Department of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran
g.afzali@gmail.com

Shahriar Mohammadi

Department of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran
mohammadi@kntu.ac.ir

Received: 05/Apr/2016

Revised: 16/May/2016

Accepted: 06/Jun/2016

Abstract

Data repositories contain sensitive information which must be protected from unauthorized access. Existing data mining techniques can be considered as a privacy threat to sensitive data. Association rule mining is one of the utmost data mining techniques which tries to cover relationships between seemingly unrelated data in a data base.. Association rule hiding is a research area in privacy preserving data mining (PPDM) which addresses a solution for hiding sensitive rules within the data problem. Many researches have been done in this area, but most of them focus on reducing undesired side effect of deleting sensitive association rules in static databases. However, in the age of big data, we confront with dynamic data bases with new data entrance at any time. So, most of existing techniques would not be practical and must be updated in order to be appropriate for these huge volume data bases. In this paper, data anonymization technique is used for association rule hiding, while parallelization and scalability features are also embedded in the proposed model, in order to speed up big data mining process. In this way, instead of removing some instances of an existing important association rule, generalization is used to anonymize items in appropriate level. So, if necessary, we can update important association rules based on the new data entrances. We have conducted some experiments using three datasets in order to evaluate performance of the proposed model in comparison with Max-Min2 and HSCRIL. Experimental results show that the information loss of the proposed model is less than existing researches in this area and this model can be executed in a parallel manner for less execution time

Keywords: Big Data; Association Rule; Privacy Preserving; Anonymization; Data Mining.

1. Introduction

Data mining is the process of extracting hidden but useful knowledge from large data bases [1]. Nowadays different sources are creating data with high speed [2]. Distributed infrastructures such as cloud computing present the opportunity to store large volume data bases for further analysis and knowledge discovery.

Big data mining is the capability of extracting desired information from large data bases or data streams [3]. Association rule mining is one of the most important data mining techniques. However, misuse of this technique may lead to disclosure of sensitive information about users [4,5]. Many algorithms have been proposed in the literature for rule hiding [6,7,8,9,10]. Most of them are based on the idea of modifying main data base to decrease the support or confidence value of sensitive association rules. The main drawback of existing works is the undesired side effect of removing some item-set on non-sensitive association rules.

In this paper, we use anonymization techniques as an alternative for removing some repeated instance of frequent item-sets. The main idea of the proposed model is that removing frequent item-sets (which is used in existing related works) has undesired side effect on new entrance data. But, by using data anonymization any necessary change can be applied to existing

anonymization level to support this new data. In other word, it is possible to change (increase or decrease) the anonymization level by new data entrance. As in big data mining, we deal with dynamic datasets, with any new data entrance, association rules can change. So, the proposed model we replace removing some instance of association rules with rule anonymity.

The remainder of this paper is organized as follow: next section reviews the related works. In section III, the proposed approach for big data association rule hiding is described. Experimental results of comparing the performance of our approach with previous works are described in section IV. At last, section V concludes this paper.

1.1 Related Work

A. Big Data

In term of definition, big data refers to high volume of structured, semi-structured and unstructured data with high velocity which can be mined for information [3]. Big data mining refers to the capability of extracting information from massive datasets that due to specific features cannot be done using existing data mining techniques [1].

In many situations, it is infeasible to store this huge amount of data, so the knowledge extraction should be

* Corresponding Author

done real-time. For processing big data, a cluster of computers with high computing performance is needed and this framework would be practical with paralleling tools such as MapReduce [11].

B. Anonymity

Information dissemination is usually with the risk of sensitive information disclosure [12]. Data usually contain sensitive information and this proves the importance of employing anonymity approaches [12,13]. Generally, there are three techniques for anonymization which include generalization, suppression and randomization. Different approaches for anonymization such as k-anonymity, l-diversity, t-closeness and etc. use these techniques.

In generalization, values of attributes are replaced with a more general one [14]. For example, if the value of attribute "age" is equal to 16, it can be replaced with appropriate range such as (10-20).

Suppression refers to stop releasing the real value of an attribute. In this way, occurrence of the value is replaced with a notation such as "*", this means that any value can be replaced instead [15]. For example, if the related value of an attribute is equal to 56497, it can be replaced with 5649*.

Randomization refers to substitution of real value with a random value. In this technique, noise is added to data so that real value of attributes is masked [16].

In this paper proposed model generalization technique is used for anonymity, while suppression technique is not suitable for quantitative data because in quantitative data we cannot substitute some parts of the data with "*". A secure randomization technique needs a defined and not reversible function. Therefore for each data, related assigned noise should be saved to make it possible to retrieve the real value, if necessary. Therefore, this technique imposes significant overhead to systems.

C. Association Rule Hiding

Association rule mining is an interesting approach to find out unknown relations between variables in large databases [6]. However, misuse of these techniques may cause disclosure of sensitive information [17]. So, many researchers worked on hiding sensitive association rules. The main purpose of association rule hiding approaches is to hide sensitive rules, without any side effect on non-sensitive rules. Le et al. [8] proposed HSCRIL model as a heuristic approach to hide a set of association rules from relational databases in retail industry. The main steps of their proposed algorithm are: identification of victim items that their modifications have least impact on other frequent item-sets, determination of minimum number of transactions which should be modified, and removing victim items from specified transactions. In this research, generation set of frequent item-sets is maintained. This generation set causes least impact on non-sensitive item-sets during sensitive rule hiding. The main result of this model is an acceptable information loss. But, this model is based on determined generation sets; however, in big

data mining, with new data entrance, generation sets will change. So, this idea cannot be applicable for big data.

Max-Min2 model of Moustakides and Verykios [18], used Max-Min theorem in association rule hiding. The main idea of this theorem is to maximize the minimum gain. In fact, they are trying to maximize sensitive rule hiding while at the same time minimize the side effect on non-sensitive rules. This model hides sensitive association rules by decreasing the support of sensitive item-sets. Results of this research show that the information loss of this model is less than existing related works. This model tried to hide sensitive association rules by removing some instances of them. However, in dynamic data sets, sensitive association rules will change with new data entrances and removing them cannot be a good idea.

Wang et al. in [19] proposed a model in which two algorithms are used to hide sensitive association rules. They used ISL (Increase support of left hand side) and DSR (decrease support of right hand side) to achieve their purpose. Removing sensitive association rules by these two algorithms causes mentioned problems of Max-Min2 algorithm.

In the research of [20] by Wang et al., all existence transactions are represented in the form of a binary matrix. In this matrix, if item i participates in transaction j , D_{ij} will be 1, otherwise it is equal to 0. Then, based on the defined threshold of support value in this system, matrix S would be determined so that $D' = S * D$. In this definition, D is the matrix related to the main database, S is the hiding matrix, and D' is the matrix related to hidden database. Based on the "volume" feature of big data, defining related matrix is time consuming and needs high storage capacity.

Dasseni et al. [21] considered the hiding of both sensitive association rules and frequent item-sets. They develop three strategies for this purpose: Increasing the support of left hand side (LHS), decreasing the support of right hand side (RHS), and decreasing the support of right and left hand sides, simultaneously. In this paper, three strategies cause less undesired effect on non-sensitive association rules. However, main disadvantage of this model is similar to Max-Min2.

Jung et al. [22] use Hadoop for association rule hiding in large scale datasets. Privacy threats which are considered in this paper are related to the flow of data to untrusted cloud service providers. So, at the first step, association rules are determined. Then, some noises are added to the item-sets to prevent frequent item-set disclosure of them. This model can prevent exposure of sensitive data without data utility degradation, but adding noise to data causes endures computing cost to systems and is not suitable for big data mining and real time processing.

Xu et al [23] concentrate on information security on big data analysis. They identify four types of users involved in data mining application. Namely, data provider, data collector, data miner and decision maker. For each group, security threats are defined and appropriate solutions considered, too.

In this paper's proposed model, anonymization technique is used to hide sensitive information. So, at first,

two criteria are defined to select best item-set(s) for anonymization. Then, based on these two criteria, in any sensitive association rule, the item-set with the least undesired side effect is selected in order to be hidden.

For selected item-set(s), quasi-identifier attributes are anonymized in appropriate level. In other word, in this proposed model, none of repeated item-sets would be removed from database and only sensitive values would

be hidden. So, if with new data entrance, each association rule changes between “sensitive” and “non-sensitive” mode, only by changing anonymity level, main purpose which is retrieving related information or hiding information, would be acquired.

Table 1 summarizes these related works.

Table 1. summary of related work

Author(s)	Method	Information hiding	Association rule hiding
[8] Le et al.	- Identification of informative items - Specification of generation set related to frequent item-sets. - Removing determined item-set form database	No	Yes
[18] Moustakides & Verykios	- Using Max-Min theorem to maximum information hiding with minimum side effect. - Reducing support of frequent item-sets to less than defined threshold.	No	Yes
[19] Wang et al.	- Decreasing support value of frequent item-sets to less than defined threshold. - Decreasing confidence value of frequent item-sets to less than defined threshold. - Utilizing ISL (Increase Support of Left hand side) and DSR (Decrease Support of Right hand side) functions to achieve mentioned purposes.	No	Yes
[20] Wang et al.	- Binary indicator matrix of items in transactions, named as D. - Hiding matrix S is determined based on the defined threshold for support. - Matrix D' related to hidden data set, is determined based on S and D	No	Yes
[21] Dasseni et al.	- Sensitive information hiding besides association rule hiding. - Increasing the support value of LHS (Left Hand Side). - Decreasing the support value of RHS (Right Hand Side). - Decreasing the support of RHS (Right Hand Side) and LHS (Left Hand Side), simultaneously.	Yes	Yes
[22] Jung et al.	- Determine sensitive association rules. - Adding noise to sensitive association rule to hide them from undefined user, without significant information loss.	No	Yes
Proposed model	- Sensitive information hiding besides association rule hiding. - Using anonymity approach for hiding sensitive association rules.	Yes	Yes

As mentioned below, as in the proposed model, anonymity technique is used instead of removing instances of association rules, if with any new data entrance, each association rule changes from sensitive to non-sensitive (or vice versa), we can update dataset easily. This feature makes the proposed model appropriate for dynamic datasets. While in all of existing models, authors only has concentrated on static datasets.

2. Proposed Model for Big Data Association Rule Hiding

As mentioned, association rules should not be disclosed since they may be used to infer sensitive information. Many researches have done in association rule hiding which most of them have significant drawbacks:

- Undesired side effect of hiding sensitive association rules on non-sensitive rules.

- The impossibility of using in big data analysis.

To solve these mentioned problems, anonymity techniques could be used for rule hiding as an alternative for deleting some of the most repeated items. In this paper two criteria are defined in order to support new data entrance in our big data base which are represented below. It is notable that features such as parallelization and scalability which considered in this model make it suitable for big data analysis.

The proposed model consists of three main steps which are described in follow:

Step 1: Association Rule Mining

There are many association rule mining algorithms such as Apriori [24] or FP-growth [25]. Let $\alpha(H)$ be the support value of item-set H, this item-set is called frequent item-set if $\alpha(H) > \sigma$, which σ is the defined support threshold. An association rule $A \rightarrow B$ is considered as a sensitive association rule if $\alpha(A \rightarrow B) \geq \sigma$ and

$\beta(A \rightarrow B) \geq \delta$, which $\beta(X)$ refers to the confidence value of this rule and δ is the defined confidence threshold.

Step 2: Best Item-set Selection

Because of the velocity feature of big data, selection of the best item(s) for anonymization should be done based on the two criteria:

- Undesired side effect of anonymization on other existing non-sensitive association rules.
- Undesired side effect of anonymization on probable new entrance data.

The Best approach is to decrease these values as much as possible.

Suppose that we want to hide a rule such as $A \rightarrow B$. The main problem is to determine the best item-set for anonymization. For this, anonymization effect of each right or left hand side item should be evaluated based on the two mentioned criteria and then, the item with the least side effect is selected.

At first, Association rules are sorted based on their confidence value. Then, these factors are used for the best item selection.

The First criterion has a static view on data set (without new data entrance). So, information loss which is caused by this anonymization could be computed with formula presented in (1).

$$\text{InfoLoss} = \frac{N_i}{N_i + N_j} \quad (1)$$

In formula (1), N_i is the number of non-sensitive association rules which A is involved in, while N_j is the number of sensitive association rules which A is involved in.

In the second criterion, we have dynamic view on our data set. In this manner, the best item is one which has greater chance to convert related non-sensitive association rules to sensitive association rule. This can cause lower information loss. So, the difference between defined confidence threshold and confidence value of existing non-sensitive association rules can be considered as the second criterion for the best item selection. This measure can be evaluated based on the formula presented in (2).

$$\text{DoC} = \sqrt{\sum_{i=1}^n (C_i - CL)^2} \quad (2)$$

In (2), C_i is the confidence value of i 'th non-sensitive association rule which A is involved in, while CL is the defined confidence threshold.

Finally, the best item selection could be done by combining InfoLoss and DoC values, but with appropriate effective weight, as (3);

$$\text{BI} = \alpha_1 * \text{InfoLoss} + \alpha_2 * \text{DoC} \quad (3)$$

The item with less BI value could be selected as the best item for anonymization. In (3), α_1 and α_2 are effective weights and their values can be changed based on the importance ratio of related criterions in each specific context. By default, α_1 and α_2 have same value and are equal to 0.5.

Step 3: Data Anonymization

As mentioned, generalization technique is used as the proposed anonymization technique. Attributes of each item-set can be classified in three categories: identifier attributes are attributes containing identifying information such as Social Security Number (SSN); sensitive attributes are set of attributes that contain personal privacy information and should be protected; quasi-identifier (QI) attributes are attributes that do not contain identifying attributes, but can be linked to other information to cause identification disclosure [8].

So, in this model, after selecting the best item-set for anonymization, exact value of sensitive and identifier attributes would be removed and then quasi-identifier attributes of this item-set would be generalized to an acceptable level.

The pseudo code of the proposed model is shown in figure 1.

```

Initialize list of sensitive association rules
While sensitive association rules lists is not empty
{
Sort sensitive association rules in decreasing
order of confidence value
Select the association rule with the maximum
confidence value
While selected association rule is not anonymized
yet
{
Calculate InfoLoss and DoC for each item-set
of association rule
Calculate the BI for each item-set
Choose the item-set with maximum BI value
Anonymize selected item-set
Remove this association rule from sensitive
association rules
}
}

```

3. Parallelization of the Proposed Method for Big Data Mining

As said before, in order to facilitate the implementation of the proposed model for big data processing, features such as parallelism should be considered in this model. Distributed computing infrastructures, such as cloud computing, can provide the required infrastructure for this purpose. Now, it is required that besides considering tree structure for our database, as shown in figure 2, basic operations such as association rule mining to be done in a distributed and parallel manner.

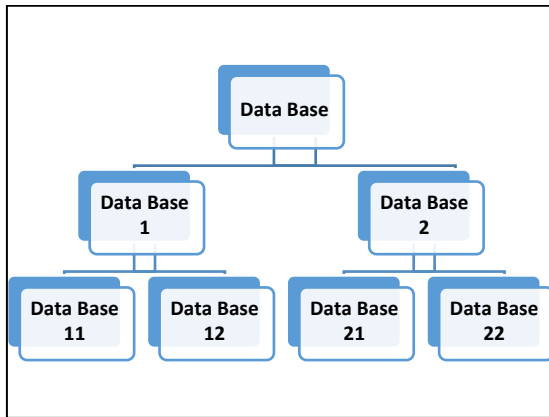


Fig. 2. proposed hierarchical structure for big data base

In addition to defining a tree structured data set, proper changes should be considered in the defined threshold of support and confidence values. Ideally, data set generator of each association rule is divided equally to slave nodes (nodes which are responsible for data storing and running the computations). In this manner, defined threshold of support and confidence values in each node is changed based on formula presented in (4).

$$\text{Threshold}_{\text{new}} = \text{Threshold}_{\text{old}} * \frac{1}{n} \quad (4)$$

In (4), $\text{threshold}_{\text{old}}$ is the defined threshold for support and confidence parameters, n is the number of slave nodes (leaf nodes in hierarchical tree structure), and $\text{threshold}_{\text{new}}$ is the new defined threshold in each data node.

Normally, it is possible that the distribution of the main data set on data nodes would not be according to the mentioned ideal form. In this manner, if in any slave node, the computed support and confidence values of each association rule is higher than the new threshold, this rule may be a sensitive association rule. So, existence of this rule in other slave nodes should be checked and according to this information, this rule would be defined as a sensitive or non-sensitive association rule.

Appropriate tree structure for data set (as shown in figure 2) can facilitate scalability feature, too. In this structure, every node can extend the number of its children. Therefore, the number of slave nodes can be extended until required computing power reached.

4. Evaluation

In order to evaluate the proposed model performance, some experiments have been done and the results are compared with Max-Min2. Max-Min2 algorithm has gained better results in minimizing undesired side effect compared with other existing association rule hiding approaches.

A. Dataset Description

Experiments have been done using three datasets. First dataset named Brijs dataset, contains market basket data from a Belgian retail supermarket store. Dataset contains 88162 transactions and 16469 product IDs.

Other two datasets are BMS-WebView-1 and BMS-WebView-2. These datasets are well-known datasets in association rule mining and contain click-stream data which are collected from two e-commerce web sites. The main goal of these two data sets are to determine the association between products which are viewed by visitors.

In order to increase the volume of datasets and make the dataset suitable for big data analysis, some instances of transactions are sampled randomly and repeated. Each database is divided into six partitions. Size of the first partition is equal to 500K and other partitions are added in next phases to this database in order to simulate the data stream feature of big data.

B. Experiment Process

As mentioned above, the proposed model is compared with Max-Min2 and HSCRIL algorithms. Three metrics which are used for this comparison are: percentage of lost rules, ghost rules, and false rules, where

Lost rule: a non-sensitive association rule which are lost during association rule hiding process and are not in the released database [8].

Ghost rule: a non-sensitive association rule which cannot be mined from main database but can be mined from released database [8].

False rule: a sensitive association rule which cannot be hidden using the proposed association rule hiding process [8]. Figures 3,4 and 5 compare the performance of the Max-Min2, HSCRIL and the proposed model.

In these figures, part a, shows the lost rules of these models in each dataset, part b, shows the ghost rules of the these models and part c, is related to the false rules which are produced by them. As shown in figure 3.a, at first, number of lost rules in the proposed model is higher than Max-Min2 and HSCRIL models; but it starts to work better as new data arrives. The main reason is that these models have static view on database. For example, consider at time t_1 , rule $A \rightarrow B$ is considered as a sensitive association rule and the appropriate item-set is removed from some transactions in database. Now, if with the entrance of new data, the confidence value of this rule in the main database is decreased to be less than defined confidence threshold, this rule is a non-sensitive rule and should not be hidden. However, there is not the chance to retrieve this removed rule.

It should be noticed that another approach is to check the main database (which is not hidden) in order to retrieve such non-sensitive hidden rule. It is clear that because of the huge volume of data in big data mining, this approach is very time consuming and would be an impractical way.

Number of ghost rules produced by the proposed model is less than MaxMin2, in all of datasets.

Any of these models would not produce false rules. So, the percentage of false rule for all of them is equal to zero.

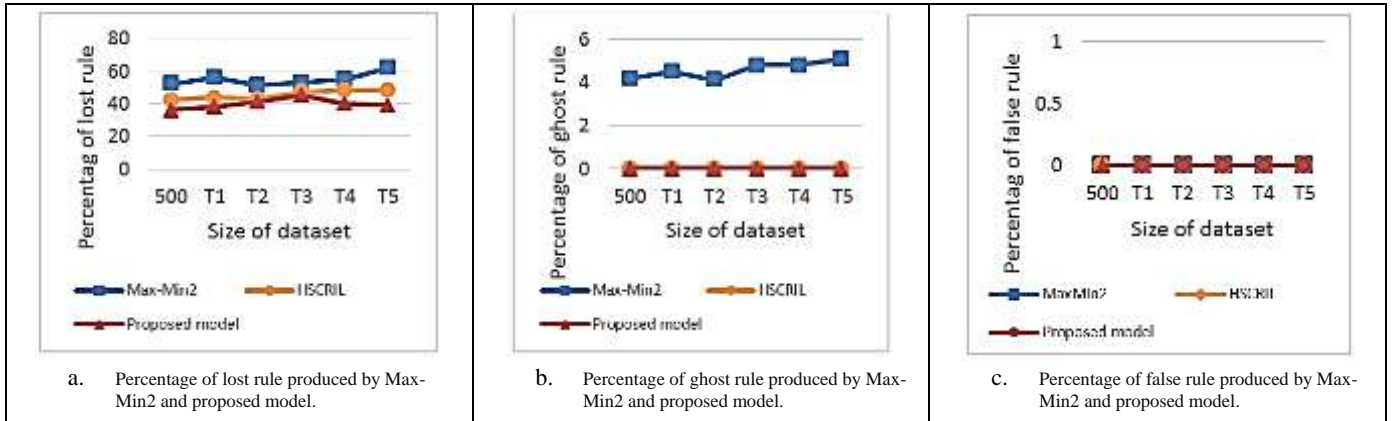


Fig. 3. comparison of the proposed model, HSCRIL and MaxMin2, Brijis dataset.

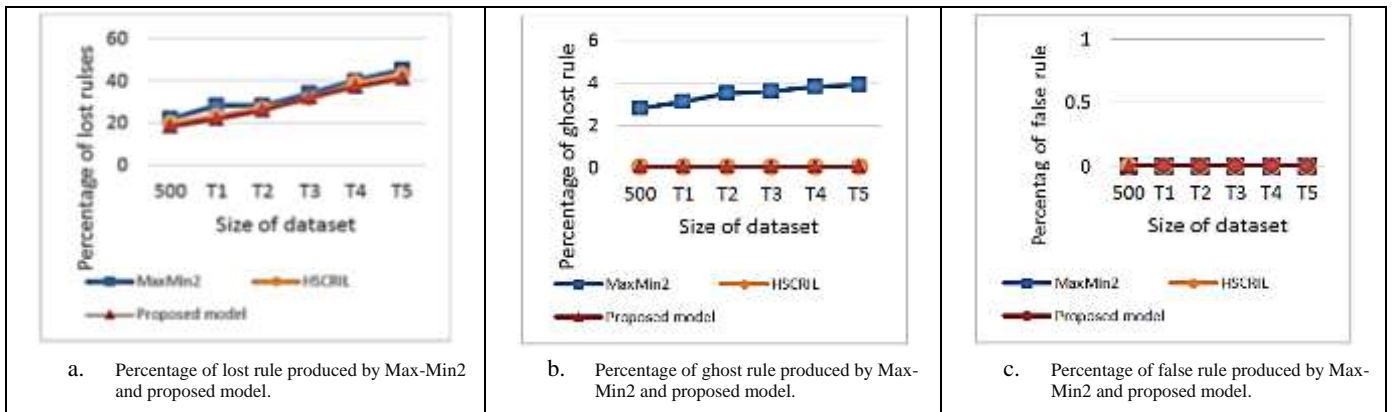


Fig. 4. comparison of the proposed model, HSCRIL and MaxMin2, BMS-WebView-1 dataset

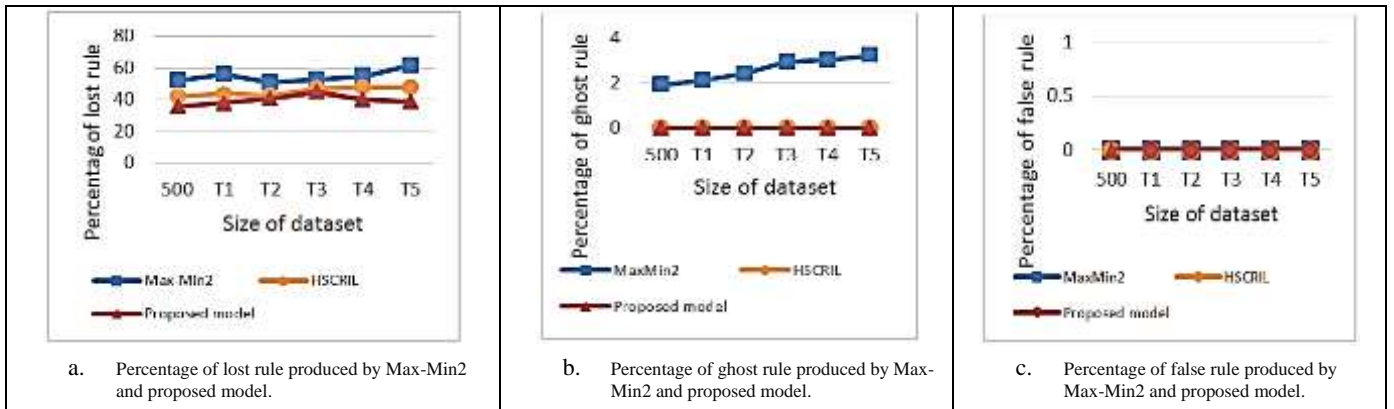


Fig. 5. comparison of the proposed model, HSCRIL and MaxMin2, BMS-WebView-2 dataset.

Percentage of released lost rules, ghost rules and false rules in Brijis, BMS-WebView-1 and BMS-WebView-2 datasets are mentioned in table 2,3 and 4.

In order to evaluate the scalability and parallel processing capability features of the proposed model, multi thread processing is used to simulate the distributed computing infrastructure. Number of thread has been changed from 4 to 10. At each manner, defined threshold of the support and confidence values changed based on the number of threads and formula 4. Execution time of the proposed model is shown in figure 6. As shown in

figure 6, as the number of the threads increases, execution time of the proposed model will decrease.

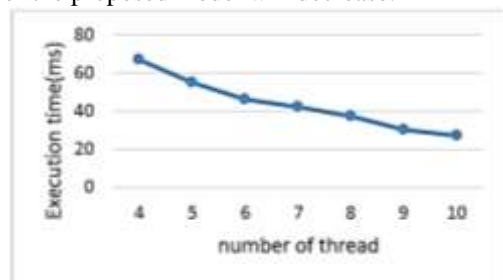


Fig. 6. execution time vs number of threads.

Table 2. Percentage of lost rule

	Brijs dataset						BMS-WebView-1 dataset						BMS-WebView-2 dataset					
Max-Min2	40	39	42	45	40	43	22	28	28	34	40	45	52	56	51	53	55	62
HSCRIL	33	34	37	31	29	31	20	23	27	39	39	43	42	44	43	47	48	48
Proposed model	30	26	28	25	22	29	18	22	26	32	37	41	36	38	41	45	40	39

Table 3. Percentage of ghost rule

	Brijs dataset						BMS-WebView-1 dataset						BMS-WebView-2 dataset					
Max-Min2	4.2	4.5	4.1	4.8	4.8	5.1	2.8	3.1	3.5	3.6	3.8	3.9	1.9	2.1	2.4	2.9	3	3.2
HSCRIL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Proposed model	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4. Percentage of false rule

	Brijs dataset						BMS-WebView-1 dataset						BMS-WebView-2 dataset					
Max-Min2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HSCRIL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Proposed model	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

5. Conclusion

Association rule mining is a data mining technique which besides its benefits in discovering unclear relationships between data, will result privacy violation. Association rule hiding can help to protect sensitive association rules to be discovered. Many different techniques have been considered to hide sensitive association rules but most of them try to select item-sets and remove them, in order to decrease the confidence value of the related rule(s) to be less than the defined threshold. In this model, instead of removing some instances of the frequent item-sets, item-sets are assigned to appropriate anonymity level. None of existing approaches can be executed in a parallel and scalable manner, to be appropriate for big data mining. Besides, removing item-

sets from the database can cause serious information loss as new data stream arrives. In this research, new big data association rule hiding technique is presented which tries to decrease undesired side effect of sensitive rule hiding on non-sensitive rules in data streams. Features such as parallelism and scalability are embedded in the proposed model to provide the facility of implementing this model for huge volume of data. Empirical evaluations show that the proposed model have less number of lost rules and ghost rules in data stream. Therefore, the performance of this model is better than other existing researches and embedded features such as parallelism and scalability can make it suitable for big data mining. So, it can be concluded that the proposed model is more effective in big data mining than existing rule hiding approaches.

As future work, we will try to decrease undesired side effect of the proposed model to gain less information loss.

References

- [1] C.L.P.Chen and Ch.Zhang. "Data Intensive Applications, Challenges, Techniques, and Technologies: A Survey on Big Data". Information Science, Vol.275, pp.314-347, 2014.
- [2] O.Kwon. N.Lee. and B.Shin. "Data Quality Management, Data Usage Experience and Acquisition Intention of Big Data Analytics", International Journal of Information Management, Vol.34, No.3, pp 387-394, 2014.
- [3] A.Cuzzocrea. C.K.S. Leung and R.K.Mackinnon. "Mining Constrained Frequent Item-Sets from Distributed Uncertain Data", Future Generation Computer Systems, Vol.37, pp 117-126. 2014.
- [4] X.Zhang. Ch.Liu. S.Nepal. Ch.Yang. W.Dou. and J.Chen. "A Hybrid Approach for Scalable Sub-Tree Anonymization over Big Data using MapReduce on Cloud", Journal of Computer and System Science, Vol.80, No.5, pp 1008-1020, 2014.
- [5] Y.Li. M.Chen. Q.Li. and W.Zhen. "Enabling Multilevel Trust in Privacy Preserving Data Mining", IEEE Transaction on Knowledge and Data Engineering, Vol.24, No.9, pp 1589-1612, 2012.
- [6] Y.H.Wu. C.Chiang and A.L.P.Chen. "Hiding Sensitive Association Rules with Limited Side Effects", IEEE Transaction on Knowledge and Data Engineering, Vol.19, No.1, pp 29-42, 2007.
- [7] A.Gkoulalas-Divanis and V.S.Verykios. "Exact Knowledge Hiding through Database Extension", IEEE Transaction on Knowledge and Data Engineering, Vol.21, No.5, pp 699-713, 2009.
- [8] H.Q.Le. S.Arch-int. H.X.Nguyen and N.Arch-int. "Association Rule Hiding in Risk Management for Retail Supply Chain Collaboration", Computer in Industry, Vol.64, No.4, pp776-784, 2013.

- [9] Y.Ch.Li. J.S.Yeh. and Ch.Chang. "MCIF: An Effective Sanitization Algorithm for Hiding Sensitive Patterns on Data Mining", *Advanced Engineering Informatics*, Vol.21, No.3, pp 269-280, 2007.
- [10] B.N.Keshavamurthy. D.Toshniwal. and B.K.Eshwar. "Hiding Co-Occurring Prioritized Sensitive Patterns over Distributed Progressive Sequential Data Streams", *Journal of Network and Computer Applications*, Vol.35, No.3, pp1116-1129, 2012.
- [11] X.Wu. X.Zhu. G.Wu. and W.Ding. "Data Mining with Big Data", *IEEE Transaction on Knowledge and Data Engineering*, Vol.26, No.1, pp 97-107, 2013.
- [12] M.E.Nergiz. and M.Z.Gok. "Hybrid K-Anonymity", *Computers & Security*, Vol.44, pp 51-63, 2014.
- [13] B.Li. E.Erdin. M.H.Gunes. G.Bebis. T.Shipley. "An Overview of Anonymity Technology Usage", *Computer Communication*, Vol.36, No.12, pp 1269-1283, 2013.
- [14] A.Monreale. G.Andrienko. N.Andrienko. F.Giannotti. D.Pedreschi. S.Rinzivillo. and S.Wrobel. "Movement Data Anonymity through Generalization", *Transactions on Data Privacy*, Vol.3, No.2, 2010.
- [15] S.Kisilevich. L.Rokach. Y.Elovici. and B.Shapira. "Efficient Multidimensional Suppression for K-Anonymity", *IEEE Transaction on Knowledge and Data Engineering*, Vol.22, No.3, pp 334-347, 2010.
- [16] G.Zhang. Y.Yang. X.Liu. and J.Chen. "A Time-Series Pattern Based Noise Generation Strategy for Privacy Protection in Cloud Computing, International Symposium on Cluster, Cloud and Grid Computing (CCGrid), pp 458-465, 2010.
- [17] H.Wang. "Quality Measurement for Association Rule Hiding", *AASRI Procedia*, Vol.5, pp 228-234, 2013.
- [18] G.V.Moustakides. and V.S.Verykios. "A MaxMin Approach for Hiding Frequent Item Sets", *Data & Knowledge Engineering*, Vol.65, No.1, pp 75-89, 2008.
- [19] S.Wang. B.Parikh. and A.Jafari. "Hiding Informative Association Rule Sets", *Expert Systems and Applications*, Vol.33, No.2, pp 316-323, 2007.
- [20] Ch.Wang. S.Tseng. and T.Hongm. "Flexible Online Association Rule Mining Based on Multidimensional Pattern Relations", *Information Science*, Vol.167, No.12, pp 1752-1780, 2006.
- [21] E.Dasseni. V.S.Verykios. A.K.Elmagarmid. and E.Bertino. "Hiding Association Rules by Using Confidence and Support", *Information Hiding Lecture Notes in Computer Science*, Vol.2137, pp 369-383, 2001.
- [22] K.Jung. S.Park. S.Cho. and S.Park. "A Novel Privacy Preserving Association Rule Mining using Hadoop", *The Third International Conference on Data Analytics*, 2014, pp 131-137.
- [23] L.Xu. C.Jiang. J.Wang. J.Yuan. and Y.Ren. "Information Security in Big Data: Privacy and Data Mining". *IEEE Access*, vol.2, pp. 1149-1176, 2014.
- [24] Ch.Borgelt. and R.Kruse. "Introduction of Association Rules: Apriori Implementation", *Compsat, Physica-Verlog Heidelberg*, pp 395-400.
- [25] Ch.Borgelt. "An Implementation of the FP-Growth Algorithm", *Proceeding of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, 2005, pp 1-5.

Golnar Assadat Afzali is a M.Sc graduate of Information Technology Engineering at K.N.Toosi University of Technology. She received her B.Sc degree from Isfahan University of Technology (IUT). Her research interests include Network Security, Trust and Privacy and Big Data mining.

Shahriar Mohammadi is a former senior lecturer at the University of Derby, UK. He received his Ph.D in Computer Science (Network Security) from University of Salford, Manchester UK in 1993. Then, while he used to be a Network consultant, he worked in UK universities of Salford and Derby of the UK for more than fifteen years as a lecturer and senior lecturer. He currently is a lecturer in the Industrial Eng. Department of the University Of K.N.Toosi, of Iran. His main research interests and lectures are in the fields of Networking, Data Security, Network Security, e-commerce and e-commerce Security. He has published more than hundred and twenty papers in various journals and conferences as well as seven books.

COGNISON: A Novel Dynamic Community Detection Algorithm in Social Network

Hamideh Sadat Cheraghchi*

Department of Computer Engineering and Science, Shahid Beheshti University, Tehran, Iran
h.s.cheraghchi@gmail.com

Ali Zakerolhossieni

Department of Computer Engineering and Science, Shahid Beheshti University, Tehran, Iran
a-zaker@sbu.ac.ir

Received: 06/Mar/2016

Revised: 15/May/2016

Accepted: 25/May/2016

Abstract

The problem of community detection has a long tradition in data mining area and has many challenging facets, especially when it comes to community detection in time-varying context. While recent studies argue the usability of social science disciplines for modern social network analysis, we present a novel dynamic community detection algorithm called COGNISON inspired mainly by social theories. To be specific, we take inspiration from prototype theory and cognitive consistency theory to recognize the best community for each member by formulating community detection algorithm by human analogy disciplines. COGNISON is placed in representative based algorithm category and hints to further fortify the pure mathematical approach to community detection with stabilized social science disciplines. The proposed model is able to determine the proper number of communities by high accuracy in both weighted and binary networks. Comparison with the state of art algorithms proposed for dynamic community discovery in real datasets shows higher performance of this method in different measures of Accuracy, NMI, and Entropy for detecting communities over times. Finally our approach motivates the application of human inspired models in dynamic community detection context and suggest the fruitfulness of the connection of community detection field and social science theories to each other.

Keywords: Social Network; Clustering; Cognitive Modeling; Evolution.

1. Introduction

The twist from self-reported survey data to autonomous data gathering enabled by Web 2 and new technologies e.g. smart phones, email, and other smart data gathering gadgets change the dimension and order of data to be analyzed unprecedentedly. Mining such data to recognize and track linked interactions of individuals is a critical area of interest for analyst due to its wide application from cybersecurity to recommender and trade systems. This problem known as community detection problem in social network is one of the well-studied areas of research during the past decades and is linked to general data clustering problem. The existence of linked data is distinguishing feature of modern community detection in social network context versus traditional point-based data clustering.

Various challenging facets of community detection has brought different solutions to this problem from computer engineering, physics, and social science perspectives and changed it to a multi-disciplinary problem. The well-studied statistical inference-based models [1], hierarchical algorithms [2], spectral and modularity-based models [3] are among these models. Survey papers [1-3] are referred for a complete review. These techniques are designed basically to capture the communities in static networks; i.e. network data is gathered in one time step or in the case of multiple time

steps, data is mixed to have the picture of the network in one snaps. With rapid growth of online social networks, where users' joining in and withdrawing from communities are common, dynamic network evolution is recognized as a very valuable research domain for content management and recommender systems [4]. Designing such dynamic algorithms to capture different events happening in the network has its own challenges. Accounting for unforeseen change in topological structure and at the same time temporal smooth overlapping structure are among these challenges. Meanwhile, most algorithms in dynamic settings are extension of static algorithms which will be discussed in Section 2.

Interestingly, social scientists were the prime group of researchers who were always attracted to study the whole network evolution and dynamism of individuals' interaction over different time steps to find its underlying principles. There are numerous school of thoughts for exploring the principles behinds individual mechanisms leadings to versatile network dynamism and community evolution. The famous sociologist, Barry Wellman [8] introduced five main principles to explore intellectual disciplines underlying networks. These fundamental principles mainly focus on relation of individuals to each other rather than individual attitude or demographic characteristic for predicting their behavior. He emphasizes on the dynamic context of relationships and the imposed effects of relationships on everyone in the network. Further,

* Corresponding Author

there are numerous theoretical roots on why people create and maintain groups as discussed in [9] including theories of self-interest, social exchange or dependency, mutual or collective interest, homophily and cognitive theory. Prototype theory as a model for selecting the groups to join is also derived from cognitive science findings.

In this research, we propose our community detection computational model by taking advantage of related social theories devised for exploration of dynamic behavior of human in joining/leaving communities in social network. For this purpose, we base our community detection algorithm according to general approach that human uses for selection: i.e. selection according to prototypes. Then, tracking the evolution of communities is achieved by following the dynamic relationship between entities and the communities and among the communities themselves. Hence, our algorithm is placed in category of prototype-based algorithms like k-means. Prototype based algorithms needs similarity measure for assigning nodes to communities. For this, cognitive consistency theory is our choice among various cognitive theories proposed. This theory helps to explain the natural tendency of human to decrease conflicting cognitions and attach to groups whom feels similar. This is also explained in homophily theory which explains willingness of individuals to communities where there are peoples one deems to be similar. In fact, similarity helps to reduce potential conflicts and increase predictability of behaviors. These two theories are related since choosing similar other persons reduces possible conflicts and increase consistency among members. Leveraging these disciplines, we devise our algorithm called COGNISON (COGNitive inspired community detection in Social Network) to tackle community detection problem in dynamic setting without any parameter or initial settings. Besides, our algorithms works for both weighted and binary networks which make it a preferable choice for social community detection problems.

The paper is organized as follows: Section 2 presents a preview of common background knowledge in community detection context. Section 3 explains our proposed approach and Section 4 represents the experiments to evaluate COGNISON. Finally, we conclude with some highlights on our future research directions.

2. Related Work

As already discussed, traditional community detection approach are designed intrinsically to capture the communities in static networks which limits the analysis of the networks and ignores events which may be much easier to predict if one had the whole picture of the network in different time steps. Hence, a new line of research has been developed focusing on tracking communities and events happening during different time steps; i.e. dynamic dimension of communities [4]. Here, we give a summary of two main lines of research for studying the evolution of communities.

The first and the most intuitive approach uses some static community detection algorithms in each time step of the network. The changes undergone by the communities in different time steps are tracked according to some similarity/distance measures such as intersection of communities to determine the relationship among communities thereby tracking dynamicity of communities. This approach is called *independent community mining*. The main characteristic of this approach is discovering communities from the scratch in each time step using independent or equal algorithms. In the other category, there are algorithms that incorporate the information obtained in other snapshots for extracting communities of future time step and is called *incremental community detection*. This approach of algorithm improves the time and computational complexity compared to independent community detection approach [5-7]. In one dominant approach in this category, a cost-function is calculated in each time step trying to minimize the changes happening to communities in the following time step. This approach assumes that abrupt changes in subsequent time steps are unlikely and these changes have small impact on the community structure. This concept was introduced by Chakrabarti et al.[8] who coined the term *evolutionary clustering* in which two potentially contradicting criteria in an additive equation should be optimized. The first criterion is the correspondence of clustering result to current data as much as possible (*clustering quality*) and the second is keeping the shifts of the results between current clustering and previous time step as low as possible (*history quality*) to allow for *temporal smoothness* as formulated in equation 1. Notice that cost-based approaches is unable to handle drastic changes happening in the network In fact, the assumption of small changes in the network in cost-based approach limits its applicability when abrupt changes happens due to different reasons. Problems with choosing the ideal value for smoothing parameter (α in equation 1) responsible for tuning the weight to place on historic or clustering quality is also addressed in [9].

$$\text{Cost}_{\text{total}} = \alpha \text{Cost}_{\text{quality}} + (1 - \alpha) \text{Cost}_{\text{quality}} \quad (1)$$

Furthermore, there are some other *direct methods* with different definitions for quality and temporal cost. Whatever the definition is, cost functions are incorporated in different static clustering modularity [10], spectral [11], and inference-based [12] paradigms to capture the dynamic of the network. Survey papers [4,13,14] are used for a complete review. Following we review shortly some other important group of community detection paradigms.

In *Modularity-based algorithms* dense communities are recognized using modularity criterion [15,16] in which agglomerative algorithms greedily optimize modularity criterion. Modification of these algorithms is the basis of dynamic modularity based algorithms. For example, Gorke et al. [10] take the changing nodes in different time steps into a separate community and revise the membership according to some merge functions based on modularity criterion. However, modularity-based

community detection algorithms bear some weaknesses [17]: inability to handle noise and a large number of high score communities which avoid recognition of specific community structures [18] are among these problems. *Evolutionary spectral clustering* approach finds an n -dimensional placement of nodes according to a variation of adjacency matrix, e.g. eigenvalues as a cost function to regularize temporal smoothness [11] or investigate the changes to eigenvalues in different time steps [19]. The weakness of spectral-based clustering algorithms lies in high computational cost incurred during matrix multiplication which makes it a weak choice for very large networks. *Inference-based methods* are also another broad category which considers an underlying statistical model that can generate communities in the network. FacetNet [12] is the first probabilistic generative model proposed for analyzing evolution of communities and several other works have been introduced recently [20]. This category also suffers from high memory usage [21].

Usage of cognitive and social theories in clustering domain is also taking new dimensions recently. Apart from famous k-means algorithm and its derivations for example belief k-mode clustering [22] which are all placed in this category, cognitive inspired clustering are leveraged in different applications. Data dissemination based on cognitive theories [23] and linguistic clustering by prototype theory [24] are among recent approaches in this category. Human group formation based on homophily or similarity among members is also verified in different studies [25]. Now, we take advantage of social theories to introduce a new community detection algorithm in social networks applicable in both binary and weighted networks. In contrast to k-means like algorithms, the number of communities will be determined automatically according to characteristic of the network.

3. Proposed Model

Suppose a network G with n node where the interaction between each pair of members in time t is indicated by a symmetric binary/weighted link. This information can be represented in a proximity matrix W^t where W_{ij}^t denotes the proximity between member i and j according to the link weight of connected items (edge weight of binary network is 1).

Now, we describe how cognitive inspired disciplines helps to design an efficient community detection algorithm for tracking the evolution of communities. For this purpose, we leverage the approach that human uses for categorization: i.e. prototype based selection in our community detection algorithm and track the evolution of communities by following the dynamically updated structures between entities and the communities in the history of the network. In fact, communities are created and altered online in regard to the observed changes in the network. How to follow this relationship is the key to track the dynamic of communities. We explore

functionality of this algorithm in two key phases of recognition and categorization. In the *recognition* phase, members are introduced to communities present in the networks and selection process according to a cognitive inspired similarity measure happens. In the second phase of *learning*, prototypes are updated to reflect the events happened in the network and set as proper candidate for future selection scenarios. These steps are explained in details in the next sections.

3.1 Recognition Phase

In this phase, members explore the environment and a selection process takes place similar to operation of prototype theory. Each input entry i along with its interacting neighbors recorded in W_i^t finds community prototypes available from past history and decides on the best one to join. This is through similarity checking of objects against prototypes. Now, we elaborate on the details of structures and similarity measure used in COGNISON.

Input entry leverages the minimum local information available of its own id and their neighbor ids, the frequency of interaction with its neighbor and the time of interaction to construct its own input structure and prototype structure. Hence, minimal features stored in prototype structure are ID member, frequency and time step of linked data observed ($\langle id, frequency, time \rangle$). Obviously, at the beginning of the first time step, there is no community prototypes and the values of the first linked input entry are stored in the first community prototype. As the subsequent entries are entered, they are checked by a similarity measure to see whether they can be included in one of the existing prototypes or a new one should be created to accommodate properties of this entry. This process continues until all nodes in current time step are assigned to their communities. At beginning of next time steps, the available prototypes derived in previous time step are leveraged for comparison purpose. This is compatible with the idea that members tend to preserve their membership to their assigned communities in previous time steps.

Now, we elaborate on how to compute similarity measure of each data entry to available category prototypes. Assignment to one of the prototypes is steered by cognitive consistency theory. For this, we use finding of experiments directed by behavioral researchers in which to assess predictability of friends' revisit in the future, one should consider the effects of *frequency* and *spacing* pattern of previous visits as major elements for prediction of future visits [26,27]. In their definition, frequency is the rate of previous visits and recency is the time elapsed since last visits. Whenever frequency and recency of visit is high, there is high possibility of revisit. We use this concept to categorize each input entries in the most similar community prototype and derive our measure to assess similarity of input entry I to each prototype $C_i^* \in C^*$ as follows:

$$Likeness(I, C_k^*) = \sum_{j \in M} Activity_j * Recency_j \quad (2)$$

$$Recency_j = e^{\frac{t-t_j}{t}} \quad (3)$$

In this equation, similarity is computed in *Likeness* formula which is computed based on two features of common members between input entry I and examined prototype C_i^* (M denotes common members). The more common members exists in one prototype, the higher is the chance of selecting it. However, *activity* and *recency* of those common members in each prototype are other important factors for selection. *Activity* of a member in prototypes is the weights of its interaction when it is included in the prototype. Hence, more a node is observed, its activeness will be higher. For weighted networks, each observation of input entry will record the weight of interaction. Further, *recency* of each member is computed using an exponential function which takes into account the difference among current time and the last time the entry the member is observed. So, if the entry is observed in community in just current time step, this variable takes its highest value of 1 and if observed in older time steps, it acquires a value less than 1 which will decrease the previously seen activity of that member. If more than one community takes non-zero value of similarity, the one with highest measure is selected for inclusion of input entry.

3.2 Learning Phase

After the assignment of nodes to its best recognized communities, an update scheme should take place in the selected prototype to reflect the changes made due to recently added member structure assignment. If a member belongs to more than one community in the final stage, we assign the node to the community in which the node has the highest similarity. In the update process, three main feature of prototypes, i.e. $\langle id, frequency, time \rangle$ each are updated. ID of new member is added if not already present in the prototype structure and *frequency* of interaction is updated by summing up current frequencies of input entries to their old values present in the prototype. Finally, time property of the members present in the current time step is updated. In this way, activity of nodes which have not been observed in previous time step is decreased which helps the algorithms to be responsive to new events while preserving past events. This is achieved when exploiting *likeness* measure (eq. (2) and (3)) for selecting the best prototype.

4. Experimental Results

We examine the performance of the proposed algorithm on both evolving synthetic and real datasets. In the synthetic experimental section, in addition of toy example, we use the stabilized and frequently used synthetic LFR generator is used for artificial dataset generation and for real dataset experiment, the famous of MIT really mining dataset is exploited. The number of entities in synthetic and real dataset may change in

different time steps. Further, the numbers of communities differs in the intervals. Since ARTISON inherits most of its properties from representative-based algorithms, proper comparison is achieved by comparing it to other representative-based algorithms. For this reason, we choose two state-of-art evolutionary k-means algorithms specially designed for dynamic settings of network for comparison purpose. This equals to compare ARTISON with the pioneer evolutionary framework extended to k-means [28] where two temporal and quality costs are optimized with a constant smoothing factor (α in equation (1)) to capture the dynamic of the network. Further, we use another more recent evolutionary framework extended to k-means algorithm called Adaptive Evolutionary Clustering (AFFECT [29]) where optimal smoothing factor is determined automatically using a statistical approach. In all of these case, the optimum number of communities for each time step is determined by well-known silhouette width criterion [30]. This measure determines how compact the distance of communities are in a given time step and the maximum width of this measure is used to assess the number of needed communities in k-means. In addition, we use two other modern hierarchical agglomerative community detection algorithms based on modularity criterion in social network for real dataset experiments to make comparison with state of art algorithms. Louvain [31] and fast modularity [32] where both have acquired high performance in recent survey studies.

For the evaluation, we use four measures to determine the accuracy and quality of the community detection algorithms in different time steps via clustering Rand Index and F measure to indicates the amount of disagreement between discovered communities (C) and the labels of ground truth communities (C^*). F measure is a harmonic mean of precision and recall measures where precision is the ratio of relevant objects (real community member detected) to total number of objects detected and recall is the ratio of relevant objects detected to total real ground truth members. All the mentioned measures reach their best at 1 and their worse at 0 value.

$$F(C, C^*) = \frac{2|C \cap C^*|}{|C| + |C^*|} \quad (4)$$

Higher values of all of these measure are preferred. For quality of clustering we use another common measure in information theory called Entropy [33] to measure the quantity of the disorder observed in the results. Lower value of entropy is preferred which means better clustering result.

4.1 Synthetic Dataset Evaluation

In the first experiment, we use a toy synthetic dataset in which different numbers of communities appear during the test to better verify the dynamic community tracking capabilities of the proposed algorithm. Figure 1 shows the diagram of the synthetic dataset.

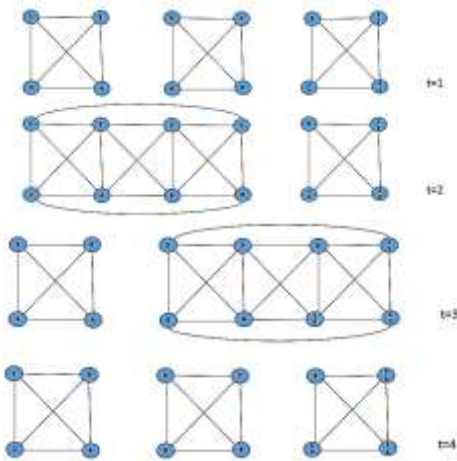


Fig. 1. An example of community evolution in four time step: in $t=1$, three communities, in $t=2$ and $t=3$ two communities and $t=4$ again three community exists.

As indicated in the figure, in the first time step, three communities are recognizable. Then, a merging happens and two communities are distinguishable in $t=2$ and $t=3$. Finally, we come back again to three communities observed first. For LFR generator, we use two experiments of switch and expand/contract events to test the performance of the algorithm. The other parameters of the generator are set as follows: average and maximum degree of node is set to 10 and 50, and minimum and maximum community size is 20 and 100 nodes respectively and the initial number of nodes is 500 nodes. The experiments are averaged over different runs since k-means based algorithms produce different results in multiple run due to initial node. In the switching event, the number of nodes during the whole experiment is fixed but they change their communities with probability of 20%. For the expansion event, we considered three expansions and two contraction events per time step by switching probability of 10% (50 nodes out of 500 nodes switch their community in each time step).

Figure 2 shows the result of competing algorithms in terms of Rand index and F measure. We presented the mean and standard error over all time steps. Since we expect a higher value for Rand index and F-measure for recognition of the preferred community detection approach, we judge COGNISON as the superior one. For the standard deviation, COGNISON in several cases has higher deviation from other algorithms but high difference of the measure compared justify this variation.

Table 1. Comparison of proposed algorithm in synthetic datasets with other protocols in two measures of a) Rand Index, b) F measure.

Dataset	Measures	COGNISON	AFFECT	Evolutionary-k-means
Toy Dataset	Rand	0.86±0.16	0.70±0.05	0.72±0.15
	F	0.80±0.25	0.58±0.14	0.60±0.26
Switch (0.2)	Rand	0.82±0.05	0.42±0.07	0.44±0.10
	F	0.39±0.19	0.21±0.10	0.31±0.13
Expand	Rand	0.86±0.06	0.36±0.15	0.39±0.14
	F	0.51±0.23	0.22±0.14	0.26±0.14

4.2 Real Dataset Evaluation

In this section, we intend to evaluate the performance of COGNISON on the Reality Mining dataset [34] commonly used for dynamic evaluation purpose [35,36]. This dataset is gathered by MIT media lab to analyze the cell phone activity of 90 participants consisting of students and staff interacting over a period of nine months. The large volume of approximately 500,000 hours of data is extracted by monitoring different cell usage of participants logged as incoming and outgoing calls, cell tower id, and any Bluetooth devices discovered during their interactions. Our experiments covers Bluetooth activity of participants which records the IDs of nearby Bluetooth devices (student or student ID) every five minutes. Affiliation of each participant is available to be used as the ground truth information as discovered by Eagle et al. [34]. Further, for finding optimal number of communities, we use silhouette measure [30].

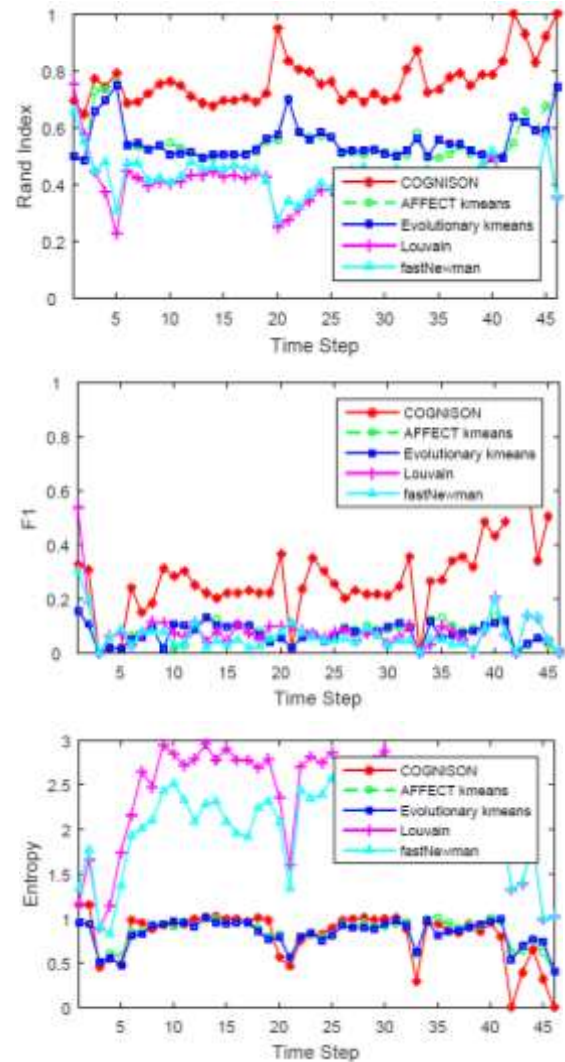


Fig. 2. Comparison of Rand Index and F measure values achieved in real dataset for the five algorithms: COGNISON, AFFECT, evolutionary k-means, Louvain and Fast Newman clustering algorithms.

Table 2. Comparison of proposed algorithm in synthetic datasets with other protocols in two measures of a) Rand Index, b) F measure.

	Rand Index	F-Measure	Entropy
COGNISON	0.75±0.06	0.26±0.07	0.76±0.27
AFFECT	0.55±0.07	0.07±0.04	0.84±0.17
Evolutionary k-means	0.55±0.07	0.07±0.04	0.83±0.15
Louvain	0.41±0.09	0.08±0.08	2.30±0.60
Fast Newman	0.43±0.08	0.07±0.06	1.98±0.45

For entropy measure, lower value shows less quantity of disorder found in community detection and is desired. As depicted in the last row of Table 2, entropy of COGNISON is lower than all.

Notice that COGNISON has several distinguishing features. The ability of discovering the number of communities intrinsically is of great advantage while the other k-means based algorithms assess the number of optimal characters as inputs or find it through some other calculations external to the algorithm (e.g. using silhouette [37] as used in our experiments) for the initialization of the algorithm. Second, the algorithm is free of any smoothing factor commonly used for evolutionary algorithms. In fact, in COGNISON, there is

no tradeoff between quality and history costs which makes it more robust to changes.

5. Conclusion and Future Works

Following the stream of works presented for dynamic community detection, specifically evolutionary clustering algorithms, we proposed another online dynamic community detection algorithm in social network context called COGNISON. While the initialization of each time step takes community snapshot of the previous time step into account, different mechanisms for link weighting cause the enforcement of strong links and weakening of weak links not present in the previous steps. This helps to solve a big challenge in most community detection algorithms; i.e. knowing the number of communities to pass as an input to the algorithm. The experimental results displayed the good performance of this algorithm against the state of art evolutionary algorithms and encourage the ongoing works on dynamic community detection to take more advantage of cognitive-inspired paradigms.

References

- [1] D. J. MacKay, *Information theory, inference and learning algorithms*: Cambridge university press, 2003.
- [2] K. Sasirekha and P. Baby, "Agglomerative hierarchical clustering algorithm—A Review," *International Journal of Scientific and Research Publications*, vol. 3, 2013.
- [3] T. N. Dinh and M. T. Thai, "Community detection in scale-free networks: approximation algorithms for maximizing modularity," *Selected Areas in Communications, IEEE Journal on*, vol. 31, pp. 997-1006, 2013.
- [4] T. Aynaud, E. Fleury, J.-L. Guillaume, and Q. Wang, "Communities in evolving networks: Definitions, detection, and analysis techniques," in *Dynamics On and Of Complex Networks*, Volume 2, ed: Springer, 2013, pp. 159-200.
- [5] M. Takaffoli, J. Fagnan, F. Sangi, and O. R. Zaïane, "Tracking changes in dynamic information networks," in *Computational Aspects of Social Networks (CASoN)*, 2011 International Conference on, 2011, pp. 94-101.
- [6] D. Greene, D. Doyle, and P. Cunningham, "Tracking the evolution of communities in dynamic social networks," in *Advances in Social Networks Analysis and Mining (ASONAM)*, 2010 International Conference on, 2010, pp. 176-183.
- [7] T. Falkowski, J. Bartelheimer, and M. Spiliopoulou, "Mining and visualizing the evolution of subgroups in social networks," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 2006, pp. 52-58.
- [8] D. Chakrabarti, R. Kumar, and A. Tomkins, "Evolutionary clustering," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 554-560.
- [9] J. Zhang, Y. Song, G. Chen, and C. Zhang, "On-line Evolutionary Exponential Family Mixture," in *IJCAI*, 2009, pp. 1610-1615.
- [10] R. Görke, P. Maillard, A. Schumm, C. Staudt, and D. Wagner, "Dynamic graph clustering combining modularity and smoothness," *Journal of Experimental Algorithmics (JEA)*, vol. 18, p. 1.5, 2013.
- [11] Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng, "Evolutionary spectral clustering by incorporating temporal smoothness," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 153-162.
- [12] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng, "Facetnet: a framework for analyzing communities and their evolutions in dynamic networks," in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 685-694.
- [13] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, "Community detection in social media," *Data Mining and Knowledge Discovery*, vol. 24, pp. 515-554, 2012.
- [14] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng, "Analyzing communities and their evolutions in dynamic social networks," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, p. 8, 2009.
- [15] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical review E*, vol. 69, p. 066133, 2004.
- [16] S. Cafieri, P. Hansen, and L. Liberti, "Locally optimal heuristic for modularity maximization of networks," *Physical Review E*, vol. 83, p. 056105, 2011.
- [17] E. Eaton and R. Mansbach, "A Spin-Glass Model for Semi-Supervised Community Detection," in *AAAI*, 2012.
- [18] B. H. Good, Y.-A. de Montjoye, and A. Clauset, "Performance of modularity maximization in practical contexts," *Physical Review E*, vol. 81, p. 046106, 2010.
- [19] H. Ning, W. Xu, Y. Chi, Y. Gong, and T. S. Huang, "Incremental spectral clustering by efficiently updating the eigen-system," *Pattern Recognition*, vol. 43, pp. 113-127, 2010.
- [20] L. Tang, H. Liu, and J. Zhang, "Identifying evolving groups in dynamic multimode networks," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, pp. 72-85, 2012.
- [21] M. B. Hastings, "Community detection as an inference problem," *arXiv preprint cond-mat/0604429*, 2006.

- [22] S. B. Hariz, Z. Elouedi, and K. Mellouli, "Clustering approach using belief function theory," in *Artificial Intelligence: Methodology, Systems, and Applications*, ed: Springer, 2006, pp. 162-171.
- [23] M. Conti, M. Mordacchini, and A. Passarella, "Design and performance evaluation of data dissemination systems for opportunistic networks based on cognitive heuristics," *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, vol. 8, p. 12, 2013.
- [24] H. Zhao and Z. Qin, "Clustering Data and Vague Concepts Using Prototype Theory Interpreted Label Semantics," in *Integrated Uncertainty in Knowledge Modelling and Decision Making*, ed: Springer, 2015, pp. 236-246.
- [25] N. F. Johnson, C. Xu, Z. Zhao, N. Ducheneaut, N. Yee, G. Tita, et al., "Human group formation in online guilds and offline gangs driven by a common team dynamic," *Physical Review E*, vol. 79, p. 066117, 2009.
- [26] M. M. Krasnow, A. W. Delton, J. Tooby, and L. Cosmides, "Meeting now suggests we will meet again: Implications for debates on the evolution of cooperation," *Scientific reports*, vol. 3, 2013.
- [27] T. Pachur, L. J. Schooler, and J. R. Stevens, "We'll Meet Again: Revealing Distributional and Temporal Patterns of Social Contact," 2014.
- [28] Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng, "On evolutionary spectral clustering," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, p. 17, 2009.
- [29] K. S. Xu, M. Kliger, and A. O. Hero Iii, "Adaptive evolutionary clustering," *Data Mining and Knowledge Discovery*, vol. 28, pp. 304-336, 2014.
- [30] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis* vol. 344: John Wiley & Sons, 2009.
- [31] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, 2008.
- [32] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, p. 066111, 2004.
- [33] Y. Yao, "Information-theoretic measures for knowledge discovery and data mining," in *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, ed: Springer, 2003, pp. 115-136.
- [34] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 15274-15278, 2009.
- [35] J. More and C. Lingam, "Current trends in reality mining," ed: IRJES, 2013.
- [36] H. Zhang, R. Dantu, and J. W. Cangussu, "Socioscope: Human relationship and behavior analysis in social networks," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 41, pp. 1122-1143, 2011.
- [37] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53-65, 1987.

Hamideh Sadat Cheraghchi received the BSc degree from Azad University, South Tehran branch, Tehran, Iran in 2006 MSc. from Azad University, Qazvin branch in 2009 and she is currently a PhD candidate in computer architecture in the department of Computer Engineering and Science at Shahid Beheshti University, Iran. Her research focuses on data mining in social network under the supervision of Dr. Ali Zakerolhosseini.

Ali Zakerolhosseini received the BSc degree from university of Coventry, UK, in 1985, MSc from the Bradford University, UK, in 1987, and PhD degree in Fast transforms from the University of Kent, UK, in 1998. He is currently an assistant professor in the department of Computer Engineering and Science at Shahid Beheshti University, Iran. His research focuses on Reconfigurable device and multi classifiers. His current research interests are Data Security, Cryptography and Social Data Mining.

Analysis and Evaluation of Techniques for Myocardial Infarction Based on Genetic Algorithm and Weight by SVM

Hodjatollah Hamidi*

Department of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran
h_hamidi@kntu.ac.ir

Atefeh Daraei

Department of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran
adaraei@mail.kntu.ac.ir

Received: 22/Apr/2016

Revised: 23/May/2016

Accepted: 31/May/2016

Abstract

Although decreasing rate of death in developed countries because of Myocardial Infarction, it is turned to the leading cause of death in developing countries. Data mining approaches can be utilized to predict occurrence of Myocardial Infarction. Because of the side effects of using Angioplasty as main method for diagnosing Myocardial Infarction, presenting a method for diagnosing MI before occurrence seems really important. This study aim to investigate prediction models for Myocardial Infarction, by applying a feature selection model based on Wight by SVM and genetic algorithm. In our proposed method, for improving the performance of classification algorithm, a hybrid feature selection method is applied. At first stage of this method, the features are selected based on their weights, using weight by SVM. At second stage, the selected features, are given to genetic algorithm for final selection. After selecting appropriate features, eight classification methods, include Sequential Minimal Optimization, REPTree, Multi-layer Perceptron, Random Forest, K-Nearest Neighbors and Bayesian Network, are applied to predict occurrence of Myocardial Infarction. Finally, the best accuracy of applied classification algorithms, have achieved by Multi-layer Perceptron and Sequential Minimal Optimization.

Keywords: Artificial Neural Network; Sequential Minimal Optimization; REPTree; Knowledge Discovery in Databases; Myocardial Infarction.

1. Introduction

Myocardial Infarction (MI) is a prominent cause of death all around the world [1]. In spite of reduction in mortality rates because of cardiovascular diseases in developed countries, it is known as a significant cause of death in developing countries [2]. Based on [3] it is likely to ischemic heart disease become the most common cause of death by 2020 [3]. Myocardial Infarction is turned to the main cause of death for Iranian people above 35 years [4]. As mentioned in [5], Ischemic heart disease is known as the main cause of death in Western countries. Ischemic heart disease is happened due to Atherosclerosis. Atherosclerosis is the process of depositing Cholesterol and fat in the vessels especially in people with genetic susceptibility, overweight and obesity, inactive way of life, high blood pressure, and people who consume a lot of cholesterol and fat. This is lead to production of atherosclerotic plaques which cause the fully or partially blockage of blood [5]. Based on Ministry of Health and Medical Education report in 2012, the rate of mortality because of Myocardial Infarction is estimated 85 per 100,000 [2]. Myocardial Infarction means the death of heart muscle due to a sudden blockage of a coronary artery by a blood clot [6]. Thus, immediately after coronary occlusion, blood flow is stopped in the coronary vessels beyond the blocked site with the exception of a

very small amount of collateral blood flow in surrounding vessels. So an area of the heart muscle with zero blood flow or with very low blood flow to do muscle action has been infarcted. This process is called Myocardial Infarction [5].

Data mining methods transform raw data into useful information [7]-[8]. Therefore, data mining can be regarded as a tool for acquiring knowledge from raw and meaningless data in the medical field, and in the field of Myocardial Infarction in particular. Data mining is a part of knowledge discovery in databases (KDD) [9]; in [10] data mining is introduced as a KDD process includes three stages: Data Preprocessing, Data Modeling and Data Post Processing [10]. Data modeling tasks in the data mining process are divided into two categories: predictive tasks and descriptive tasks. The algorithms of the predictive category include classification algorithms for discrete data, and regression algorithms, for continuous data [11], which are learned through supervised learning process [7].

The algorithms which determine the class labels based on the training data with particular labels are called classification algorithms [7]. Early detection of occurrence Myocardial Infarction seems to be very important because of its important role of mortality in the world and Iran. Angiography is a costly and non-invasive method for detecting the blockage in vessels and may have complications for patients such as bleeding at the insertion

* Corresponding Author

site to the vein, stroke, vascular injury, renal failure and even death. Thus, due to side effects of drugs and angiography [12], using prediction methods could consider really important. The aim of this study is early prediction of Myocardial Infarction using the classification algorithm and a proposed feature selection. In addition, a combined feature selection method was used along with the classification techniques. To the best of our knowledge, using genetic algorithm along with feature selection method has not been used so far. The algorithms Naïve Bayes, Support Vector Machine, Artificial Neural Network, K-Nearest Neighbors, Sequential Minimal Optimization, REPTree, Random Forest and Bayesian Network are used for reaching to this purpose. The paper is organized as follows: In second section the related researches in this field are reviewed. Section three describe the Myocardial Infarction data set. Section four explain the proposed model. Experimental Results comes in fifth section. In the section sixth the results are analyzed. The section seventh contains conclusion.

2. Literature Review

Many researches have been worked in this field. In a study, Baxt et al. in 2002, proposed a model for prediction of acute Myocardial Infarction. They used artificial Neural Network method. Data set used in this study consisted of 2204 patient with 40 attributes. The results showed that the Neural Network has a high potential for use in prediction of acute Myocardial Infarction [13]. The authors in [14] presented a system which detects abnormality in the heart and its walls. They used a real data set which contains 141 samples. After Feature selection, Support Vector Machine (SVM) algorithm was applied to the data. The results show that the selection of 3 main features selected by feature selection leads to high efficiency.

Conforti et al. in [15] employed a Support Vector Machine with 5 core functions include linear, Gaussian, Laplacian, polynomial and sigmoid. The aim was to categorize the patterns for detection of acute Myocardial Infarction (AMI). They used a dataset consisted of 242 patients with chest pain. 105 features in this data set were divided to 4 categories: history, ECG, electrocardiogram, and blood test. The best accuracy in 7 different Feature selection modes, on average, was equal to 85.85% for linear and polynomial function, 82.64% for the Gaussian, 86.43% for Laplace and 82.4% for the sigmoid. In [16], is used a Neural Network for diagnosing heart attacks. UCI datasets is used in this paper, consisted of 13 features. After preprocessing, clustering is used to determine categories. After the rules extracted in clusters using Mafia algorithm, a three layered Neural Network is applied on data. In [17], a model is presented to predict Myocardial Infarctions and coronary artery bypass graft. The algorithm C4.5 decision tree is used for classification. The data set consisted of 1200 patients. Size of the data set is reduced to 369 after excluding the inappropriate

samples. After applying C4.5 tree, they extracted the relevant rules and the most important factors affecting the occurrence of Myocardial Infarction. In 2010, Arif et al. used Back Propagation Neural Network for detection Myocardial Infarction MI and determine its location. The main features was Q and T waves as well as ST-Elevation or ST-Depression. The data set used in this study consisted of 148 MI patients. Neural Network algorithm was used for both detection MI and determining its location. Sensitivity and specificity measures were used to assess the results of detection of MI (97.5% and 99.1%, respectively) [18]. In [19], 5 classification algorithms is employed to develop a model for prediction of heart attack The algorithms used for this purpose included J48 decision tree, Bayesian Network, Naïve Bayes(NB), CART and REPTREE. 11 features were investigated in this study. According to the experimental results, J48, Naïve Bayes and CART classification methods achieved higher accuracy (99.07%) and proved to be better than Naïve Bayes techniques and Bayesian Network. In 2014, the authors used a model of multi-layered feed-forward Neural Network algorithm predicting heart attack, which utilized genetic algorithm for initialization and discovering the optimal weights for multi-layered feed forward Neural Network. The data set used in this study, obtained from UCI, which includes 270 cases and 13 features. Finally, in the evaluation stage, the propose method obtained accuracy of 88% [20]. In [21] is attempted to detect and locate MI using 4 algorithms include Probabilistic Neural Network (PNN), K-Nearest Neighbors (K-NN), Multi-layer Perceptron (MLP), and Naïve Bayes. The dataset used in this study included information of 549 patients with 15 features. After applying the algorithms to the data the highest accuracy for the detection of MI was obtained by the Naïve Bayes method, which was equal 94.74%. The obtained accuracy was lower for the diagnosis of MI. The author in [22] aimed to classify ECG signals as healthy or Myocardial Infarction. The data set used in this study was the PTB data set, included 82 healthy cases and 367 cases of Myocardial Infarction. After applying two methods of artificial Neural Network and Support Vector Machine on the data set, the results showed that SVM has higher accuracy than the artificial Neural Network. Sharma et al. in 2015, proposed a model for diagnosing Myocardial Infarction using K-Nearest Neighbors and SVM. The kernel functions Lin and RBF used for SVM algorithm. The accuracy obtained for K-NN was equal 81%. The accuracy resulted in using SVM was higher, so that SVM with Lin function obtained the accuracy 89% and SVM with RBF achieved to highest accuracy equal to 96% [23]. Kora and Kalva in 2015, proposed a model of Neural Networks (LN and SCG), K-NN and SVM combined with improved BAT algorithm for diagnosing heart attack. They used ECG features of PTB data base. This data base, used in this study, consisted of 52 normal cases and 148 MI. BAT algorithm was used for feature extraction. The results showed that, the highest accuracy

belongs to Neural Network (LM) which is equal to 98.1%. The other algorithms' accuracy consisted of 87.9% for Neural Network (SCG), 65.1% for K-NN and 76.74% for SVM [24].

3. Myocardial Infarction Data Set Description

In this study, a dataset consisted of the information collected from 519 visitors to Shahid Madani Specialized Hospital of Khorram Abad, include 222 without MI and 297 with MI, is used. After reviewing patient records and consulting with specialists, and based on [3] and [12] books, features such as, Troponin I, CRP, LDH and three main heart arteries include LAD, RCA and LCX, were added to the features in articles reviewed. The dataset contains 52 features (51 predictor features and 1 goal feature), presented in table 1.

4. The Proposed Model

4.1 Preprocessing Process

Preprocessing method is an important part of data mining process which prepare the raw data for data mining [25]. Raw data is usually incomplete, noisy and incompatible [26] because of huge size of databases, integration of several various datasets or human errors [27]. In this study the missing values are filled using the same available values. Moreover, normalization, as a method for converting the data into a suitable form for data mining, is used in this study. In this study, the Min - Max normalization method on the interval of [0, 1] is used.

Table 1. The Features of Myocardial Infarction Data Set

Features	Range
Age	28-93
Sex	M, F
Weight	43-120
Body Mass Index (BMI)	16-42
Systolic Blood Pressure (SBP)	80 – 210
Diastolic Blood Pressure (DBP)	40 – 190
Heart Rate	50 – 190
Family History	Yes , No
Smoke	Yes , No
Obesity	Yes , No
Hypertension	Yes , No
Chronic Renal Failure (CRF)	Yes , No
Cerebrovascular Accident (CVA)	Yes , No
Congestive Heart Failure (CHF)	Yes , No
Dyslipidemia (DLP)	Yes , No
Blood Pressure	Yes , No
Edema	Yes , No
Fatigue and weakness	Yes , No
Lung Rales	Yes , No
Typical Chest Pain	Yes , No
Distribution of pain to arms and neck	Yes , No
Dyspnea	Yes , No
Atypical CP (Atypical Chest Pain)	Yes , No
Non-anginal CP (Non-anginal Chest Pain)	Yes , No
Exertional CP (Exertional Chest Pain)	Yes , No

Features	Range
Troponin I	Pos.-Neg.
C-reactive protein (CRP)	Pos. – Neg.
Total Cholesterol	22 – 480
Fasting Blood Sugar (FBS)	23 – 550
Lactate Dehydrogenase (LDH)	30 – 1000
Creatine Phosphokinase (CPK)	8 – 650
Creatinine (Cr)	0.3 – 18
Triglyceride (TG)	70 – 400
Low Density Lipoprotein (LDL)	38 – 182
High Density Lipoprotein (HDL)	21 – 102
Ejection Fraction (EF)	10 – 80
Lymphocyte	2- 81
Platelet	90 – 1700
Blood Urea Nitrogen (BUN)	10 – 80
Erythrocyte Sedimentation Rate (ESR)	4 – 76
Hemoglobin	3.5 – 23
Potassium (K)	2.5 – 12.5
Sodium (Na)	62 – 192
White Blood Cells (WBC)	3200- 20000
ST Elevation	Yes , No
ST Depression	Yes , No
T inversion	Yes , No
Poor R Progression	Yes , No
LAD (left Anterior Descending artery)	Yes , No
LCX (Left Coronary Artery)	Yes , No
RCA (Right Coronary Artery)	Yes , No

4.2 Proposed Feature Selection Method

Feature selection is one of the most common used methods of dimensionality reduction. In this methods after removing the irrelevant features, the best features is kept and used [9].

At the first stage of the proposed feature selection method used in this study, weight by SVM is applied. Using this operator, first the features' weights are specified. These features, then, are used in Genetic Algorithm, at the second stage. Genetic algorithm finds the best hypothesis by searching a hypothesis space. The initial hypotheses called a population is randomly generated and a fitness function is used for evaluating each hypothesis. Hypotheses with greater fitness have the higher probability of being chosen to create the next generation. Some of the best hypotheses may be retrained at the next generation, the other operations, crossover and mutation are used to generate new hypotheses. The size of population is same for all generations [28]. To the best of our knowledge, in subject of Myocardial Infarction prediction, the two using Genetic Algorithm along with feature selection is not used so far.

4.3 Classification Algorithms

Classification algorithms employed in this study, to assess the performance of proposed feature selection, consist of REPTree, Random Forest, Bayesian Network, Support Vector Machine, Multi-layer Perceptron, K-Nearest Neighbors, Sequential Minimal Optimization and Naïve Bayes.

4.3.1 RRPTree

REPTree utilize the regression tree logic and creates many trees in various iterations. Next it chooses the best

tree from all created trees. In general, Reduced Error Pruning Tree (REPTree) is a fast decision tree learning method that create the trees in iterations and prunes them using reduced error pruning [29].

4.3.2 Random Forest

Random forest is a classifier contain of a combination of trees. Each tree is produced of a random vector sampled of the input vector. Every tree considered as a unit, which vote for the most popular class to classify instance [30].

4.3.3 Bayesian Network

Bayesian Network is a network based on the relationship between attributes. It utilizes statistic techniques to represent probability this relationships. This algorithm, similar to Naïve base, uses Bayes rule [31].

4.3.4 Multi-Layer Perceptron (MLP)

Neural Network is capable of predicting new observations, from earlier ones, after executing the learning process using the past data [32]. One of the mostly used learning algorithm for Neural Networks is Multilayer Perceptron [33].

4.3.5 K-Nearest Neighbors (K-NN)

K-Nearest Neighbors algorithm is based on comparing a given test instance with training instances which are similar to it. When an unknown instance is given, the algorithm searches the pattern space for the k training instances that are closest to the unknown instance. These k training instances are the k “nearest Neighbors” of the unknown instance [26].

4.3.6 Support Vector Machine (SVM)

SVM is a classification algorithm, which is based on the statistical learning theory [34]. This algorithm maps input vectors to higher dimensional spaces where maximal separating hyper-planes are constructed. Two parallel hyper-planes are constructed on both sides of a data searching hyper-plane. The separating hyper-plane is the one that maximizes distance between two parallel hyper-planes [35].

4.3.7 Sequential Minimal Optimization (SMO)

SMO is used for learning SVM algorithm. This algorithm decrease the time of obtaining the weights for SVM, in optimization problems. This low time is due to using the serial optimizing methods and the linear memory that this methods require. SMO doesn't need ant matrix storage. Since no matrix algorithms are employed in SMO, its sensitivity to numerical problems' precision is low [36].

4.3.8 Naïve Bayes (NB)

The Naïve Bayes algorithm is a probabilistic classifier based on conditional probability. It means, the Naive Bayes classifier uses probability to classify the new instance [9]. The algorithm uses of all the features contained in the dataset, and based on its main assumption, Conditional

independence. In fact, it considers the features equally important and independent of each other [37].

5. Experimental Results

In this study, version 7.0.1 of RapidMiner is used for classification. To evaluate the performance of the classification models accuracy, sensitivity and specificity can be used, which are obtained using K-fold cross validation. They are the most popular measures for evaluating the classification performance.

Classification accuracy is one of the most common metrics for evaluating performance of the model. It is the ratio of TP and TN obtained by model to the total number of instances [38], as shown by equation (3):

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

where TN, TP, FP and FN represents the number of true negatives, true positives, false positives and false negatives, respectively [39].

Sensitivity and specificity are also two performance measures for evaluating of the efficiency of a classification technique. Sensitivity, or true positive rate, is the ratio of positive instances have been truly classified as positive. Specificity is the ratio of negative instances have been truly classified as negative [40].

5.1 Results

This section present the experimental results of the proposed classification models. These algorithms are compared to each other in 2 different statues, using or not using of feature selection method. The input parameter for selection of features using weight by SVM is set to “top p%”, where p value is equal to 0.7. It means the top 70% of all features are selected.

For genetic algorithm the best results are obtained when the number of generations is set to 10 and the population size set to 7. The other parameters are used at the software default. After using Genetic Algorithm, it selects 32 features from the 36 features selected using weight by SVM. The weights of the features selected in first stage and selected features after implementing second stage of feature selection is shown in Table 2. Third column of this table represent the selected features at end of the feature selection model.

Thus, the final features selected by proposed feature selection method include: CHF, Troponin I, CRP, Total cholesterol, LDL, HDL, ESR, Lymphocyte, BMI, sex, Family History, Obesity, CRF, CVA, FBS, Cr, TG, BUN, Hemoglobin, Na, EF, Edema, Lung rales, Typical chest pain, Dyspnea, Non-anginal CP, Exertional CP, ST Depression, T inversion, ST Elevation, Poor R Progression, LAD.

It is worth noting that the parameter K is set to 4. The other algorithms are implemented in defaults of software. Table 3 shows the results for implemented classification algorithms on

Myocardial Infarction dataset, in two states: using proposed feature selection method and not using feature election.

Table 2. Selected Features by Proposed Feature Selection Method

Attribute	Weights of selected features after first stage of FS	Selection status after second stage of FS
ST Elevation	1.0	✓
Troponin I	0.98	✓
ST Depression	0.96	✓
CRF	0.94	✓
Exertional CP	0.92	✓
T inversion	0.9	✓
Obesity	0.88	✓
K	0.86	×
TG	0.84	✓
HDL	0.82	✓
ESR	0.8	✓
FBS	0.78	✓
Lymphocyte	0.76	✓
Edema	0.74	✓
Hemoglobin	0.72	✓
Lung rales	0.7	✓
BUN	0.68	✓
EF	0.66	✓
Dyspnea	0.64	✓
Poor R Progression	0.62	✓
LDL	0.6	✓
Non-anginal CP	0.58	✓
Typical chest pain	0.56	✓
Atypical CP	0.54	×
heart rate	0.52	×
Na	0.5	✓
Family History	0.48	✓
CVA	0.46	✓
CHF	0.44	✓
sex	0.42	✓
Total cholesterol	0.4	✓
DLP	0.38	×
LAD	0.36	✓
CRP	0.34	✓
BMI	0.32	✓
Cr	0.3	✓

Table 3. Performance of Classification Algorithms

Algorithm		Accuracy	Sensitivity	Specificity
NB	not using FS	95.95	95.96	95.95
	using FS	96.72	96.97	96.40
MLP	not using FS	96.33	95.96	97.75
	using FS	97.50	97.31	97.52
SVM	not using FS	96.15	95.96	96.40
	using FS	95.76	96.30	95.50
K-NN	not using FS	95.75	93.27	99.10
	using FS	97.69	96.30	99.55
SMO	not using FS	96.15	96.30	95.95
	using FS	97.30	97.31	97.30
REPTree	not using FS	94.61	93.94	95.50
	using FS	95.38	93.94	97.30
Random Forest	not using FS	90.37	97.98	80.18
	using FS	93.45	97.64	87.84
Bayesian Network	not using FS	95.57	95.29	95.95
	using FS	96.34	95.62	97.30

6. Analysis of the Results

As shown in Table 3, K-NN, MLP and SMO algorithms have the best accuracy. Besides, although the accuracy of these algorithms are the best, the other algorithms have achieved the accuracy above 93%. The chart in Figure 1 provide a comparison of accuracy of algorithms in two state using or not using of feature selection. From this chart it is obvious that using proposed feature selection method improved the accuracy of all the algorithms, except SVM. It is noting that although RF algorithm have achieved the lowest accuracy in comparison with the other algorithms, the 3% improvement of the accuracy of this algorithm after using feature selection method is notable.

The results also show that using the feature selection method have enhanced the sensitivity of algorithms, which means the tendency of algorithms to classify MI cases is increased. In terms of specificity measure, except for SVM, feature selection have led to higher results; it means that the trend of algorithms to predict healthy cases in increased. It is worth mentioning that using the feature selection method have caused a much improvement in specificity for RF algorithm, which is about 9%.

The results achieved in this study in comparison to the studied, reviewed in section 2, Show the better performance of proposed method. In compared to [14 -15] and [22 - 24], which have used SVM, despite of reduction of accuracy after using feature selection, our implemented model with SVM algorithm achieved higher accuracy.

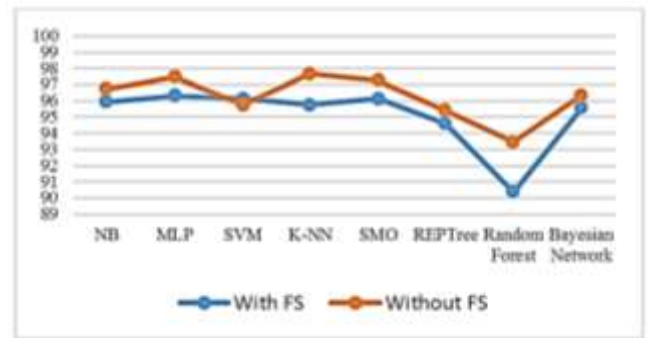


Fig. 1. Comparison of classification algorithms accuracy in two state: using or not using proposed feature selection

Moreover, our model using K-NN in compared to [21] and [23-24] have reached better accuracy, so that the accuracy of our model using K-NN is about 9% higher than the reviewed papers. The model in this study using MLP have reached better accuracy in comparison with the results of the models proposed in [13], [20 -22], so that the accuracy of our model using MLP is about 9% higher than the best accuracy which is resulted in [21].

Although Naïve Bayes algorithm's result in this study is higher than the accuracy in [21], it is lower than the accuracy resulted in [19]. Also the accuracy of Bayesian Network and REPTree in our study is lower than the accuracies achieved in [19]. This being low of accuracy in some algorithms in our study may be due to the difference between the number and types of features and cases used in studies which can affect the results.

7. Conclusion

In this study a new feature selection method is proposed which utilized Weight by SVM and Genetic Algorithm. Eight classification algorithms, include Naïve Bayes, Multi-layer perceptron, Support Vector Machine, K-Nearest Neighbors, Bayesian Network, SOM, Random forest and REPTree, are applied to a real Myocardial Infarction. Findings showed that using this feature selection method can lead to higher accuracy for classification algorithms in predicting Myocardial Infarction. After applying the proposed feature selection method, a subset of the features selected and the results showed increasing in accuracy, sensitivity and specificity in all algorithms, except SVM. Among the investigated algorithms, K-NN (K=4), MLP and SOM algorithms, had the best accuracies, 97.69%, 97.50% and 97.30% respectively. Overall, it can be concluded that applying the feature selection method, used in this study, along

with classification algorithms is almost a confident method for predicting Myocardial Infarction. High sensitivity and specificity, in addition to high accuracy of the algorithms, can be considered as a benefit of the proposed model. One weakness can be considered about work may be the large number of MI cases in compared to normal cases; because, in general, disease datasets are usually imbalanced and the number of disease cases are much less than normal cases. This problem could be effective on the results. In the future we are going to employ more classification algorithms using the proposed feature selection method by considering more normal cases than healthy cases. Besides, we want to make changes in our feature selection method using the other evolutionary method and weighting methods and compared the results. As a suggestion this method can be used for the other health problems, such as diseases prediction.

References

- [1] A.S. Go, D. Mozaffarian, V.L. Roger, E.J. Benjamin, J.D. Berry, W.B. Borden, D.M. Bravata, S. Dai, E.S. Ford, C.S. Fox, and S. Franco. "Heart Disease and Stroke Statistics--2013 Update: A Report From the American Heart Association". *Circulation*, Vol. 127, pp. e6-e245, 2012.
- [2] A. Ahmadi, H. Soori, Y. Mehrabi, K. Etemad, T. Samavat, and A. Khaledifar. "Incidence Of Acute Myocardial Infarction In Islamic Republic Of Iran: A Study Using National Registry Data In 2012". *Eastern Mediterranean health journal*, Vol. 21, pp. 5-12, 2015.
- [3] C. Wiener, C. Brown, A. Hemnes and T. Harrison. *Harrison's principles of internal medicine*. New York: McGraw-Hill Medical, 2012, pp. 455-456.
- [4] F. Mohammadi, A. Taherian, M. Hosseini and M. Rahgozar. "Effect of Home-Based Cardiac Rehabilitation on Quality of Life in the Patient with Myocardial Infarction". *Journal of Rehabilitation*, Vol. 7, pp. 11-19, 2006. [In Persian]
- [5] J. E. Hall. *Guyton and Hall Textbook of Medical Physiology*. New York: Saunders, 2015, pp. 264-266.
- [6] R. Dhingra, J. Shaw and L. A. Kirshenbaum. "molecular regulation of apoptosis signaling pathway in heart" in *Apoptosis: Modern Insights into Disease from Molecules to Man*, 1st ed., V. R. Preedy, Ed. Florida: CRC Press, 2010, pp. 382-385.
- [7] N. Esfandiari, M. Babavalian, A. Moghadam and V. Tabar. "Knowledge discovery in medicine: Current issue and future trend". *Expert Systems with Applications*, Vol. 41, pp. 4434-4463, 2014.
- [8] M. Jabbar, B. Deekshatulu and P. Chandra. "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm". *Procedia Technology*, Vol. 10, pp. 85-94, 2013.
- [9] S. Kumar and G. Sahoo. "Classification of Heart Disease Using Naïve Bayes and Genetic Algorithm". *Computational Intelligence in Data Mining*, Vol. 2, pp. 269-282, 2014.
- [10] U. Fayyad and R. Uthurusamy. "Data mining and knowledge discovery in databases". *Communications of the ACM*, Vol. 39, pp. 24-26, 1996.
- [11] P. Tan, M. Steinbach and V. Kumar. *Introduction to data mining*. Boston: Pearson Addison Wesley, 2005.
- [12] I. Benjamin, R. C. Griggs, E. J Wing and J. Fitz. *Andreoli and Carpenter's Cecil Essentials of Medicine*. New York: Saunders, 2015, pp. 93-102.
- [13] W. Baxt, F. Shofer, F. Sites and J. Hollander. "A neural computational aid to the diagnosis of acute Myocardial Infarction". *Annals of Emergency Medicine*, Vol. 39, pp. 366-373, 2002.
- [14] M. Qazi, G. Fung, S. Krishnan, J. Bi, R. Bharat Rao and A.S. Katz. "Automated heart abnormality detection using sparse linear classifiers". *Engineering in Medicine and Biology Magazine*, Vol. 26, pp. 56-63, 2007.
- [15] D. Conforti, D. Constanzo and R. Guido. "Medical decision making: A case study within the cardiology domain". *Journal on Information Technology in Healthcare*, Vol. 5, pp. 343-356, 2007.
- [16] S. Patil and Y. Kumaraswamy. "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network". *European Journal of Scientific Research*, Vol. 31, pp. 642-656, 2009.
- [17] M. Karaolis, J. Moutiris, L. Papaconstantinou and C. Pattichis. "Association rule analysis for the assessment of the risk of coronary heart events." in *Engineering in Medicine and Biology Society. EMBC 2009. Annual International Conference of the IEEE*, 2009, pp. 6238 - 6241.
- [18] M. Arif, I. Malagore and F. Afsar. "Automatic Detection and Localization of Myocardial Infarction Using Back Propagation Neural Networks", in *4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE)*, 2010, pp. 2151-7614.
- [19] H. Masethe and M. Masethe, "Prediction of Heart Disease using Classification Algorithm," in *Proceedings of the World Congress on Engineering and Computer Science*, San Francisco, USA, 2014.

- [20] N. Krishnaraj and R. Vinothkumar. "Heart Disease Prediction using GA and MLBPN". *International Journal of Applied Management & Business Utility*, Vol. 2, pp. 17-24, 2014.
- [21] N. Safdarian, N. Dabanloo and G. Attarodi. "A new pattern recognition method for detection and localization of Myocardial Infarction using t-wave integral and total integral as extracted features from one cycle of ECG signal". *Journal of Biomedical Science and Engineering*, Vol. 07, pp. 818-824, 2014.
- [22] N. Bhaskar. "Performance analysis of Support Vector Machine and Neural Networks in detection of Myocardial Infarction". *Procedia Computer Science*, Vol. 46, pp. 20-30, 2015.
- [23] L. Sharma, R. Tripathy and S. Dandapat. "Multiscale Energy and Eigenspace Approach to Detection and Localization of Myocardial Infarction". *IEEE Transactions on Biomedical Engineering*, Vol. 62, pp. 1827-1837, 2015.
- [24] P. Kora and S. Kalva. "Improved Bat algorithm for the detection of Myocardial Infarction". *SpringerPlus*, Vol. 4, pp. 1-18, 2015.
- [25] U. Fayyad and R. Uthurusamy. "Data mining and knowledge discovery in databases". *Communications of the ACM*, Vol. 39, pp. 24-26, 1996.
- [26] N. G. B. Amma. "Cardiovascular disease prediction system using genetic algorithm and Neural Network," in *International Conference on Computing, Communication and Applications (ICCCA)*, 2012, pp. 1-5.
- [27] Han J, Kamber M, Pei J. *Data mining: concepts and techniques*. Morgan Kaufmann, 2011.
- [28] M. Pacharne and V. Nayak. "Feature Selection Using Various Hybrid Algorithms for Speech Recognition," in *Computational Intelligence and Information Technology*, 1st ed., V. Das and N. Thankachan, Ed. Berlin: Springer Berlin Heidelberg, 2011, pp. 652-656.
- [29] S. Kalmegh. "Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News". *IJSET - International Journal of Innovative Science, Engineering & Technology*, Vol. 2, pp. 438-446, 2015.
- [30] M. Pal. "Random forest classifier for remote sensing classification". *International Journal of Remote Sensing*, Vol. 26, pp. 217-222, 2005.
- [31] Y. Wang and J. Vassileva, "Bayesian Network-based trust model." in *International Conference on Web Intelligence, IEEE/WIC*, 2003, pp.372-378.
- [32] R. Ganesh Kumar and Y. Kumaraswamy. "Performance Analysis Of Soft Computing Techniques For Classifying Cardiac Arrhythmia". *Indian Journal of Computer Science and Engineering (IJCSE)*, Vol. 4, pp. 459-465, 2014.
- [33] M. Gardner and S. Dorling. "Artificial Neural Networks (the multilayer perceptron)—a review of applications in the atmospheric sciences". *Atmospheric Environment*, Vol. 32, pp. 2627-2636, 1998.
- [34] K. Polat, S. Güneş and A. Arslan. "A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine". *Expert Systems with Applications*, Vol. 34, pp. 482-487, 2008.
- [35] S. Gunn. "Support Vector Machines for Classification and Regression". University of Southampton, Technical Report, 1998.
- [36] J. Platt. "Sequential minimal optimization: A fast algorithm for training Support Vector Machines". Microsoft Research, Technical report MSR-TR-98-141998.
- [37] D. Hand and K. Yu. "Idiot's Bayes: Not So Stupid after All?", *International Statistical Review/Revue Internationale de Statistique*, Vol. 69, p. 385, 2001.
- [38] R. Alizadehsani, J. Habibi, B. Bahadorian, H. Mashayekhi, A. Ghandeharioun, R. Boghrati and Z. Alizadeh Sani. "Diagnosis of Coronary Arteries Stenosis Using Data Mining". *Journal of Medical Signals and Sensors*, Vol. 2, pp. 153-159, 2012.
- [39] A. Onan. "A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer". *Expert Systems with Applications*, Vol. 42, pp. 6844-6852, 2015.
- [40] M. Heydari, M. Teimouri, Z. Heshmati and S. Alavinia, "Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran", *International Journal of Diabetes in Developing Countries*, 2015, pp.1-7.

Hodjatollah Hamidi born 1978, in shazand Arak, Iran, He got his Ph.D in Computer Engineering. His main research interest areas are Information Technology, Fault-Tolerant systems (fault-tolerant computing, error control in digital designs) and applications and reliable and secure distributed systems and E- Commerce. Since 2013 he has been a faculty member at the IT group of K. N. Toosi University of Technology, Tehran Iran. Information Technology Engineering Group, Department of Industrial Engineering, K. N. Toosi University of Technology.

Atefeh Daraei born in Khorram Abad, Lorestan Iran. She received her B.Sc in Information Technology in 2011 from University College of Nabi Akram, Tabriz, Iran. She is currently an M.Sc student in Information Technology (E-Commerce), at K. N. Toosi University of Technology, Tehran Iran. Her research interests include Machine learning, Knowledge Discovery and Data Mining and Customer Relationship Management.

Optimization of Random Phase Updating Technique for Effective Reduction in PAPR, Using Discrete Cosine Transform

Babak Haji Bagher Naeeni*
Department of Electrical Engineering, IRIB University, Tehran, Iran
bnaeeni@yahoo.com

Received: 01/Dec/2015

Revised: 12/Apr/2016

Accepted: 25/May/2016

Abstract

One of problems of OFDM systems, is the big value of peak to average power ratio. To reduce it, any attempt have been done amongst which, random phase updating is an important technique. In contrast to paper, since power variance is computable before IFFT block, the complexity of this method would be less than other phase injection methods which could be an important factor. Another interesting capability of random phase updating technique is the possibility of applying the variance of threshold power. The operation of phase injection is repeated till the power variance reaches threshold power variance. However, this may be a considered as a disadvantage for random phase updating technique. The reason is that reaching the mentioned threshold may lead to possible system delay. In this paper, in order to solve the mentioned problem, DCT transform is applied on subcarrier outputs before phase injection. This leads to reduce the number of required carriers for reaching the threshold value which results in reducing system delay accordingly.

Keywords: Orthogonal Frequency Division Multiplexing (OFDM); Peak-to-Average Power Ratio (PAPR); Random Phase Updating; Discrete Cosine Transform.

1. Introduction

If OFDM subcarriers would be summed in an inphase manner, a high signal peak would be produced in the time domain. As the signal variety range is an important factor in performance of telecommunication facilities (e.g. Transmitter/ Receiver), reduction in this range would lessen the system implementation expenses (1). High signal peaks in time domain can be troublesome in two aspects. One problem is reduction in analog to digital converters performance by expanding their input range and increasing the quantization noise. The other problem is a drop in performance level of RF amplifier in the transmitter side. In case of higher values of peak to average ratio of power in transmitted signal, trying to reduce the average values before the transmission process will affect the performance of amplifier and also raises the error probability in receiver; otherwise transmitted signal would be distorted because the peak signal power would be higher than the linear range of amplifier. Recently, PAPR overheads have been really considered as a problem, since the OFDM has had an important role in communication systems. In the last decade many solutions, such as Coding methods [3], Tone-reservation[4], Selective Mapping [5], Random phase injection algorithm [6] etc., has been proposed to solve the PAPR problem in OFDM systems. In this article we tended to improve the random phase updating technique (which is one of the phase injection methods) using discrete cosine transform as novel technique.

2. PAPR Problem and Random Phase Updating Technique

An important issue in multi carrier systems, especially OFDM systems, is the high ratio of peak to average power (PAPR), which is one of the inherent characteristics of the transmitted signal.

Generally, high peaks can be seen in transmitted signal when the phase of subcarriers signals would be summed constructively. PAPR is defined mathematically in the following review.

An OFDM signal can be written as follows [6]:

$$s(t) = \sum_{i=-\infty}^{+\infty} \sum_{m=0}^{M-1} b_m(i) e^{j2\pi(m/T)(t-iT)} p(t-iT) \quad (1)$$

Where T is the OFDM symbol duration $b_m(i)$ is the symbol of the mth sub channel at time interval iT. Which is ± 1 for BPSK modulation, $p(t)$ is a rectangular function with amplitude one and duration T, and M is the number of carriers. The OFDM signal of (1) in the time interval of $0 \leq t \leq T$ can be written as

$$s(t) = \sum_{m=0}^{M-1} b_m e^{j2\pi(m/T)t} \quad (2)$$

The power of $s(t)$ is:

$$p(t) = |s(t)|^2 = \sum_{m=0}^{M-1} \sum_{n=0}^{M-1} b_m b_n^* e^{j(2\pi(m-n)/T)t} \quad (3)$$

* Corresponding Author

The PAPR of the OFDM signal is written as:

$$PAPR = \frac{\text{Max}\{p(t)\}}{\text{Mean}\{p(t)\}} \quad (4)$$

The variation of the instantaneous power of OFDM signal from the average is:

$$\Delta p(t) = p(t) - E\{p(t)\} \quad (5)$$

And accordingly, the power variance (PV) of OFDM signal, denoted by ρ , can be written as [3]:

$$\rho = \frac{1}{T} \int_0^T (\Delta p(t))^2 dt = \sum_{i=1}^{M-1} |R_{bb}(i)|^2 \quad (6)$$

Where $R_{bb}(i)$ is the autocorrelation function of the sequence b_m .

$$R_{bb}(i) = \sum_{m=0}^{M-1-i} b_m b_{m+i}^* \quad (7)$$

The power variance ρ is a good measure of the PAPR. PV and PAPR are related to each other according to the following relationship:

$$Q\left(\frac{PAPR - 1}{\sqrt{\rho}}\right) + Q\left(\frac{1}{\sqrt{\rho}}\right) = \beta \quad (8)$$

Where β denotes the probability that $p(t)$ be less than or equal to p_{max} and

$$Q(y) = \frac{1}{\sqrt{2\pi}} \int_y^\infty e^{-u^2/2} du \quad (9)$$

From (7) it is seen that for a fixed β the OFDM signal with high PAPR has a high value of ρ . Because of the less computational burden in calculation of ρ , [see (5), (6)], in this method concentrate on the power (variance and assess its value for the random phase updating algorithm. However, using (7) the corresponding value of PAPR can also be obtained As shown in Fig. 1 in the random phase

updating algorithm for each carrier a random phase is generated and assigned to that carrier. Using (2) the OFDM signal with phasing is written as:

$$s(t) = \sum_{m=0}^{M-1} b_m e^{j2\pi((m/T)t + \phi_m)} \quad (10)$$

Where $2\pi\phi_m$ is the m th subcarrier phase shift. Adding random phases to each subcarrier will change the power variance of OFDM signal. In the random phase updating algorithm, the phase of each subcarrier is updated by a random increment as:

$$(\phi_m)_i = (\phi_m)_{i-1} + (\Delta\phi_m)_i \quad m = 0, 1, \dots, M - 1 \quad (11)$$

Where i is the iteration index and $(\Delta\phi_m)_i$ is the phase increment of m th subcarrier at i th iteration. In the random phase updating method, without loss of generality, the initial phase, i.e., $(\Delta\phi_m)_0$, can be considered zero. Consequently, a random phase increment is generated and the phase is updated by adding the increment to the phase of that subcarrier. Flow chart of the algorithm for this iterative phase updating is shown in Fig. 2. In Fig. 2(A) a certain threshold for PV is set and for Fig.2 (B) a limited number of iterations is allowed:

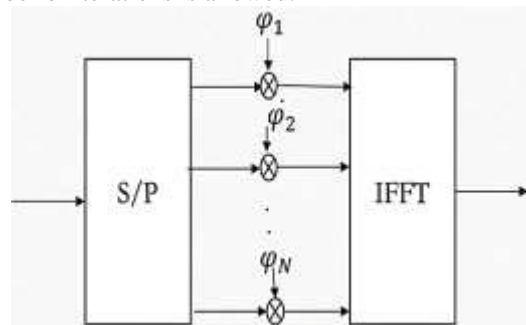


Fig. 1. Block diagram of OFDM with phasing showing the principle of adding phase shifts to the OFDM symbols.

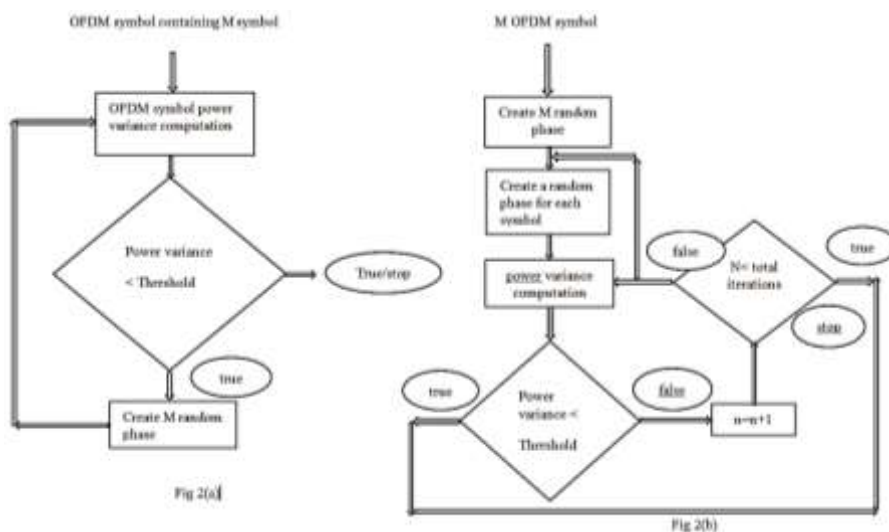


Fig. 2. generating and injecting random phase for OFDM system

In this method different distributions for the random phase increments have been considered and their influence on the PV has been investigated. Two distributions are Gaussian ($\Delta\phi_m = N(0, x^2)$) and Uniform ($\Delta\phi_m = \text{Unif}[0, x]$), where $x \in \{0, 0.25, 0.5, 0.75\}$. Results are shown in Table I.

Table I. power variance and number of iterations for the random phase updating algorithm with uniform and Gaussian distributions of phase increments

	Uniform Distribution				Normal Distribution			
	Power Variance		No. of iteration		Power variance		No. of iteration	
x	Mean	Std. dv	Mean	Std. dv	Mean	Std. dv	Mean	Std. dv
0.1	71.89	9.02	33.38	42.5	72.12	8.97	39.05	50
0.25	68.85	9.59	58.56	10.6	69.08	9.57	9.54	11.8
0.5	67.57	9.99	4.93	6.16	67.64	9.92	5.15	6.39
0.75	67.47	10	4.62	5.75	67.45	10	4.67	5.81
1.0	67.41	10	4.64	5.8	67.36	10	4.64	5.76

It is seen that there is no significant difference in the PV results when Gaussian or Uniform distribution is considered for the phase increments. In the rest of the method the uniform distribution has been chosen for the distribution of phase increments. The influence of different variances of the phase increments on the reduction of OFDM signal has been investigated. Results indicate a connection between phase shift variance and the PV number of iterations required reaching the threshold. Simulations have been carried out for different number of carriers as well as different PV thresholds. As shown in Fig. 3 when variance of phase shift increments is small more number of iterations is required.

This can be clearly justified. When standard deviation of phase increments is small the generated phases are likely not good to reduce the PAPR. But when the standard deviation of phase increments is large, the random phase increments have larger variations and it is more likely that their values be proper to decrease the PV. As seen in Fig. 4 by increasing the standard deviation of phase increments the number of iterations to reach the threshold decreases. Meanwhile, the lower the PV threshold the more the number of iterations. That is quite clear since lower threshold or smaller PV needs more iterations to select the proper phases for the subcarriers. From Fig. 3 the influence of different number of carriers on the number of iterations for different variances of phase shifts is also clear. It is obvious that increasing number of carriers from 8 to 48 slightly changes the number of iterations of the algorithm. As shown in Fig. 4, and unlike the number of carriers, the threshold level has a significant effect on the number of iterations of the algorithm. Efficiency of the algorithm is mainly related to the selected threshold level and consequently number of iterations and

not the number of carriers. This is why in Section IV the dynamic reduction of threshold is proposed.

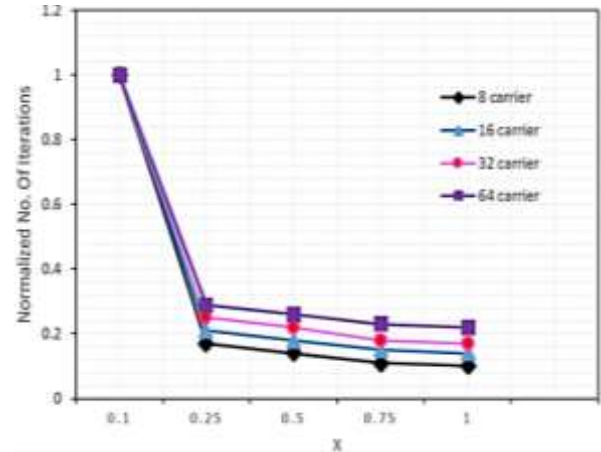


Fig. 3. Normalized mean number of iterations versus phase shift variance parameters α ; for $M = \{8, 16, 32, 48\}$ BPSK OFDM signal simulated with random phase updating algorithm (Fig. 2(A)).

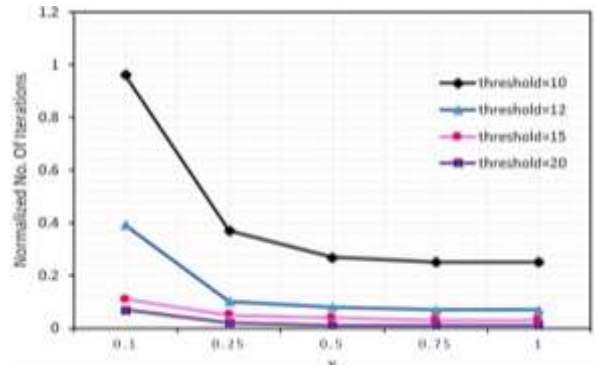


Fig. 4. Normalized mean number of iterations versus phase shift variance parameter α ; for 8-carrier OFDM system and different threshold levels, simulated with random phase updating algorithm of Fig. 2(A).

Reducing of the PAPR with phasing implies a high degree of complexity and side information. For large number of carriers the computational burden for the calculation of PV is increased [see (5)]. Besides, because of more carriers more phases are involved in the algorithm which leads to more side information. The phase shifts have to be known at the transmitter and receiver. To lessen the problem, the quantization and grouping of the random phase increments has been carried out. Quantization of the phase shift to BPSK (i.e., 2) or QPSK (i.e., 4) type phase shifts (i.e., $\Delta\phi_m \in \{0, 0.5\}$ or $\{0, 0.25, 0.5, 0.75\}$, respectively) each phase shift which leads to a reduced complexity of the algorithm. Grouping means subcarriers are bundled and all subcarriers in the same bundle (group) get the same phase shift increment (see Fig. 5). By grouping the complexity of the algorithm is further reduced. Simulations were carried out for a 16-carrier OFDM for two and four levels of phase quantization and different number of iterations (Fig. 6). Results shown in Fig. 6 indicate that rounding of the phase increments to two levels does not change the variance and reduces the mean of PV. Grouping for 16 carrier BPSK-OFDM was examined with

2 groups of 8 carriers, 4 groups of 4 carriers and 8 groups of 2 carriers and for different

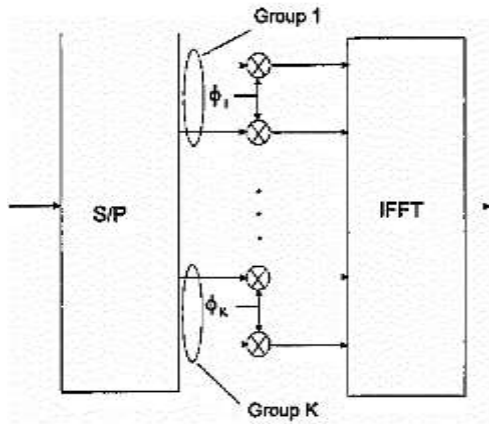


Fig. 5. Block diagram illustrating the grouping of the phases.

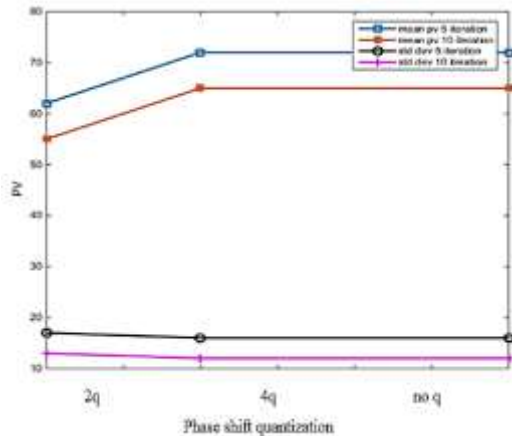


Fig. 6. Mean and standard deviation of power variance versus the phase quantization level for different number of iterations. OFDM with 16 carriers and BPSK modulation.

3. Discrete Cosine Transform

Discrete cosine transform is a reversible transform [7]. Input of this transform should be a vector and the process returns it's discrete cosine Fourier transform as an output. This way autocorrelation of the input series would be dropped so that the PAPR would be reduced [8].

Discrete cosine transform is defined as equation (12):

$$y(k) = w(k) \sum_{n=1}^N x(n) \cos\left(\frac{\pi(n-1)(k-1)}{2N}\right), \quad (12)$$

$$k = 1, 2, \dots, N$$

Which in equation:

$$w(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k = 1 \\ \sqrt{\frac{2}{N}}, & 2 \leq k \leq N \end{cases}$$

Since the discrete cosine transform reduces the inter-symbol correlation, it can also reduce the PAPR. This transform is really interested because of it produces low complexity and also it has no destructive effects on BER [9].

4. Combination of Discrete Cosine Transform and Updating Random Phase Techniques

One of the important advantages of random phase updating technique, is the relation between OFDM symbol power variance and PAPR. The complexity of this technique is less than that of most phase injection methods, because opposed to PAPR method, the power variance is computable before IFFT block. Another capability of random phase updating technique, is applying threshold power variance. In this method, phase injection would be continued until the time that power variance reaches to its threshold. However, it is possible that the power would not reach to its threshold value in a pre-determined time. This could be a disadvantage for random phase updating technique.

Therefore another technique is required as well as random phase updating for reducing both PAPR and complexity, currently. It is worth mentioning that in the proposed system, DCT of subcarriers is calculated before phase injection. This results in reducing symbols correlation which leads to power spectral density reduction. After phase injection, random phase updating is applied to new subcarriers. This leads to more reduction of PAPR compared to random phase updating. Also, the number of iterations for reaching threshold power variance would be reduced. The block finally the BER would be reduced.

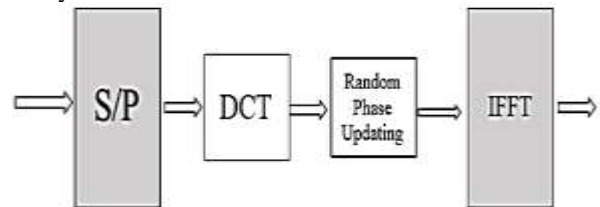


Fig. 7. The block diagram of combining random phase updating and DCT

5. Simulation Results

In the simulation results, it will be shown that the number of iterations required for reaching threshold power variance, PAPR value and bit error rate parameter, would be improved using the proposed combination technique.

5.1 Reducing the Number of Iterations Required for Reaching Threshold Power Variance

As discussed in section 4, reaching a desired threshold power variance requires to additional delay in random phase updating technique. This is one of the most important disadvantages of the mentioned technique. In order to, solve this problem, simultaneous using both DCT pre-coder and random phase updating is proposed in this paper. The proposed scheme is simulated using MATALAB. As it is shown in pictures (8 and 9), applying discrete cosine transform besides updating random phase technique, causes a considerable reduction in the numbers of repetitions required to reach the power variance (rather than updating random phase technique itself).

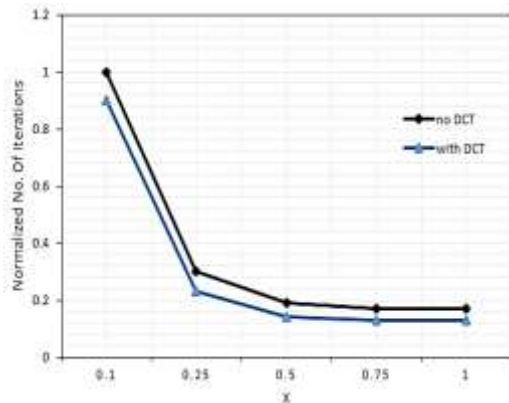


Fig. 8. Effect of applying DCT in numbers of repetitions required to reach the power variance in an 8 subcarrier OFDM system.

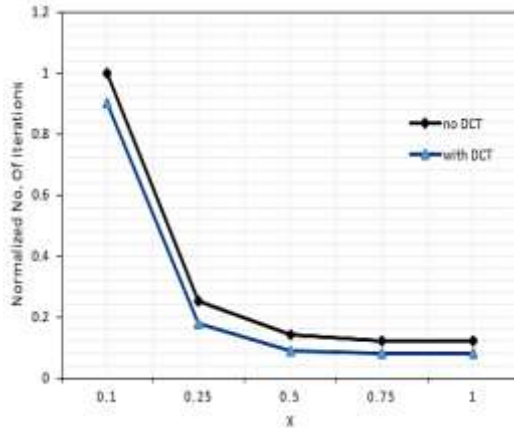


Fig. 9. Effect of applying DCT in numbers of repetitions required to reach the power variance in a 32-subcarrier OFDM system.

As shown in pictures (10-13), discrete cosine transform and updating random phase technique together, reduces the numbers of repetitions required to reach the power variance. Simulation has been done on a 16-carrier OFDM system using BPSK modulation for a 10 dB power variance (Picture (10)); a 12 dB power variance (Picture (11)); a 15 dB power variance (Picture (12)) and a 22 dB power variance (Picture (13)).

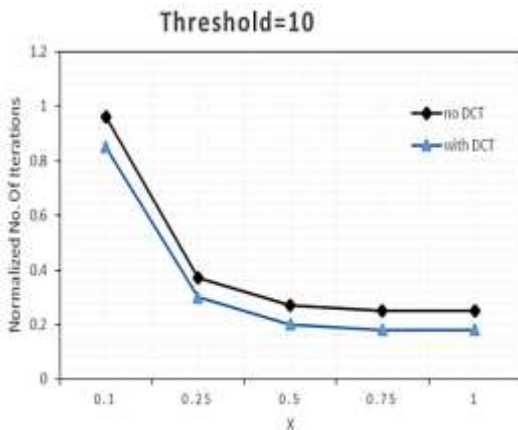


Fig. 10. Effect of applying DCT besides updating random phase technique in numbers of repetitions required to reach the power variance of 10 dB in a 16-subcarrier OFDM system using BPSK modulation.

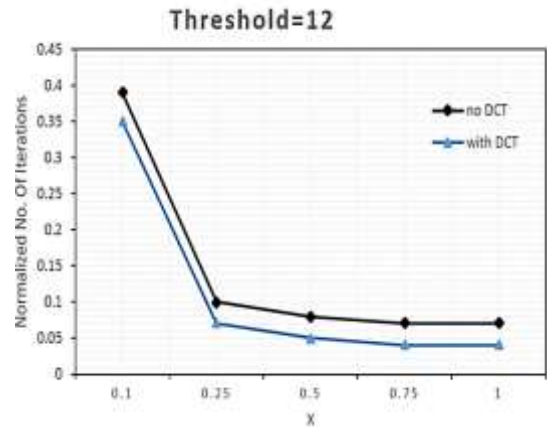


Fig. 11. Effect of applying DCT besides updating random phase technique in numbers of repetitions required to reach the power variance of 12 dB in a 16-subcarrier OFDM system using BPSK modulation.

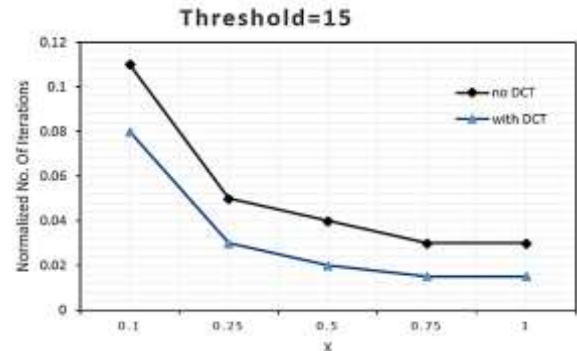


Fig. 12. Effect of applying DCT besides updating random phase technique in numbers of repetitions required to reach the power variance of 12 dB in a 16-subcarrier OFDM system using BPSK modulation.

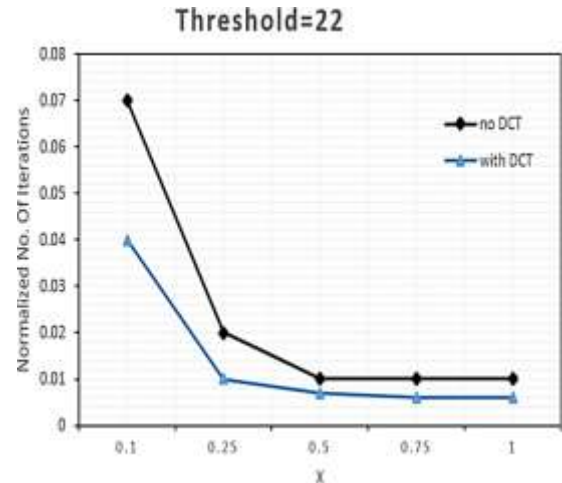


Fig. 13. Effect of applying DCT besides updating random phase technique in numbers of repetitions required to reach the power variance of 22 dB in a 16-subcarrier OFDM system using BPSK modulation.

5.2 Reducing PAPR Value

Based on (7), power variance value has a forward correlation with PAPR. Therefore, using the proposed combination scheme, PAPR would be reduced with a value of 1.8 db more than that of random phase updating method. This, can be verified using CCDF diagram.

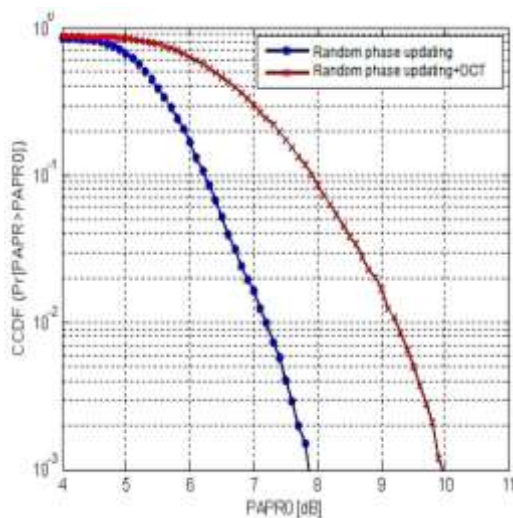


Fig. 14. Comparison between reducing PAPAR in two methods.

5.3 Improving BER

In order to send a signal through a wireless communication channel, it is required that the transmitting signal be amplified at the sender side. It should be mentioned that the power amplifiers often have a limited range. Therefore, the signal power could not be increased with a desired value. As a result, BER at the receiver side would be increased due to signal attenuation. However, since using the proposed combination scheme PAPR would be decreased, BER would be decreased too. Based on Fig.15, BER of the proposed scheme decreased with a value of 2.1 db compared to random phase updating technique.

6. Conclusion

In the proposed scheme, the DCT of subcarriers should be calculated before phase injection. Then, random phase updating should be applied to new sub-carriers. The result of this modification is as follows a. reduction of the numbers of iterations required to reach the threshold power variance, b. reducing PAPR with a value of 1.8 db, compared to random phase updating technique. And c. improvement of BER.

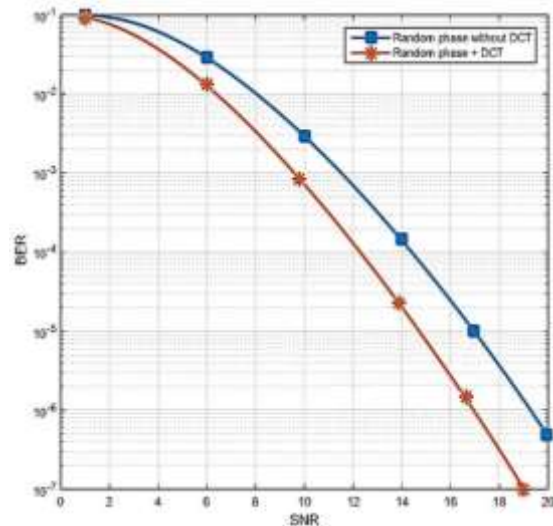


Fig. 15. BER reduction in an OFDM system using the proposed combination scheme.

References

- [1] V. K. Singh, A. Goel, A. Sharma. "Reducing Peak to Average Power Ratio of OFDM by Using Selective Mapping". *International Journal of Research in Information Technology*, vol. 2, No.4, pp. 400-407, April 2014.
- [2] T. Jiang and G. Zhu, "Complement block coding for reduction in peak to average power ratio of OFDM signals." *IEEE Communications Magazine*, vol. 43, no. 9, pp. s17-s22, Sept. 2005.
- [3] Md. Ibbrahim Abdullah et.al, "Comparative Study of PAPR Reduction Techniques in OFDM". *ARNP journal of system and software*, vol. 1, no. 8, pp. 263-269, nov 2011.
- [4] P. Phoomchusak, C. Pirak, "Adaptive tone-reservation PAPR technique with optimal subcarriers allocation awareness for multi-user OFDMA systems," in *Proc. ICACT*, 2011, pp.814-818.
- [5] H. Chen, H. Liang. "Combined Selective Mapping and Binary Cyclic Codes for PAPR Reduction in OFDM Systems." *IEEE Transactions on Wireless Communications*, vol.6, pp.3524-3528, Oct. 2007.
- [6] Nikookar, H., and Lidsheim, K.S., "Random phase updating algorithm for OFDM transmission with low PAPR." *IEEE Transaction on Broadcasting*, Vol.48, Jun 2002.
- [7] H. Ochiai and H. Imai, "Performance of the deliberate clipping with symbol selection for strictly band-limited OFDM systems." *IEEE J. Sel. Areas Commun.*, vol. 18, pp.123-128, Nov. 2000.
- [8] Jayalath, A., Tellambura, C., "Adaptive PTS approach for reduction of peak-to-average power ratio of OFDM signal." *Electronics Letters*, Vol.36, pp. 1226-1228, 2000.
- [9] S. Gazor, R. AliHemmati, "Tone reservation for OFDM systems by maximizing signal-to-distortion ratio." *IEEE Transactions on Wireless Communications*, Vol.11, pp. 762-770, 2012.

Babak Haji Bagher Naeeni received the B.Sc and M.Sc in electrical engineering from central Tehran branch, Islamic Azad university, Tehran, Iran, and from south Tehran branch, Islamic Azad university, Tehran, Iran, in 1993 and 1999, respectively. He is assistant professor of electrical engineering at IRIB university. His research interests include multiuser detection, channel equalization, artificial intelligence.

Nonlinear State Estimation Using Hybrid Robust Cubature Kalman Filter

Behrouz Safarinejadian*

Department of Electrical and Electronic Engineering, Shiraz University of Technology, Shiraz, Iran
safarinejad@sutech.ac.ir

Mohsen Taher

Department of Electrical and Electronic Engineering, Shiraz University of Technology, Shiraz, Iran
m.taher@sutech.ac.ir

Received: 26/Feb/2015

Revised: 16/Mar/2015

Accepted: 09/May/2016

Abstract

In this paper, a novel filter is provided that estimates the states of any nonlinear system, both in the presence and absence of uncertainty with high accuracy. It is well understood that a robust filter design is a compromise between the robustness and the estimation accuracy. In fact, a robust filter is designed to obtain an accurate and suitable performance in presence of modelling errors. So in the absence of any unknown or time-varying uncertainties, the robust filter does not provide the desired performance. The new method provided in this paper, which is named hybrid robust cubature Kalman filter (CKF), is constructed by combining a traditional CKF and a novel robust CKF. The novel robust CKF is designed by merging a traditional CKF with an uncertainty estimator so that it can provide the desired performance in the presence of uncertainty. Since the presence of uncertainty results in a large innovation value, the hybrid robust CKF adapts itself according to the value of the normalized innovation. The CKF and robust CKF filters are run in parallel and at any time, a suitable decision is taken to choose the estimated state of either the CKF or the robust CKF as the final state estimation. To validate the performance of the proposed filters, two examples are given that demonstrate their promising performance.

Keywords: Uncertainty; State Estimation; Cubature Kalman Filter (CKF); Robust CKF; Hybrid Robust CKF.

1. Introduction

One of the essential problems in control theory and signal processing is the problem of dynamic state estimation. Mostly, the extended Kalman filter (EKF) is used to estimate the states of nonlinear systems [1,2]. In EKF, it is necessary to calculate Hessian and Jacobian matrices at each iteration. Therefore, this filter has linearization error and is not suitable for highly nonlinear functions. Cubature Kalman filter (CKF) has been presented to dominate the limitations of EKF [3-5]. Moreover, the CKF has a higher approximation accuracy compared to the EKF and has an easier algorithm to implement, because it avoids weighty calculation of the Jacobian and Hessian matrices. Because of these advantages, CKF has been used in many applications [6-8]. When the process and measurement models of a system are known, we will have the best performance for each estimator (EKF or CKF). Nevertheless, in many physical systems, the obtained model is an approximate with parametric uncertainty and unknown external input. Furthermore, the process and measurement noises may be colored and biased instead of being zero mean and white. Recently, uncertainties in the plant model are considered and several robust filtering methods have been provided, including robust Kalman filter [9,10], robust minimum variance filters [11,12], smooth variable structure filter [13], risk-sensitive filter [14] and so on. These filters are different in terms of the uncertainty present in the plant.

Unfortunately, finding an effective and accurate model for the plant in the presence of uncertainty is often difficult.

The present work solves this problem and acts in such a way that does not require any specific information about the plant uncertainty. The proposed modification in CKF is capable of preserving filter performance in the presence of uncertainty and unknown disturbance in the plant description. In other words, a new robust nonlinear filter will be proposed. In addition, these uncertainties do not have an upper limit (like a bounded norm).

It is well understood that a robust filter design is a compromise between estimation accuracy and robustness. A robust filter is designed to obtain an accurate and suitable performance considering the errors of modelling [15-17]. Thus, in the absence of any unknown or time-varying uncertainties, the robust filter does not provide the desired performance. This issue has caused that in this paper, a hybrid robust CKF is proposed so that detects and adjusts itself with respect to the uncertainty of the system. The hybrid robust CKF consists of both filters: a CKF and a robust CKF that run in parallel. At each time step, choosing one of these two types of filters to estimate the final state is related to the uncertainty and based on the amount of the normalized innovation corresponding to the two filters (CKF and robust CKF). Therefore, it is expected that the hybrid robust CKF has a desirable performance even in the absence of uncertainty.

This paper is configured as follows. In Section 2, the problem formulation is briefly described. The estimation

* Corresponding Author

of uncertainty using a low pass filter is investigated in Section 3. Section 4 and Section 5 include the proposed robust CKF and hybrid robust CKF, respectively. In Section 6, simulation results and comparison of the proposed algorithms with CKF algorithm are given. Finally, in Section 7, conclusion is presented.

2. Problem Formulation

Consider the discrete-time nonlinear stochastic control system.

$$x_k = f(x_{k-1}, u_{k-1}) + w_{k-1} \quad (1)$$

$$z_k = h(x_{k-1}) + v_k \quad (2)$$

Where $x_k \in R^n$ is the stochastic state vector, $u_k \in R^p$ is a known control input and $z_k \in R^m$ is the measurement vector, w_{k-1} and v_k are zero mean-white Gaussian noises with covariance matrices Q_{k-1} and R_k , respectively. The process noise w_{k-1} denotes any kind of uncertainty which disturbs the system. Moreover, w_{k-1} and v_k are assumed to be uncorrelated. The problem is to estimate the system states using the measurements z_k .

3. Estimation of Uncertainty Using a Low Pass Filter

As said before, the process noise w_{k-1} represents any kind of uncertainty which disturbs the nominal system $x_k = f(x_{k-1}, u_{k-1})$. In the state transition equation (1), a quantity equal to the uncertainty w_{k-1} is expressed as

$$w_{eq,k-1} = x_k - f(x_{k-1}, u_{k-1}) \quad (3)$$

Then, an estimate of this uncertainty at the existing step is written as follows

$$\hat{w}_k = \hat{w}_{eq,k-1}(\hat{x}_k) = \hat{x}_k - f(\hat{x}_{k-1}, u_{k-1}) \quad (4)$$

To estimate the uncertainty, a low pass filter is used as [18].

$$\hat{w}_k = \prod \hat{w}_{eq,k-1}(\hat{x}_k) = \prod (\hat{x}_k - f(\hat{x}_{k-1}, u_{k-1})) \quad (5)$$

where \prod is a diagonal matrix with low pass filter entries. While a lot of candidates are possible in the selection of the \prod matrix; here, the following first-order discrete filters have been applied.

$$k_i(z) = \frac{b_i}{1 - a_i z^{-1}} \quad i = 1, 2, \dots, n \quad (6)$$

Therefore, the matrix \prod will take the form

$$\prod(z) = \text{diag}(k_1, k_2, \dots, k_n) \quad (7)$$

By expanding (7), the state-space realization of (5) is written as follows

$$\hat{w}_k = A \hat{w}_{k-1} + B(\hat{x}_k - f(\hat{x}_{k-1}, u_{k-1})) \quad (8)$$

where, $A = \text{diag}(a_1, \dots, a_n)$ and $B = \text{diag}(b_1, \dots, b_n)$. The discrete LPFs in (6) should have a unity DC gain, that is $k_i(1) = 1 (i = 1, \dots, n)$. Thus, $a_i + b_i = 1$ and naturally $A + B = I$.

4. Robust Cubature Kalman Filter

In this section, the cubature Kalman filter is introduced briefly and then it is extended to develop a new robust cubature Kalman filter.

4.1 Cubature Kalman Filter

All of known filters have problems such as divergence in high-dimensions. Dimension problem is related to state space and is observed in higher dimension state space models. In recent years, the cubature Kalman filter was proposed in order to solve divergence and dimension problems [3]. This filter is a nonlinear method to estimate the system states with an upper dimension limit based on the spherical-radial cubature rule [4]. There is no need to make a derivative in the cubature rule. Thus, it is not necessary to compute the Jacobian and Hessian matrices. CKF algorithm is demonstrated in [3,4] completely.

4.2 Robust Cubature Kalman Filter

The robust CKF is a combination of the traditional CKF with a new uncertainty estimator. This filter is designed to estimate the states of any nonlinear system in presence of various types of uncertainties including, the parameter uncertainty or the unknown input. This filtering algorithm is similar to the traditional CKF, except that the estimated uncertainty \hat{w}_{k-1} and an uncertainty estimation error covariance $P_{w,k-1}$ are included in the algorithm. The robust CKF consists of the following steps.

1. Initialize ($\hat{x}_{k-1|k-1}$ and $p_{k-1|k-1}$)
2. Initial density at time $k-1$ can be decomposed as follows

$$P_{k-1|k-1} = S_{k-1|k-1} S_{k-1|k-1}^T \quad (9)$$

3. Obtain the cubature points

$$X_{i,k-1|k-1} = S_{k-1|k-1} \xi_i + \hat{x}_{k-1|k-1} \quad \begin{matrix} (i = 1, 2, \dots, m) \\ m = 2n \end{matrix} \quad (10)$$

4. In this step, the cubature points pass through the nonlinear function $f(\cdot)$.

$$X_{i,k|k-1} = f(X_{i,k-1|k-1}, u_{k-1}) + \hat{w}_{k-1} \quad (11)$$

5. The predicted mean is computed at the time update

$$\hat{x}_{k|k-1} = \frac{1}{m} \sum_{i=1}^m X_{i,k|k-1} \quad (12)$$

6. The predicted covariance is computed at the time update

$$P_{k|k-1} = \frac{1}{m} \sum_{i=1}^m X_{i,k|k-1} X_{i,k|k-1}^T - \hat{x}_{k|k-1} \hat{x}_{k|k-1}^T + P_{w,k} \quad (13)$$

7. The covariance matrix achieved from the previous step is decomposed again as

$$P_{k|k-1} = S_{k|k-1} S_{k|k-1}^T \quad (14)$$

8. Obtain the new cubature points

$$X_{i,k|k-1} = S_{k|k-1} \xi_i + \hat{x}_{k|k-1} \quad (15)$$

9. The cubature points pass through the nonlinear function $h(\cdot)$.

$$Y_{i,k|k-1} = h(X_{i,k|k-1}, u_{k-1}) \quad (16)$$

10. The predicted observation is computed at the measurement update

$$\hat{y}_{k|k-1} = \frac{1}{m} \sum_{i=1}^m Y_{i,k|k-1} \quad (17)$$

11. The predicted innovation covariance is calculated at the measurement update

$$P_{yy,k|k-1} = \frac{1}{m} \sum_{i=1}^m Y_{i,k|k-1} Y_{i,k|k-1}^T - \hat{y}_{k|k-1} \hat{y}_{k|k-1}^T + R_k \quad (18)$$

12. The predicted cross covariance is

$$P_{xy,k|k-1} = \frac{1}{m} \sum_{i=1}^m X_{i,k|k-1} Y_{i,k|k-1}^T - \hat{x}_{k|k-1} \hat{y}_{k|k-1}^T \quad (19)$$

13. The CKF gain is

$$K_k = P_{xy,k|k-1} P_{yy,k|k-1}^{-1} \quad (20)$$

14. Compute the updated state and covariance

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k (y_k - \hat{y}_{k|k-1}) \quad (21)$$

$$P_{k|k} = P_{k|k-1} - K_k P_{yy,k|k-1} K_k^T \quad (22)$$

15. Update the uncertainty estimation

$$\hat{w}_k = A \hat{w}_{k-1} + B (\hat{x}_k - f(\hat{x}_{k-1}, u_{k-1})) \quad (23)$$

16. Update the uncertainty estimation error covariance (derived in lemma 1)

$$P_{w,k} = P_{w,k-1} + (BK_k) R (BK_k)^T + Q_{k-1} \quad (24)$$

Block diagram of the proposed robust CKF is shown in Figure 1.

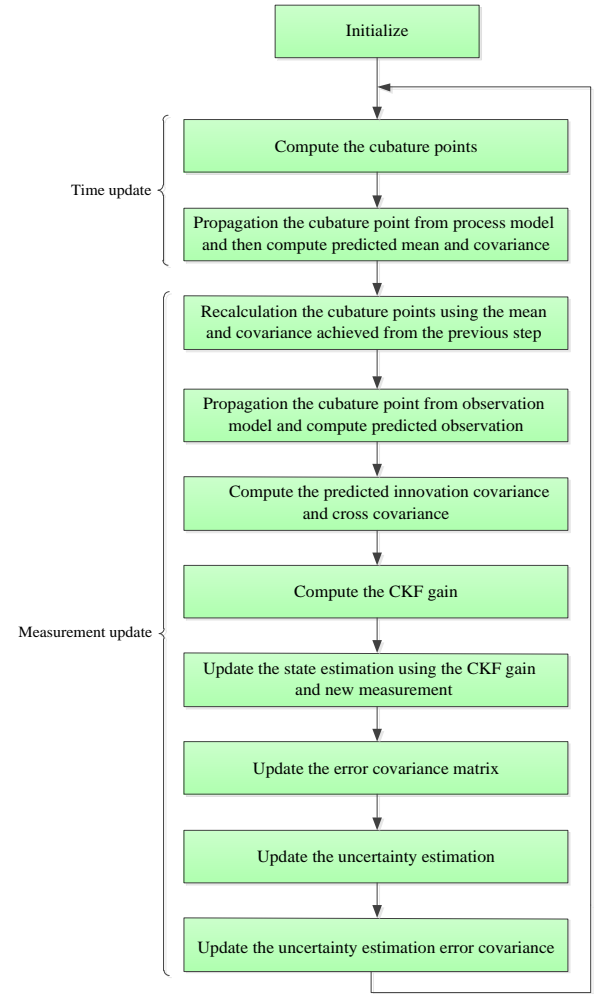


Fig. 1. Block diagram of the proposed robust CKF.

Lemma 1. The uncertainty estimation error covariance can be written as follows

$$P_{w,k} = P_{w,k-1} + (BK_k) R (BK_k)^T + Q_{k-1} \quad (25)$$

Proof: We have

$$\hat{x}_{k|k-1} = f(\hat{x}_{k-1|k-1}, u_{k-1}) + \hat{w}_{k-1} \quad (26)$$

Using this equation, the estimated uncertainty in (8) can be obtained as follows

$$\begin{aligned} \hat{w}_k &= A \hat{w}_{k-1} + B (\hat{x}_k - f(\hat{x}_{k-1}, u_{k-1})) = \\ &= A \hat{w}_{k-1} + B (\hat{x}_k - \hat{x}_{k|k-1}) + B \hat{w}_{k-1} = \hat{w}_{k-1} + B (\hat{x}_k - \hat{x}_{k|k-1}) \end{aligned} \quad (27)$$

Since $A + B = I$.

From equation (21) we have

$$\hat{x}_{k|k} - \hat{x}_{k|k-1} = K_k (y_k - \hat{y}_{k|k-1}) \quad (28)$$

Now, using (27) and (28) we obtain

$$\hat{w}_k = \hat{w}_{k-1} + B(\hat{x}_k - \hat{x}_{k|k-1}) = \hat{w}_{k-1} + BK_k(y_k - \hat{y}_{k|k-1}) \quad (29)$$

Also, we can write

$$w_k = w_{k-1} + w_k - w_{k-1} = w_{k-1} + \Delta w_k \quad (30)$$

where, Δw_k is the difference in uncertainty between the $k - 1$ th and k th instant and is assumed to be white Gaussian with zero mean and covariance Q_{k-1} .

Subtracting (29) from (30), gives

$$(w_k - \hat{w}_k) = (w_{k-1} - \hat{w}_{k-1}) - BK_k(y_k - \hat{y}_{k|k-1}) + \Delta w_k \quad (31)$$

$$\tilde{w}_k = \tilde{w}_{k-1} - BK_k \tilde{y}_k + \Delta w_k \quad (32)$$

where, $(y_k - \hat{y}_{k|k-1})$ can be approximated by the measurement noise v_k . Noises \tilde{w}_k , v_k and Δw_k are uncorrelated and the uncertainty error covariance is as follows.

$$E[\tilde{w}_k \tilde{w}_k^T] = P_{w,k} \quad (33)$$

Finally, using equations (32) and (33), the uncertainty estimation error covariance, $p_{w,k}$ can be derived as

$$P_{w,k} = P_{w,k-1} + (BK_k)R(BK_k)^T + Q_{k-1} \quad (34)$$

5. Hybrid Robust Cubature Kalman Filter

A robust filter will be designed in this section to keep a balance between uncertainty and estimation error; therefore, it has a desired performance in the presence or absence of any unknown input or uncertainty. In fact, the CKF and robust CKF run in parallel in the proposed method. At any instant, a proper decision is taken to choose either the estimated state of the CKF or the robust CKF as the final state estimation.

The presence of uncertainty shows itself as a big innovation. A simple assessment method for this problem is based on the normalized innovation which is given as

$$\varepsilon_k = \lambda_k^T \delta_k \lambda_k \quad (35)$$

where, λ_k is the innovation and δ_k is the innovation covariance. At each time step, ε_k is computed by both of the filters. Finally, the filter is selected as the final state estimator that has less innovation. Instead of a statistical decision making, a moving average of the normalized innovations is considered as follows

$$\varepsilon_k^s = \sum_{i=k-s+1}^k \varepsilon_k(i) \quad (36)$$

where s is the moving window's width. Suppose at time k , the CKF state estimation is $x_{ckf}(k|k)$ and covariance matrix is $p_{ckf}(k|k)$, the robust CKF state

estimation is $x_{Rckf}(k|k)$ and covariance matrix is $p_{Rckf}(k|k)$ and the corresponding amounts for the hybrid robust CKF are $x_{HyRckf}(k|k)$ and $p_{HyRckf}(k|k)$, respectively. The output of robust CKF is one of the two types of the CKF or the robust CKF. However, as stated above, the output of this filter is the output of the filter (CKF or robust CKF) that has less innovation. The innovation of the CKF and the robust CKF are demonstrated by $\varepsilon_{k,ckf}^s$ and $\varepsilon_{k,Rckf}^s$, respectively. In this type of filter, decision-making is performed as follows

$$[x_{ckf}(k|k), P_{ckf}(k|k)] = CKF[x_{ckf}(k-1|k-1), P_{ckf}(k-1|k-1)]$$

$$[x_{Rckf}(k|k), P_{Rckf}(k|k)] = ROBUSTCKF[x_{Rckf}(k-1|k-1), P_{Rckf}(k-1|k-1)]$$

if $\varepsilon_{k,ckf}^s > \gamma \varepsilon_{k,Rckf}^s$

$$[x_{HyRckf}(k|k), P_{HyRckf}(k|k)] = [x_{Rckf}(k|k), P_{Rckf}(k|k)]$$

else

$$[x_{HyRckf}(k|k), P_{HyRckf}(k|k)] = [x_{ckf}(k|k), P_{ckf}(k|k)]$$

end

where, γ is a scalable parameter that depends on the permissible amount of uncertainty in the system.

Block diagram of the proposed hybrid robust CKF is shown in Figure 2.

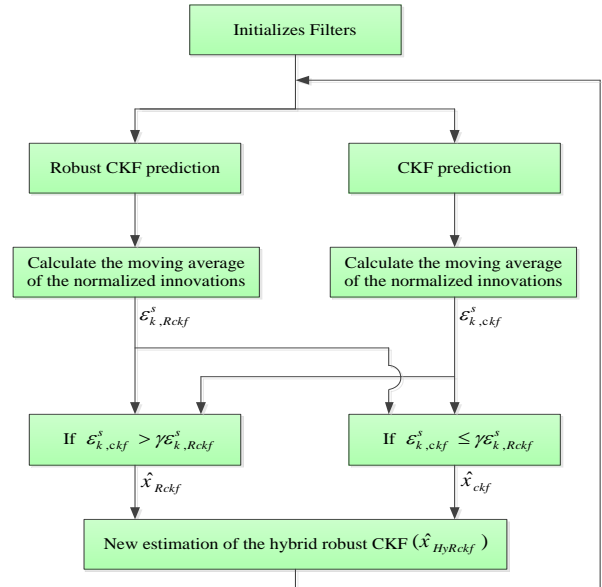


Fig. 2. Block diagram of the proposed hybrid robust CKF.

6. Simulation Results

In this section, two examples are given to show the performance of the proposed filters, in comparison with the traditional CKF. The first example is related to a ballistic target motion model with unknown ballistic coefficient and aerodynamic forces adopted from [19]. The second example is related to the Euler-discretized van der Pol oscillator that is adopted from [20].

Example 1) The ballistic target motion model with unknown ballistic coefficient is given by [19].

$$s_{k+1} = \Phi s_k + Gf(s_k) + G \begin{bmatrix} 0 \\ -g \end{bmatrix} + w_k \quad (37)$$

where s_k is the target state vector given as follows

$$s_k = [x_k \quad \dot{x}_k \quad y_k \quad \dot{y}_k \quad \beta_k] \quad (38)$$

x_k and y_k are target positions, \dot{x}_k target velocity along x axis, and \dot{y}_k target velocity along y axis and B_k is the unknown ballistic coefficient that evolves through time as follows.

$$\beta_k = \beta_{k-1} + w_k^\beta \quad (39)$$

where w_k^β is a sequence of independent, identically distributed (IID) Gaussian variables with zero mean and variance \tilde{q} . Furthermore, g is the gravity acceleration and the matrices Φ and G are as follows

$$\Phi = \begin{bmatrix} 1 & \Delta & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & \Delta & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad G = \begin{bmatrix} \frac{\Delta^2}{2} & 0 \\ \Delta & 0 \\ 0 & \frac{\Delta^2}{2} \\ 0 & \Delta \\ 0 & 0 \end{bmatrix} \quad (40)$$

where Δ is the time interval between two consecutive radar measurements. w_k is a sequence of IID Gaussian random vectors, with zero mean and a covariance matrix as follows.

$$Q = \begin{bmatrix} \frac{q\Delta^3}{3} & \frac{q\Delta^2}{2} & 0 & 0 & 0 \\ \frac{q\Delta^2}{2} & q\Delta & 0 & 0 & 0 \\ 0 & 0 & \frac{q\Delta^3}{3} & \frac{q\Delta^2}{2} & 0 \\ 0 & 0 & \frac{q\Delta^2}{2} & q\Delta & 0 \\ 0 & 0 & 0 & 0 & \tilde{q}\Delta \end{bmatrix} \quad (41)$$

where q is a positive real number and \tilde{q} is the variance of w_k^β in (39). Finally, $f(s_k)$ is the nonlinear function in (37) that denotes the ballistic coefficient β_k and is given by

$$f(s_k) = -0.5 \frac{g}{\beta_k} \rho(y_k) \sqrt{\dot{x}_k^2 + \dot{y}_k^2} \begin{bmatrix} \dot{x}_k \\ \dot{y}_k \end{bmatrix} \quad (42)$$

where $p(\cdot)$ is the air density, which is defined as follows.

$$\rho(y_k) = c_1 \exp(-c_2 y) \quad (43)$$

with

$$\begin{cases} c_1 = 1.227, c_2 = 1.093 \times 10^{-4} & \text{for } y < 9144m \\ c_1 = 1.754, c_2 = 1.49 \times 10^{-4} & \text{for } y \geq 9144m \end{cases}$$

The two dimensional observation vector is as $z_k = [r_k \quad \varepsilon_k]^T$, where r_k is the measured range and ε_k is the elevation angle. Measurement equation is expressed as follows

$$z_k = \begin{bmatrix} \sqrt{x_k^2 + y_k^2} \\ \arctan(\frac{y_k}{x_k}) \end{bmatrix} + v_k \quad (44)$$

v_k is a Gaussian random sequence with zero mean and covariance matrix

$$R = \begin{bmatrix} \sigma_r^2 & 0 \\ 0 & \sigma_\varepsilon^2 \end{bmatrix} \quad (45)$$

To simulate the trajectory, the parameter values are chosen as $g = 9.8$, $\Delta = 2s$, $q = \tilde{q} = 5$, $\sigma_r = 150m$, $\sigma_\varepsilon = 150rad$. All three filters are initialized as follows

$$x_0 = \hat{x}_{0|0} = \begin{bmatrix} 232000m \\ 2290 \cos(190^\circ) m/s \\ 88000m \\ 2290 \sin(190^\circ) m/s \\ 40000 kg \cdot m^{-1} \cdot s^{-2} \end{bmatrix}$$

$$P_{0|0} = \text{diag}([1000^2 m^2, 50^2 m^2 \cdot s^{-2}, 1000^2 m^2, 50^2 m^2 \cdot s^{-2}, 2500^2 kg^2 \cdot m^{-2} \cdot s^{-4}])$$

To evaluate the performance of the proposed filters, two cases are considered:

A) It is assumed that the ballistic coefficient and aerodynamics forces are completely unknown, so instead of $f(\cdot)$ in (37) zero is replaced.

B) It is assumed that the ballistic coefficient and aerodynamics forces are completely known.

For the robust CKF, the parameters are selected as $A = 0.8I_{5 \times 5}$, $B = 0.2I_{5 \times 5}$ initial mean of uncertainty is $\hat{w}_0 = [0 \ 0 \ 0 \ 0 \ 0]^T$ and initial uncertainty covariance is assumed to be $P_{w,0} = Q$. In addition, the scaling parameter and the moving window width are $\gamma = 2.5$ and $s = 5$ respectively. The root mean square error (RMSE) index is used to compare the performance of the three filters. This index is defined as

$$RMSE(x) = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \hat{x}_i)^2} \quad (46)$$

For all of the three filters, the corresponding root mean square error (RMSE) curves for the estimated target position are shown in Figs (3-6). These figures have been achieved through the implementation of 100 Monte Carlo runs. It can be seen in these figures that when the ballistic coefficient and aerodynamics forces are unknown (figs 3-4), the robust CKF and hybrid robust CKF give better results than the CKF. But when the ballistic coefficient and aerodynamics forces are known (figs 5-6), the hybrid CKF follows the traditional CKF and the performance of these two filters is better than the robust CKF. So, both in presence and in the absence of any uncertainty, the hybrid CKF has a promising performance.

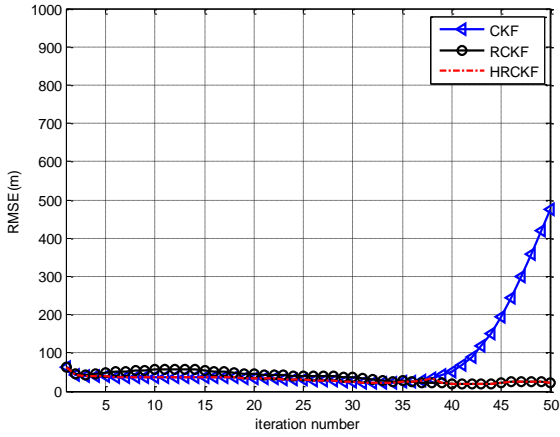


Fig. 3. RMS error along x-axis when the ballistic coefficient and aerodynamics forces are unknown.

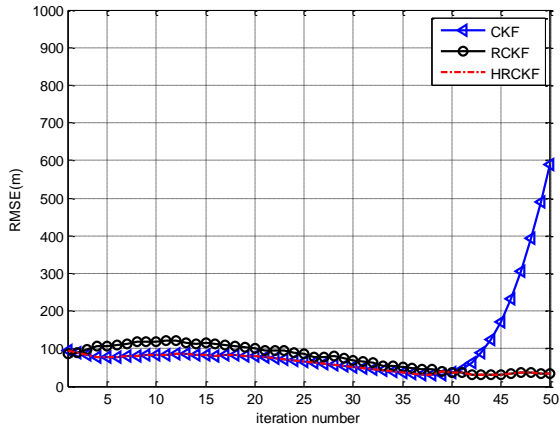


Fig. 4. RMS error along y-axis when the ballistic coefficient and aerodynamics forces are unknown.

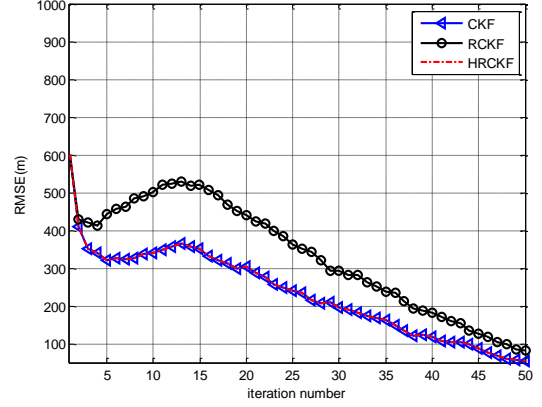


Fig. 5. RMS error along x-axis when the ballistic coefficient and aerodynamics forces are known.

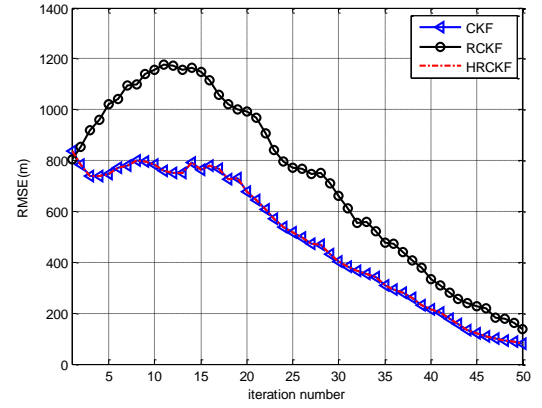


Fig. 6. RMS error along y-axis when the ballistic coefficient and aerodynamics forces are known.

Example 2) The Euler-discretized van der Pol oscillator is given as follows [20].

$$\begin{bmatrix} x_{1,k} \\ x_{2,k} \end{bmatrix} = \begin{bmatrix} x_{1,k-1} + Tx_{2,k-1} \\ -Tx_{1,k-1} + (T+1-Tx_{1,k-1}^2)x_{2,k-1} \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_{k-1} + w_{k-1} \quad (47)$$

where, $x_k = [x_{1,k} \ x_{2,k}]^T$ is the state vector, u_{k-1} is the external input assumed to be unknown, w_{k-1} is a white Gaussian noise with zero mean and covariance $Q = 10^{-6}I_{2 \times 2}$ and T is the sampling interval. Measurement equation is given as follows

$$z_k = \begin{bmatrix} 1 & 1 \end{bmatrix} x_k + v_k \quad (48)$$

The measurement noise v_k is a zero mean white Gaussian with covariance $R = 0.04$. The input u_{k-1} is given by.

$$u_{k-1} = \begin{cases} T \sin(2kT), & \text{if } kT < 10s \text{ or } kT > 30s \\ T \sin(2kT) + 0.5, & \text{if } 10s < kT < 20s \\ T \sin(2kT) - 0.5, & \text{if } 20s < kT < 30s \end{cases} \quad (49)$$

where $T = 0.1$. In this simulation, we set $x_0 = [1 \ 1]^T$, $\hat{x}_{0|0} = [0.5 \ 1.5]^T$ and $P_{0|0} = 0.5I_{2 \times 2}$.

To evaluate the performance of the proposed filters, two cases are considered:

A) It is assumed that the input u_{k-1} is unknown, so zero is replaced with u_{k-1} in (47).

B) It is assumed that the input u_{k-1} is known.

The parameters in the robust CKF algorithm are selected as $B = 0.2I_{2 \times 2}$ and $A = 0.8I_{2 \times 2}$. The initial mean of uncertainty is assumed to be $\hat{w}_0 = [0 \ 0]^T$ and initial uncertainty covariance is $P_{w,0} = Q$. The scaling parameter and the moving window width are selected as $\gamma = 1.5$ and $s = 4$ respectively, for the hybrid robust CKF. Performance of the three filters has been compared according to the RMSE of $x_{2,k}$ for 60 Monte Carlo simulation runs. Simulation results show that when the input is unknown, the robust CKF and hybrid robust CKF give better results in comparison with the CKF (fig 7). Furthermore, when the input is known, the CKF and hybrid robust CKF give better results in comparison with the robust CKF (fig 8). So, as it can be seen in the figures, in both cases, the hybrid robust CKF provides the best performance among all of the filters.

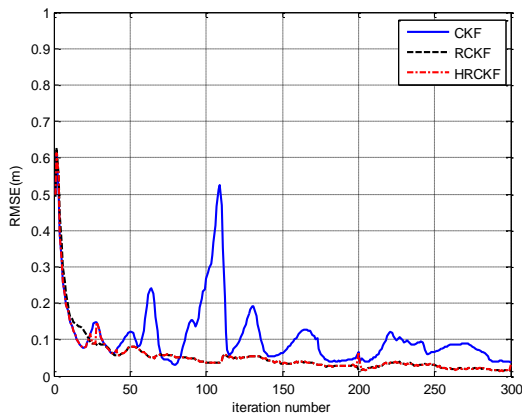


Fig. 7. RMS estimation error of x_2 with unknown input.

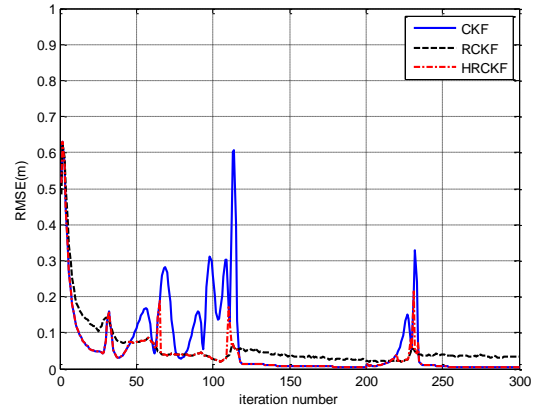


Fig. 8. RMS estimation error of x_2 with known input.

7. Conclusion

In the presented work, two novel methods of state estimation in nonlinear systems were proposed named robust CKF and hybrid robust CKF. The robust CKF was designed by including the uncertainty estimator in the traditional CKF which produces reliable estimates in presence of large modelling errors. The hybrid robust UKF was proposed to maintain a balance between uncertainty and estimation error. The hybrid robust CKF detects the uncertainty and adapts the system accordingly. Two examples have been considered to compare the performances of the proposed filters and it was found that the hybrid robust CKF provides the best results for any nonlinear system in the presence or absence of uncertainty.

References

- [1] R. Grover and P. Y. C. Hwang, Introduction to random signals and applied Kalman filtering. Wiley N. Y., 1992.
- [2] M. S. Grewal and A. P. Andrews, Kalman filtering: theory and practice using MATLAB. John Wiley & Sons, 2011.
- [3] I. Arasaratnam and S. Haykin, "Cubature Kalman filters," IEEE Trans. On Autom. Control, vol. 54, no. 6, pp. 1254–1269, Jun. 2009.
- [4] I. Arasaratnam, "Cubature Kalman filtering: theory & applications," Ph. D. Thesis, 2009.
- [5] B. Safarinejadian, M. A. Tajeddini, and A. Ramezani, "Predict time series using extended, unscented, and cubature Kalman filters based on feed-forward neural network algorithm," 3rd International Conference on Control Instrumentation and Automation, (ICCIA), 2013, pp. 159–164.
- [6] M. Havlicek, K. J. Friston, J. Jan, M. Brazdil, and V. D. Calhoun, "Dynamic modeling of neuronal responses in fMRI using cubature Kalman filtering," Neuroimage, vol. 56, no. 4, pp. 2109–2128, 2011.
- [7] D. Macagnano and G. T. F. de Abreu, "Multitarget tracking with the cubature Kalman probability hypothesis density filter," Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), 2010, pp. 1455–1459.
- [8] K. P. B. Chandra, D.-W. Gu, and I. Postlethwaite, "Cubature Kalman filter based localization and mapping," World Congress, 2011, pp. 2121–2125.
- [9] F. Yang, Z. Wang, and Y. Hung, "Robust Kalman filtering for discrete time-varying uncertain systems with multiplicative noises," IEEE Trans. On Autom. Control, vol. 47, no. 7, pp. 1179–1183, 2002.
- [10] Z. Dong and Z. You, "Finite-horizon robust Kalman filtering for uncertain discrete time-varying systems with uncertain-covariance white noises," IEEE Signal Process. Lett., vol. 13, no. 8, pp. 493–496, 2006.
- [11] U. Shaked and C. E. de Souza, "Robust minimum variance filtering," IEEE Trans. On Signal Process., vol. 43, no. 11, pp. 2474–2483, 1995.
- [12] Y. Theodor and U. Shaked, "Robust discrete-time minimum-variance filtering," IEEE Trans. On Signal Process., vol. 44, no. 2, pp. 181–189, 1996.
- [13] S. Habibi, "The smooth variable structure filter," Proc. IEEE, vol. 95, no. 5, pp. 1026–1059, 2007.
- [14] S. Dey and J. B. Moore, "Risk-sensitive filtering and smoothing via reference probability methods," IEEE Trans. On Autom. Control, vol. 42, no. 11, pp. 1587–1591, 1997.
- [15] H. Li and M. Fu, "A linear matrix inequality approach to robust H_∞ filtering," IEEE Trans. On Signal Process., vol. 45, no. 9, pp. 2338–2350, 1997.

- [16] R. S. Mangoubi, *Robust estimation and failure detection: A concise treatment*. Springer Science & Business Media, 2012.
- [17] L. Xie, Y. C. Soh, and C. E. de Souza, "Robust Kalman filtering for uncertain discrete-time systems," *IEEE Trans. On Autom. Control*, vol. 39, no. 6, pp. 1310–1314, 1994.
- [18] S. J. Kwon, "Robust Kalman filtering with perturbation estimation process for uncertain systems," *IEE Proc.-Control Theory Appl.*, vol. 153, no. 5, pp. 600–606, 2006.
- [19] M. G. S. Bruno and A. Pavlov, "Improved sequential Monte Carlo filtering for ballistic target tracking," *IEEE Trans. On Aerosp. Electron. Syst.*, vol. 41, no. 3, pp. 1103–1108, 2005.
- [20] B. Teixeira, J. Chandrasekar, H. J. Palanthandalam-Madapusi, L. Torres, L. A. Aguirre, and D. S. Bernstein, "Gain-constrained Kalman filtering for linear and nonlinear systems," *IEEE Trans. On Signal Process.*, vol. 56, no. 9, pp. 4113–4123, 2008.

Behrouz Safarinejadian received the B.Sc and M.Sc degrees from Shiraz University, Shiraz, Iran, in 2002 and 2005, respectively, and the Ph.D degree from Amirkabir University of Technology, Tehran, Iran, in 2009. Since 2009, he has been with the Faculty of Electrical and Electronic Engineering, Shiraz University of Technology, Shiraz, Iran. His research interests include computational intelligence, control systems theory, estimation theory, statistical signal processing, sensor networks, and fault detection.

Mohsen Taher received the B.Sc degree in electrical engineering from Quchan University of Advanced Technologies Engineering, Quchan, Iran, in 2010 and the M.Sc degree in electrical engineering from Shiraz University of Technology, Shiraz, Iran, in 2016. He is currently with the Control Engineering Department, Shiraz University of Technology. His research interests include distributed state estimation, sensor networks and consensus filter.

Quality Assessment Based Coded Apertures for Defocus Deblurring

Mina Masoudifar

Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
mi.masoudifar@stu.um.ac.ir

Hamid Reza Pourreza*

Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
hpourreza@um.ac.ir

Received: 02/Jan/2016

Revised: 14/May/2016

Accepted: 23/May/2016

Abstract

A conventional camera with small size pixels may capture images with defocused blurred regions. Blurring, as a low-pass filter, attenuates or drops details of the captured image. This fact makes deblurring as an ill-posed problem. Coded aperture photography can decrease destructive effects of blurring in defocused images. Hence, in this case, aperture patterns are designed or evaluated based on the manner of reduction of these effects. In this paper, a new function is presented that is applied for evaluating the aperture patterns which are designed for defocus deblurring. The proposed function consists of a weighted sum of two new criteria, which are defined based on spectral characteristics of an aperture pattern. On the basis of these criteria, a pattern whose spectral properties are more similar to a flat all-pass filter is assessed as a better pattern. The weights of these criteria are determined by a learning approach. An aggregate image quality assessment measure, including an existing perceptual metric and an objective metric, is used for determining the weights. According to the proposed evaluation function, a genetic algorithm that converges to a near-optimal binary aperture pattern is developed. In consequence, an asymmetric and a semi-symmetric pattern are proposed. The resulting patterns are compared with the circular aperture and some other patterns in different scenarios.

Keywords: Coded Aperture; Blur; Defocus; Computational Imaging; Image Quality Assessment.

1. Introduction

In imaging with a lens-based camera, focal plane is defined. Therefore, if an image is captured of a scene with varying depth, then out of focus regions are blurred. The amount of blurring depends on the size of aperture. If the size of aperture is extended, then depth of field is decreased and defocus blur is increased. Hence, smaller apertures are desired to decrease the scale of blur. On the other hand, by growing the resolution of camera systems, size of pixels has been reduced. Accordingly, wider apertures are needed to obtain light required for these small pixels and maintain signal-to-noise ratio in the captured image. As a result, there is a challenge between the size of circular aperture and the scale of blur in conventional cameras [1]. Coded aperture photography is a field of computational imaging that can be used for gathering light of a wider aperture with less destructive effects of defocusing. A coded aperture camera is a conventional camera with a mask on the aperture. This type of camera has been used in several applications such as defocus deblurring [2-4]; depth estimation [5-10]; estimating depth and image [11-13]; super-resolution [14] and so on. A comprehensive review about computational cameras is found in [15,16]. The main idea in coded aperture for deblurring is to use a mask on aperture in order to change the pattern of rays passed through it. In this way, the shape of defocus kernel is changed,

whereupon damaging effects of defocus blur is reduced. As a result, deblurring operation on the images captured by this type of camera is more successful compared to the images taken with a conventional camera. The first idea of using coded aperture imaging was introduced in the field of astronomy. Various patterns have been designed for lens-less imaging of gamma-ray or X-ray sources. Such patterns are designed with the aim of collecting more lights in order to improve signal-to-noise ratio. A comprehensive study about these techniques is found in [17]. One of the well-known patterns introduced in recent decades is modified uniform redundant array (MURA) [18]. However, these patterns, which are designed for lens-less imaging, are not suitable to use with lenses for defocus deblurring [2,3]. In lens-based imaging, the first applications of unconventional apertures were introduced in the optic field in order to increase depth of field or compensate attenuated waves [19,20]. The approaches used in these applications are principally based on the optical properties of imaging systems. In recent years, other approaches have been proposed for defocus deblurring in lens-based imaging. Hiura et al. [6] design a multi-focus coded aperture camera that simultaneously captures three images with different focus values. Four-pin-hole and two-pin-hole apertures are used for depth estimation and deblurring, respectively. Veeraraghavan et al. [3] search for a mask pattern such that the minimum magnitude of its Fourier spectrum is maximized. MURA

* Corresponding Author

pattern is used as initial pattern and then a gradient descent search is used for finding the optimum pattern. The non-binary obtained pattern is pruned to get a sub-optimal 7×7 binary pattern with a random search algorithm. This search is very time-consuming [3]. Zhou et al. [2] define an objective function aimed to minimize residual error of deblurred images defocused by a 13×13 binary coded aperture. A genetic algorithm is used to find a near optimal pattern. According to their objective function, patterns provided for each amount of noise are different. They also introduce in [11] a coded aperture pair to capture two images for both depth map and all-focus image estimation. Masia et al. [4] extend the idea proposed in [2]. A simple aggregate measure consists of normalized root mean square error (RMSE) and two perceptual metrics, including structure similarity index (SSIM) and High Dynamic Range-Visual Difference Predictor (HDR-VDP2), is defined. To evaluate an aperture pattern, a special image with nearly wide bandwidth of power spectra, is blurred by the pattern and then deblurred. Quality of the deblurred image is evaluated by the proposed aggregate measure. This value determines the fitness of the corresponding pattern. More fitness value means more appropriate pattern. A genetic algorithm is developed to find a near optimal pattern in size 11×11 . Designing masks while taking into account perceptual image quality assessment criteria is a valuable idea. However, spectral response of the selected image is not completely matched to the statistical model of natural images. Hence, decision based on the deblurring result of only one image might not lead to the best design.

Figure 1 shows a circular conventional aperture and some patterns introduced so far for different purposes.

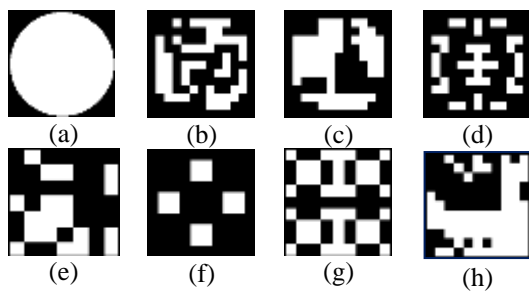


Fig. 1. Conventional aperture and some pattern designed for coded aperture. (a) Conventional aperture; aperture mask proposed by: Zhou et al. [2] for (b) $\sigma_{\text{noise}} \sim N(0,0.001)$ and (c) $\sigma_{\text{noise}} \sim N(0,0.005)$, (d) Levin et al. [5], (e) Veeraraghavan et al. [3] (f) Hiura et al. [6], (g) MURA mask [18] and (h) Masia et al. [4].

In defocus deblurring, the main aim of aperture designing is to compensate the low response of the blurring kernel to high frequencies. Indeed, if spectral response of a pattern has the maximum similarity to a flat all-pass filter, the pattern is assumed as the best. Hence, our main idea in evaluating a pattern is measuring the amount of this similarity. Based on this idea, a weighted evaluating function consists of two terms is defined. The weight of each term is determined based on the quality assessment results of deblurred images taken with some existing

aperture patterns. A genetic algorithm is applied for finding a pattern that satisfies the proposed evaluating function.

The rest of this paper is organized as follows: in Section 2, first a review of blurring problem formulation is presented. Then an aggregate measure is defined for evaluating the quality of deblurred images. Section 3 describes our method to find near optimal aperture pattern. Analysis and performance comparison is discussed in Section 4. Finally, conclusions are drawn in Section 5.

2. Background

2.1 Blurring Problem Formulation

Defocussing acts as a low-pass filter in which high frequencies or details of a captured photo are attenuated or dropped. For a simple fronto-parallel object at depth d , defocusing is defined as convolution of a defocus kernel or point-spread-function (PSF), called k^d with a sharp image (f_0) which causes a spatial invariant blur:

$$f_1 = k^d \otimes f_0 + \eta, \quad \sum_i k_i^d = 1 \quad (1)$$

where η corresponds to additive noise and k_i^d refers to the elements of k^d . The superscript d indicates the kernel k is a function of depth of scene. It is usually assumed the additive noise (η) is a Gaussian noise $N(0, \sigma^2)$ [21].

Equivalently, spatially invariant blur in frequency domain is defined as Eq.2:

$$F_1 = K^d \cdot F_0 + \zeta = |K^d| \cdot |F_0| e^{i(\varphi_{K^d} + \varphi_{F_0})} + \zeta \quad (2)$$

This multiplication means the spectrum of in-focus image (F_0) is filtered by the spectrum of the filter K^d , which is also called Modulation Transfer Function (MTF), and then the noise ζ is added. Defocus kernel resulted from a conventional circular aperture has a Bessel-like spectrum. Therefore, some spectra of the sharp image are damped or lost especially in higher frequencies. Accordingly, deblurring aimed to design an inverted filter, is assumed as an ill-posed problem [21]. If the entire scene is in-focus, then no frequency will be dropped and K^1 can be assumed as a flat all-pass filter, namely 1. Our main idea is to design a pattern whose spectral properties have the most possible similarity to a flat all-pass filter.

2.2 Quality Assessment of Deblurred Images

Image restoration has a long history in the field of image processing. Up to now, several methods have been proposed to estimate a sharp image (\hat{f}_0) from a noisy-blurred image (f_1), while many problems in this field have not been solved yet [22].

Quality of restored results is measured by image quality assessment methods. There are various methods including objective and perceptual approaches that

¹ In the rest of text, we use notation K instead of K^d which can be generalized to each depth d . In addition, for simplicity and without loss of generality, we suppose that K has a vectorized form. Therefore, we use 1-dimensional notations.

compute similarity or dissimilarity of a reference image to a test image. In our study, the deblurred image and the in-focus image are assumed as test and reference images, respectively. Usually objective quality measures or metrics, which work at pixel level (such as RMSE), are used for quality evaluation. However, human perception of image quality is not necessarily correlated with these measures [4,23].

Masia et al. [4] showed that using both types of these measures lead to more precise qualification of the restored images. As mentioned before, they used RMSE, SSIM and HDR-VDP2; but they didn't give any specific reason for choosing these measures. In this research, we use an aggregate measure of objective and perceptual metrics. This measure is used for two purposes. First, it is used in the process of designing a new aperture pattern. Then, deblurred results of images captured with different patterns are evaluated with this aggregate measure. In the rest of this section, selected measures and the reason of choosing them are briefly described.

The proposed aggregate measure includes RMSE and visual information fidelity (VIF). RMSE is a well-known objective quality assessment measure that is defined to compute the difference between two images (\hat{f}_0, f_0) with size $R \times C$ as Eq.3:

$$\text{RMSE}(\hat{f}_0, f_0) = \sqrt{\frac{1}{R \times C} \sum_{x=1}^C \sum_{y=1}^R [f_0(x, y) - \hat{f}_0(x, y)]^2} \quad (3)$$

VIF[23] is a full reference perceptual quality assessment measure for computing visual fidelity of the test image to the reference image. The reference image is modeled as the output of a stochastic "natural" source that traverses the human visual system (HVS) channel and then is processed by the brain. The information that the brain can extract from the output of the HVS channel is quantified. The same measure is computed for the test image, which may be disfigured by an image distortion type. Image distortion is modeled in wavelet domain and includes various distortion types such as blur, additive noise and global or local contrast changes. VIF computes the fidelity or similarity of the information extracted from the test image to the information extracted from the reference image. Because of the complexity of equations in computing VIF, we avoid to describe the method of computing in details and refer readers to the original publication [23].

According to [24,25], VIF is the most precise quality assessment measure if images are distorted by artifact or blur whereas RMSE and consequently peak-signal-to-noise ratio (PSNR) are good evaluators if images are distorted by noise.

In real world, deblurred images suffer from various types of error such as ringing artifacts and inverted noise (deconvolution noise) [21,23]. Since the performance of all quality assessment measures is reduced in the presence of several distortion types in an image [24], using an aggregate measure, which its terms are sensitive to

different types of distortion, improves the accuracy of quality assessment [4,26].

Accordingly, the following aggregate measure is used for assessing the quality of deblurred images that is sensitive to both artifact and noise:

$$Q = (1 - \text{RMSE}) + \text{VIF} \quad (4)$$

The value of pixel intensities are assumed to be between 0 and 1. Therefore, the range of both RMSE and VIF is [0-1]. It is clear that a larger value of Q signifies a better result. Although equal weights of quality measures were used for computing Q, other possibilities may be applicable which will be studied in our future works.

3. Aperture Pattern Design

The main object of this paper is to find a pattern that reduces ill-posedness of blurring problem in defocused images. Therefore, we must search for a pattern whose corresponding filter is similar to a flat all-pass filter. A flat all-pass filter has some explicit properties: its spectral response to all frequencies is as high as possible and this response has no fluctuation, so it has no serious drop. According to the following reason, we don't focus on phase properties of a mask. Based on Wiener restoration algorithm, if kernel K is known, phase properties of K have no effect on deblurring error. This matter can be resulted directly from the criterion introduced in [2], which is based on the amount of deblurring error obtained by Wiener filter. As a result, our criteria for evaluating a filter is defined as follows:

3.1 Defining Criteria

3.1.1 Distance of Filter to an All-Pass Filter

As mentioned in Section 1, using a coded mask on lens changes the shape and properties of the defocus kernel, thus weakening the high frequencies can be decreased. However, because the aperture is partially masked, the amount of light passed through the aperture is reduced. Therefore, in a fixed exposure time, using a mask causes to reduce the brightness of the captured image. This reduction can be modeled by decreasing the sum of kernel elements in Eq. 1, whereupon the spectral response of the corresponding filter is affected. Reduction of the brightness can be compensated slightly during the deblurring operation. However, if the transmitted light is very low, signal to noise ratio of the captured image is decreased. On the other hand, increasing the exposure time in order to compensate this reduction is not desired. This problem is one of challenges in coded aperture photography that has been discussed completely in [27,28]. Hence, we must search for a pattern whose spectral response to all frequencies is as high as possible. Therefore, to find a filter whose spectral response is as similar as possible to 1, the first criterion is defined as Euclidean distance between the spectrum of filter K and 1.

$$C_1 = \|1 - |K|\|_2 \tag{5}$$

3.1.2 Derivation of Spectral Response

Using $\|\cdot\|_2$ in computing the criterion C_1 causes to assign large penalties to some frequency components of K that have a low magnitude of spectrum. However, this criterion does not guarantee to obtain a flat spectral response. Hence, the second criterion is defined as the norm of gradient of K .

$$C_2 = \|\nabla|K|\|_2 \tag{6}$$

Obviously, less fluctuation in spectral response derives from less fluctuation in difference between K and $\mathbf{1}$. This could be easily shown by computing the derivative of C_1 with respect to the frequency component u .

$$\frac{\partial C_1}{\partial u} = -2 \frac{\partial |K|}{\partial u} (1 - |K|) = 0 \implies \begin{cases} \frac{\partial |K|}{\partial u} = 0 & (a) \\ |K| = 1 & (b) \end{cases} \tag{7}$$

Both of equations (7.a) and (7.b) emphasize that the least fluctuation of spectral response is needed to have a uniform similarity. Since K is a low-pass filter, condition (7.b) is not obtainable. Therefore, the second criterion is used to get closer as possible to the condition (7.a).

3.2 Evaluation Function

For a filter K with size $M \times M$, the range of values for C_1 and C_2 are $[0..M^2]$. However, in practice, values of C_2 are much smaller than C_1 . On the other hand, they may have different significance in pattern evaluation. Hence, the evaluation function is defined as a weighted sum of the proposed criteria:

$$F_{\text{pattern}} = a_1 C_1 + a_2 C_2 \tag{8}$$

As we discuss later, M is set to 32. A pattern with minimum value of F is assumed as the best pattern. To find the best values for the weights a_1 and a_2 , a search strategy has been used that is described in the Section 3-3.

3.3 Computing Weights

For computing weights, coded aperture imaging system is simulated with different existing aperture patterns. The capture process is simulated by multiplying the Fourier transform of the sharp image (F_0) to the defocus kernel (K^d) obtained of a pattern and then the noise ζ is added. Deblurring is performed with an improved version of Wiener algorithm proposed in [2]. This algorithm is chosen because of its appropriate quality and speed.

At first, properties of 8 patterns shown in Figure 1 are studied in 6 different blur scales varying between 3 and 8 pixels in radius. As stated in Section 4, this range of blur sizes covers an adequate range of blur scales in real scenes. To have a more precise evaluation, each blurring kernel is zero padded into a 32×32 matrix, and then its Fourier transform is computed. The values of the two proposed criteria are computed for each 48(8×6) blurring kernel (K). Then, the performance of these kernels is evaluated. For this purpose, 20 different images consist of indoor and

outdoor natural images and some resolution charts are chosen in a manner that they model the expected spectrum of natural images [29]. In this way, we could have a fair evaluation about each kernel. (The average size of used pictures is about 640×480 .) Figure 2 shows some of used pictures and the average of their spectral response.

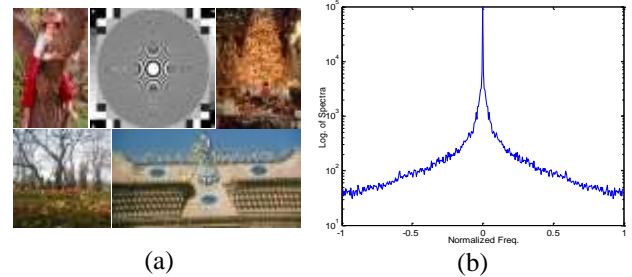


Fig. 2. (a) Some of images used for evaluating the existing apertures, (b) Log scale of average power spectrum of 20 selected images.

These images are blurred by each kernel (K) and then deblurred. The quality of each deblurred image is computed with Eq.5 and then the average of quality is computed over 20 images. Since Zhou et al.[2] declared interdependence of appropriate aperture pattern and the amount of noise, this scenario is repeated separately on 6 different levels of additive noise ($\sigma = 0.0005, 0.001, 0.005, 0.01, 0.015, 0.02$). In this way, 48 different Q value are obtained for each amount of noise. For each level of noise, a_1 and a_2 must be found in a manner that statement 9 holds true for each paired kernel (K, K'):

$$Q > Q' \iff F < F' \tag{9}$$

where F and F' refers to the fitness value of K and K' , respectively. By extending the statement 9, we have:

$$F < F' \rightarrow a_1 C_1 + a_2 C_2 < a_1 C'_1 + a_2 C'_2 \implies \tag{10}$$

$$a_1 (C_1 - C'_1) + a_2 (C_2 - C'_2) < 0 \tag{11}$$

By dividing two sides of inequality 11 on a_1 (suppose $a_1 > 0$), a simplified form is obtained:

$$(C_1 - C'_1) + b_2 (C_2 - C'_2) < 0 \tag{12}$$

Indeed, without loss of generality and just for simplicity, we can fix $a_1 = 1$ and reduce the space of solution to find another coefficient. This inequality divides 1-D search space into feasible and infeasible regions. Inequality 12 must hold true for each paired kernel (K, K') that $Q > Q'$. Hence, there is a linear inequality system that its solution gives the final feasible region in which statement 9 is true for all paired kernel. Table 1, shows individual computed feasible values of b_2 corresponding to each amount of noise.

Table 1. The range of b_2 values according to the variance of additive noise.

σ	b_2
0.0005	[15.88..19.75]
0.001	[15.88..19.75]
0.005	[15..17.5]
0.01	[9.68..15.46]
0.015	[9.68..10.33]
0.02	[6.1..9.8]

As shown in Table 1, by increasing the amount of noise, the importance of C_1 relative to C_2 increases gradually. It means in higher levels of noise, more light is needed to keep signal to noise ratio.

It must be noticed the large values of b_2 does not imply that C_1 is less significant. As mentioned earlier, in practice, values of C_1 is greater than C_2 . Therefore, b_2 nearly compensates this difference as well as determining the significance of each criterion in evaluating the fitness of a pattern.

According to Table 1, choosing different amount of b_2 leads to different patterns. However, if designing only one pattern is desired, it is preferred to choose a value of b_2 that is near to the most of feasible regions. In this way, the designed pattern will be appropriate for a wide range of noise. Regarding to Table 1, b_2 is set to 15.5. This point is inside the feasible region of $\sigma = 0.005$ while being close to the feasible region of at least three other levels of noise (0.0005, 0.001, 0.01). In fact, Table 1 clarifies why a pattern designed for $\sigma=0.005$ has better performance in other levels of noise. This is the same result that has been experimentally experienced in [4].

Accordingly, a pattern with filter K is evaluated by Eq. 13:

$$F(K) = C_1(K) + 15.5 \times C_2(K) \quad (13)$$

3.4 Mask Resolution Determination

In this study, mask resolution is determined such that each single hole provides no diffraction. According to superposition property in coded aperture imaging, if a single hole of a pattern does not provide any diffraction, then the image composed of rays passed through all the holes does not provide any diffraction [7]. Based on formula proposed in [7], a 7×7 mask is appropriate for an imaging system with aperture-diameter = 20^{mm} and pixel-size = $11.5^{\mu\text{m}}$. According to the camera specification used in our experiments, this resolution is selected for our mask. The related formula has been stated clearly in [7]. It must be mentioned higher resolution could be chosen if both diffraction and defocus were formulated in blurring function. This will be considered in our future study.

3.5 Optimization

Our goal is to obtain a pattern that minimizes the value of function F as defined in Eq.13. Generally, there are two main approaches to find the best pattern. In the first approach, Fourier transform of an initial pattern is used for finding optimal pattern. Then, a constraint linear or non-linear problem must be solved to find the answer. Inverse Fourier transform of the answer gives a non-binary pattern. However, finding the best binary pattern from the answer is very time consuming. This problem has been already reported in [3].

Another approach is using evolutionary algorithms. Genetic algorithm is the most popular type of evolutionary algorithms. A population of random binary patterns is created. The fitness value of each pattern is computed. In our case, spectral properties of each pattern determine the

fitness. Population evolves by using some breeding operators such as crossover and mutation. After some generations, the population is converged to a final pattern. This simple yet effective method has been used in [2,4].

Because the search space of patterns is very large, implementing heuristic search strategies such as random restart hill-climbing is impractical.

In this study, we implemented a genetic algorithm, which is described here in details. A generation of binary patterns with population size 1000 is created. A pattern is defined by a vector of 49 binary elements. According to [30], this size of population ensures that our search is converged to a proper solution. The fitness value of each pattern is computed based on Eq.13 (i.e. F value). Patterns are selected by the stochastic uniform method and then are evolved by crossover with $p_c = 0.8$ and mutation with $p_m = 0.05$. Selection and reproduction are repeated until the average change of fitness value over last 10 generations is less than 0.00001. In our case, convergence occurs after about 50 generations. In this way, we find a pattern shown in Figure 3.a. Like other studies, resulting pattern is not symmetric. Asymmetric patterns are not rotation-invariant. In addition, some photographers would like to use symmetric apertures. Therefore, our genetic algorithm is repeated to find a symmetric pattern. In this search, each chromosome has 16 bits length that contains nearly a quarter of a pattern with size 4×4 . The complete form of a pattern is obtained by reflecting the early version (4×4 pattern) along the last column as vertical pivot and then along the last row as horizontal pivot. The fitness is computed based on Eq.13. Because of shortening the length of each chromosome, convergence occurs in about 30 generations. As expected, the fitness value of the resulted pattern is slightly lower than the asymmetric one. Since Levin et al. [5] reported earlier that symmetric patterns might contain more zero frequencies than asymmetric ones, this result is not surprising. Figure 3.b shows the resulted pattern. This pattern is symmetric over a 90° rotation, so we call it semi-symmetric.



Fig. 3. Patterns obtained of optimization. (a) Asymmetric pattern, (b) Semi-symmetric pattern.

The transmission rate (compared to the circular aperture) of our optimized asymmetric and semi-symmetric patterns are 0.4025 and 0.3517, respectively. The transmission rate is 0.4001, 0.4067 and 0.5856 for patterns proposed by Veeraraghavan et al. [3], Zhou et al. [2] ($\sigma \sim N(0,0.001)$) and Masia et al. [4], respectively. Transmission rate of our asymmetric pattern is close to the patterns proposed in [2] and [3], which all of them are searched in a search space containing all patterns with different transmission ratio. However, in [4], transmission

ratio is fixed to 0.5856 and search space contains just patterns with the same transmission rate.

4. Analysis and Performance Comparison of Apertures

In this section, our proposed patterns are evaluated in comparison with conventional aperture and other patterns proposed for defocus deblurring in [2-4]. It must be mentioned, among several patterns proposed in [2], we choose the pattern provided for noise ($\sigma \sim N(0,0.001)$). The reason is that Zhou et al. reported that this pattern in general is more efficient than other patterns proposed by them [2].

In the rest of this section, first, the spectral response of these apertures are compared. Then, deblurring results of images captured with them are evaluated by simulation and real experiments.

4.1 Analysis in Spectral and Spatial Domain

As mentioned before, optimal apertures for defocus deblurring seek a smooth spectral response while transmitting light as much as possible. Figure 4.a shows 1D slice of Fourier transform of our patterns in comparison with other patterns and circular aperture. It shows our patterns keep high spatial frequencies and have less fluctuation. We also compare these patterns in spatial domain. For each pattern, a convolution matrix (blurring matrix) of the blurring kernel is computed. Then, singular values of each matrix are determined using SVD. The slop of singular values shows attenuation rate of information in the captured image in any direction [21]. Plotted values in Figure 4.b show that the blurring matrix of our pattern has larger singular values in higher frequencies that lead to less attenuation of details in the defocused captured image. The singular values in the semi-symmetric pattern are smaller than asymmetric one while greater than other patterns in most of frequencies. Therefore, better deblurring results are expected of both our asymmetric and semi-symmetric patterns.

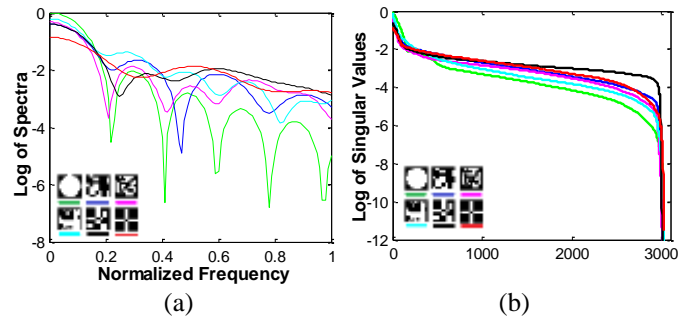


Fig. 4. Spectral and SVD analysis of patterns. Log spectrum(a) and Singular values(b) of conventional pattern(green) and patterns proposed in[3] (blue), [2] (magenta), [4](cyan) and proposed asymmetric(black) and semi-symmetric(red) patterns.

4.2 Performance Comparison of Apertures

In this part, results of our experiments are demonstrated. At first, the mentioned apertures are evaluated via simulation in different scenarios. Then deblurring results of real scenes are examined.

4.2.1 Performance Evaluation via Simulation

The imaging system is simulated as described in Section 3. Then the patterns studied in Section 4.1 are examined. For each pattern, 20 various images are blurred and then various amount of additive noise are added synthetically. These images include some outdoor images selected from an image database [31] and some indoor images taken with a handheld camera. Then, images are deblurred using the modified Wiener algorithm [2]. Tables 2-4 show results of this experiment in three different sizes of blur. Each entry of these tables indicates the average of the labeled measure over 20 images. In the smallest blur scale (blur-size = 5), the pattern proposed in [3] and our semi-symmetric pattern provide better results and our asymmetric pattern has the second rank. However, by increasing the blur scale, the proposed asymmetric pattern gives better results than other apertures. Interestingly, in many situations our semi-symmetric pattern provides better results compared to asymmetric patterns proposed in [2,4].

Table 2. Performance evaluation of three apertures across four different levels of noise (blur size = 5).

Quality σ	RMSE						VIF						Q					
	Conv.	Veera.	Z.001	Masia	Sym.	Asym.	Conv.	Veera.	Z.001	Masia	Sym.	Asym.	Conv.	Veera.	Z.001	Masia	Sym.	Asym.
0.001	0.0221	0.0089	0.0159	0.0176	0.0115	0.0097	0.8575	0.9627	0.8934	0.9427	0.9506	0.9609	1.8355	1.9538	1.8775	1.9251	1.9391	1.9511
0.005	0.0442	0.0297	0.0403	0.0397	0.0261	0.0298	0.7061	0.7747	0.6604	0.7418	0.8235	0.7832	1.6619	1.7450	1.6201	1.7020	1.7974	1.7534
0.01	0.0538	0.0435	0.0547	0.0512	0.0363	0.0430	0.6232	0.6423	0.5439	0.6132	0.7184	0.6645	1.5693	1.5989	1.4891	1.5620	1.6822	1.6215
0.02	0.0624	0.0586	0.0695	0.0643	0.0487	0.0570	0.5235	0.5100	0.4367	0.4835	0.5869	0.5360	1.4611	1.4514	1.3672	1.4192	1.5383	1.4789
Avg on σ	0.0456	0.0352	0.0451	0.0432	0.0306	0.0349	0.6776	0.7224	0.6336	0.6953	0.7699	0.7361	1.6320	1.6873	1.5885	1.6521	1.7393	1.7012

Table 3. Performance evaluation of three apertures across four different levels of noise (blur size = 13).

Quality σ	RMSE						VIF						Q					
	Conv.	Veera.	Z.001	Masia	Sym.	Asym.	Conv.	Veera.	Z.001	Masia	Sym.	Asym.	Conv.	Veera.	Z.001	Masia	Sym.	Asym.
0.001	0.0562	0.0200	0.0174	0.0261	0.0237	0.0174	0.5196	0.8519	0.8586	0.7748	0.7389	0.8768	1.4634	1.8319	1.8412	1.7487	1.7152	1.8594
0.005	0.0775	0.0461	0.0448	0.0558	0.0454	0.0431	0.3760	0.5408	0.5412	0.5218	0.5557	0.6029	1.2985	1.4947	1.4964	1.4660	1.5103	1.5598
0.01	0.0887	0.0612	0.0642	0.0707	0.0582	0.0590	0.3149	0.4631	0.4004	0.4118	0.4624	0.4659	1.2262	1.4019	1.3362	1.3411	1.4042	1.4069
0.02	0.1011	0.0786	0.0859	0.0854	0.0759	0.0755	0.2545	0.3436	0.2848	0.3143	0.3444	0.3456	1.1534	1.2651	1.1989	1.2289	1.2685	1.2701
Avg on σ	0.0809	0.0515	0.0531	0.0595	0.0508	0.0488	0.3662	0.5499	0.5212	0.5057	0.5253	0.5728	1.2854	1.4984	1.4682	1.4462	1.4745	1.5241

Table 4. Performance evaluation of three apertures across four different levels of noise (blur size = 21).

Quality σ	RMSE						VIF						Q					
	Conv.	Veera.	Z.001	Masia	Sym.	Asym.	Conv.	Veera.	Z.001	Masia	Sym.	Asym.	Conv.	Veera.	Z.001	Masia	Sym.	Asym.
0.001	0.0677	0.0337	0.0343	0.0453	0.0403	0.0328	0.3824	0.7631	0.7360	0.6055	0.6494	0.8035	1.3147	1.7293	1.7017	1.5602	1.6091	1.7707
0.005	0.0933	0.0589	0.0615	0.0748	0.0610	0.0562	0.2640	0.4888	0.4486	0.3828	0.4690	0.5300	1.1707	1.4299	1.3871	1.3080	1.4080	1.4738
0.01	0.1060	0.0729	0.0788	0.0893	0.0718	0.0689	0.2124	0.3728	0.3245	0.2962	0.3820	0.4077	1.1064	1.3000	1.2458	1.2069	1.3102	1.3388
0.02	0.1187	0.0894	0.0991	0.1035	0.0853	0.0855	0.1697	0.2735	0.2273	0.2252	0.2949	0.2949	1.0509	1.1841	1.1282	1.1217	1.2096	1.2094
Avg on σ	0.0964	0.0637	0.0684	0.0782	0.0646	0.0609	0.2571	0.4746	0.4341	0.3774	0.4488	0.5090	1.1607	1.4108	1.3657	1.2992	1.3842	1.4482

4.2.2 Performance Evaluation in Real Scenes

For real experiments, the proposed patterns are printed as well as some other patterns on a single photomask sheet. To experiment with a specific aperture pattern, it is cut out of the photomask sheet and inserted into a camera lens. In our experiment, a Canon EOS 1100D camera with an EF 50mm f/1.8 II lens is used. The assembled lenses with the proposed masks are shown in Figure 5.



Fig. 5. Lens assembled with proposed masks.

A very thin LED is used to calibrate the true PSF. The LED is mounted behind a pierced black cardboard to make a point light source. For each aperture pattern, the camera focus is set to 1.2^m. Then, the camera is moved back until 2^m in 10^{cm} increments. At each depth, an image is captured. Each image is cropped according to the surface that the point light spreads. In many cases, the size of resulting PSF is close to the blur size that can be approximated based on thin lens formula, depth and the parameters of the camera [7]. Afterward, by using some threshold values, residual light is cleared and the result is normalized. It is a common way to estimate PSF [3, 4]. To have a fair comparison, in all experiments, we use the same setting of [4]. The camera is set to F# = 2, Te = 1/20^{sec}, ISO-sensitivity = 200, resolution¹ = S3. According to [28], the illumination condition is adjusted as office room. The selected camera resolution produces images in size of 720×480 with pixel-size ≈ 30.6^{µm}. Since in computing mask resolution the pixel-size is set to 11.5^{µm}, our pattern could also be used with camera resolution = S2 without concerning about diffraction. However, we choose resolution S3 that confirms us other patterns with smaller holes, which are studied in our experiments, provide no diffraction. Regarding to the selected resolution and depth range, the size of calibrated PSFs are varies between 5 and 15 pixels which are almost equal to blur scales that were used for computing weights (Section 3). Figure 6 shows some calibrated PSFs of patterns used in our experiments.

Fig. 6. Calibrated PSFs for some of the evaluated patterns in depth 80^{cm}.

Experiment 1.

In the first test of a real scene, Circular Zone Plate Chart (CZP) is placed at different depths (10, 30, 50, 70^{cm}) and one image is captured at each depth. Imaging noise is estimated about 0.01. It is estimated by some tests on uniform and unicolor scenes. Therefore, Wiener filter with NSR = 0.01 is used for deblurring. The blurred captured images are restored with the calibrated PSFs of each pattern. Figure 7 shows deblurred results in depth 70^{cm}. Notice that the captured images have different brightness levels since different apertures absorb different amounts of light.

We also perform a quantitative analysis to compare the performances of these apertures. In each depth, a defocus image is captured. Then, without moving camera or chart, an all-focused version is also captured. After restoration, the deblurred image is aligned carefully to its corresponding focused one. Then, quality of deblurred images is assessed in comparison with their focused ones. Figure 8 shows RMSE, VIF and Q measure of the restored images. It shows both the proposed patterns give better performance compared to other apertures. However, like simulation results (see tables 2-4), in lower depths (i.e. smaller blur scale), the semi-symmetric pattern yields better results than the asymmetric one. By increasing depth, our asymmetric pattern outperforms than the semi-symmetric pattern. It can be explained by studying the frequency responses of these two masks. In smaller blur scales, the blurring kernels of both masks have almost flat frequency responses. However, the asymmetric pattern falls behind the semi-symmetric one because of a fall in its frequency response in the normalized frequency of 0.25 (see Fig. 4.a). By increasing the blur scale, the asymmetric pattern, which has higher frequency response and less fluctuation in high spatial frequencies, outperforms than the semi-symmetric one. As shown in Table 5, this result is also predicted by computing the fitness value of these two patterns for different blur scales.

Table 5. The fitness value (Eq.13) of the asymmetric and semi-symmetric patterns for different blur scales (r = 1..6). Smaller value means better fitness

r	1	2	3	4	5	6
F(Asym.)	38.67	40.33	41.65	42.58	42.19	42.46
F(Semi-sym.)	37.19	39.51	41.99	44.48	42.97	42.6

1. In Canon1100D, images could be taken in 5 different resolution (L, M, S1;S2;S3) that takes images in size 4272×2848, 3088×2056, 2256×1504, 1920×1280 and 720×480, respectively.

The results of our real experiment are similar to the ones obtained in the simulation experiment. However, there is a difference between simulation and real ones. In the real experiment, the range of VIF values are less than values obtained by simulation, while RMSE has nearly the same range of values. It shows VIF is more sensitive to visual effects of deblurred images. This sensitivity causes VIF to have more variation than RMSE. As a result, Q measure is sometimes biased to VIF value. Indeed, if we want to have more emphasis on RMSE, we should study about using a weighted sum in Eq.5.

Experiment 2.

Experiment 1 is repeated for other real scenes in different depths. Because of multiplicity of the studied apertures, the results of only three scenes are shown that contain details in different sizes and include various types of edges. Figure 9 contains a scene with details and letters in various sizes in depth 80^{cm}. Although deblurring result of each pattern has some drawbacks, our patterns provide better results. Figure 10 contains a face with some letters and curve edges in depth 60^{cm}. As shown in Figure10, deblurring results of our proposed pattern provide fewer artifacts. Figure 11 contains a scene in depth 40^{cm}.

For evaluating the deblurring results, a subjective quality assessment is also performed by assigning a score out of 10 (0: lowest, 10: highest quality) to the restored images. To this aim, ten experts evaluated the restored images. Table 6 shows the average of given scores.

Table 5. The average of subjective scores assigned to the deblurring results (Fig. 9-11) of different masks.

	Veera.[3]	Zhou001[2]	Masia[4]	Asym	Sym
Fig. 9 (Depth = 80)	5.5	5.8	5.2	6.3	6.1
Fig. 10 (Depth = 60)	8	7.9	7.8	8.3	8.35
Fig. 11 (Depth = 40)	8.3	8.15	8.1	8.35	8.3

As shown in Table 6, by decreasing the depth, all of the patterns provide acceptable results and just a few drawbacks are seen in some patterns. As a result, the main difference of the available patterns must be studied in deeper scenes in which high frequencies are more attenuated.

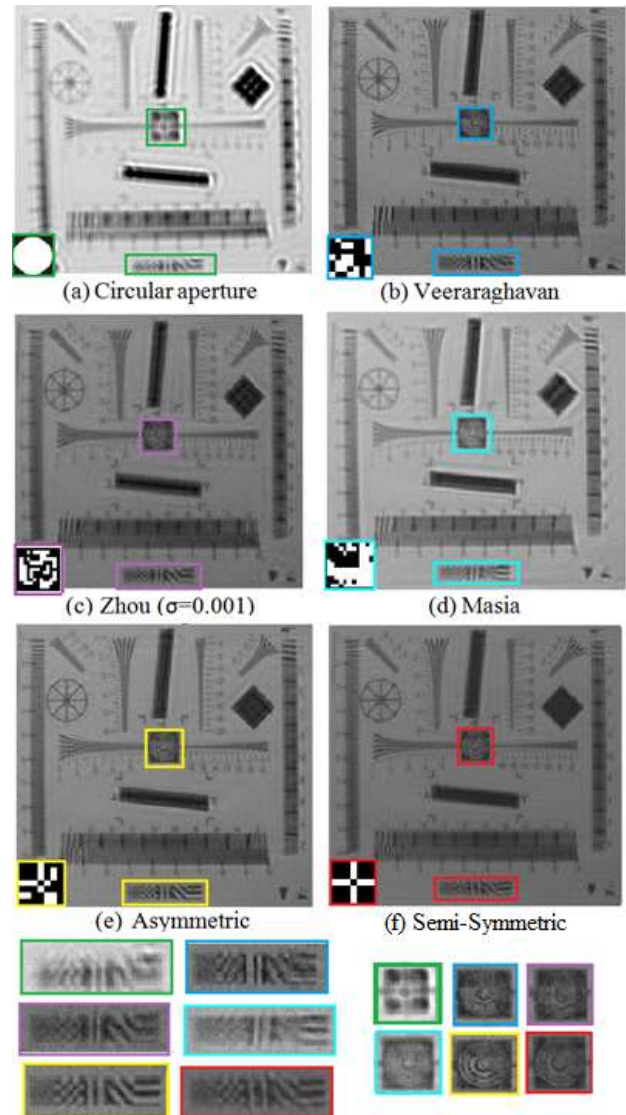


Fig. 7. (a)-(f) Deblurred result of captured images with circular aperture and some other masks in depth 70^{cm}. Bottom-left corner of each image depicts the mask used in each case Last row shows close-up of deblurred images.

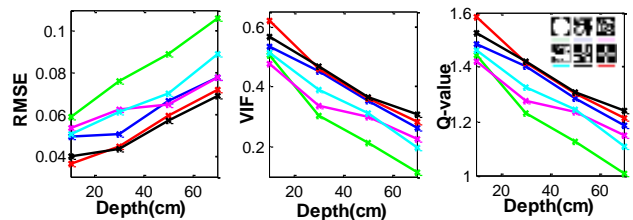


Fig. 8. Deblurring results of CZP resolution chart in 4 different depths. Q-value is computed according to Eq. 4. (Green: Circular aperture, Blue: Veeraraghavan et al. [3], Magenta: Zhou et al. [2], Cyan: Masia et al.[4], Black: Our asymmetric mask, Red: Our semi-symmetric mask.



Fig. 9. Captured image and deblurred result in depth 80cm for five different apertures (a) Veeraraghavan[3], (b) Zhou[2], (c) Masia[4], (d) Asymmetric and (e) semi-symmetric patterns. (f) Close-ups of deblurred images.

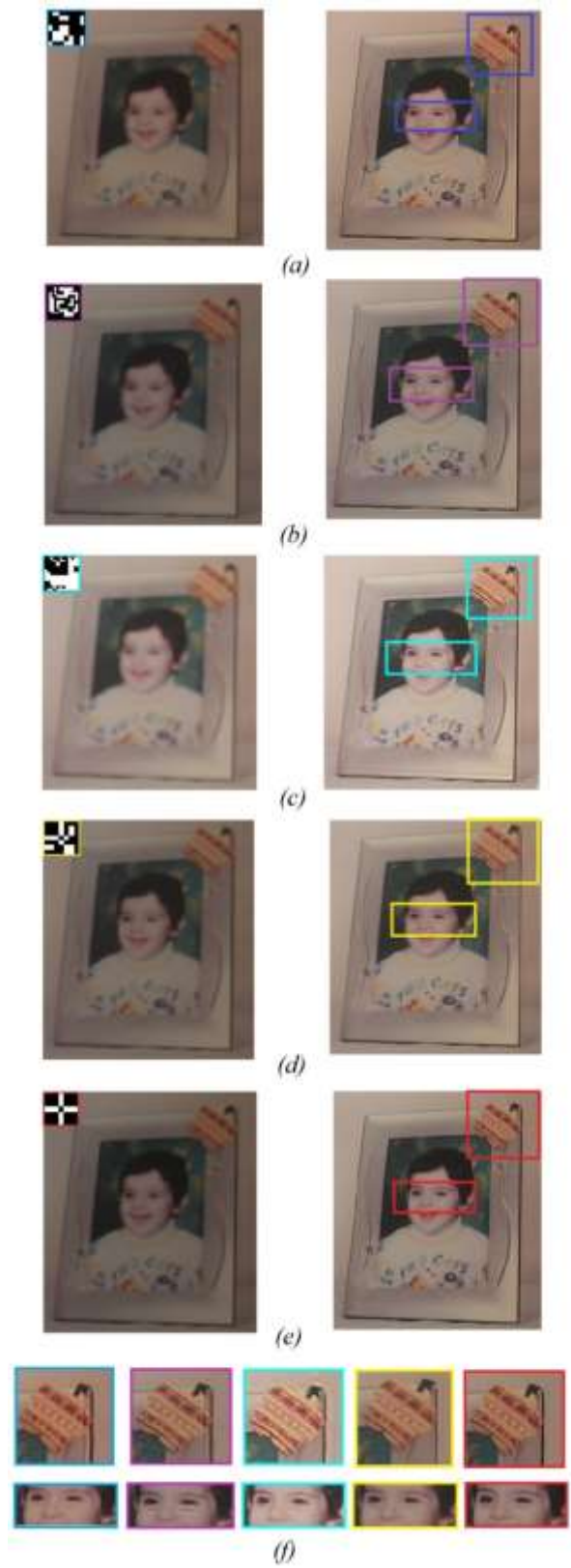


Fig. 10. (a)-(e) Captured images (left) and deblurred results (right) for 5 different aperture patterns in depth 60cm, (f) Close-ups of (a)-(e). Top-left corner of each image depicts the mask used in each case.

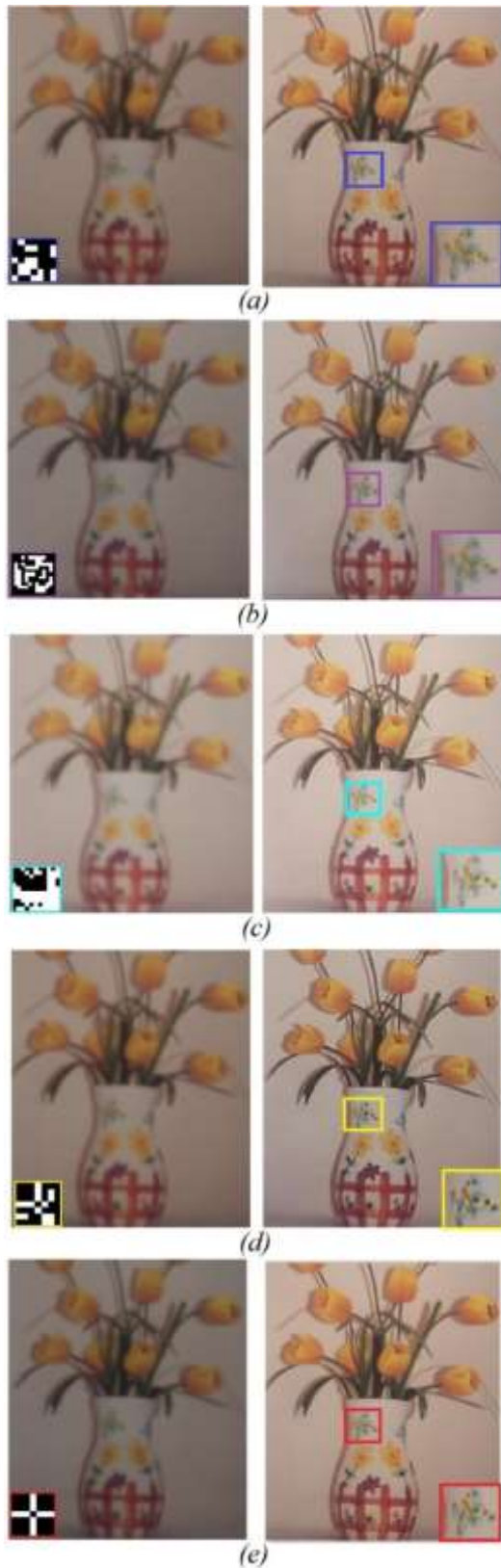


Fig. 11. (a)-(e) Captured images(left) and deblurred results (right) for 5 different aperture patterns in depth 40cm . Bottom- left corner of each image depicts the mask used in each case.

5. Conclusion and Future Work

In this paper, some new criteria are introduced to evaluate coded aperture patterns that are designed for deblurring. They are defined to measure the similarity of the derived filter of a pattern with an all-pass filter. Based on these criteria a new fitness function is proposed to evaluate aperture patterns. The coefficients used in this function are chosen so that the function has the least error in evaluating of a pattern.

To our best knowledge, the first semi-symmetric pattern for deblurring is proposed in this study. Symmetric patterns are rotation invariant. Therefore, most photographers would like to use symmetric apertures if they exist, while all existing masks are asymmetric. Our experiments show that symmetric patterns are slightly less efficient than asymmetric ones, although they provide acceptable results.

It should be mentioned, while we have proposed a semi-symmetric pattern, designing a full-symmetric pattern is still an open problem.

In this research, an aggregate measure including VIF and RMSE is introduced to assess the quality of deblurring results. Our experiments show that the sensitivity of VIF measure is more than RMSE. Therefore, the proposed aggregate measure may be biased to VIF value. Designing a weighted aggregate measure might be investigated in future studies.

References

- [1] K. Mitra, O. Cossairt, and A. Veeraraghavan, "To denoise or deblur: parameter optimization for imaging systems," in *IS&T/SPIE Electronic Imaging*, 2014, pp. 90230G-90230G-6.
- [2] C. Zhou and S. Nayar, "What are good apertures for defocus deblurring?," in *Computational Photography (ICCP)*, 2009 IEEE International Conference on, 2009, pp. 1-8.
- [3] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," *ACM Transaction on Graphics*, vol. 26, no.3, p. 69, 2007.
- [4] B. Masia, L. Presa, A. Corrales, and D. Gutierrez, "Perceptually optimized coded apertures for defocus deblurring," *Computer Graphics Forum*, vol. 31, no.6, pp. 1867-1879, 2012.
- [5] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," *ACM Transactions on Graphics*, vol. 26, no. 3, p. 70, 2007.
- [6] S. Hiura and T. Matsuyama, "Depth measurement by the multi-focus camera," in *Computer Vision and Pattern Recognition*, 1998. Proceedings. 1998 IEEE Computer Society Conference on, 1998, pp. 953-959.
- [7] M. Martinello, "Coded aperture imaging," Heriot-Watt University, 2012.
- [8] A. Sellent and P. Favaro, "Which side of the focal plane are you on?," in *Computational Photography (ICCP)*, 2014 IEEE International Conference on, 2014, pp. 1-8.
- [9] A. Sellent and P. Favaro, "Optimized aperture shapes for depth estimation," *Pattern Recognition Letters*, vol. 40, pp. 96-103, 2014.
- [10] Y. Bando, B.-Y. Chen, and T. Nishita, "Extracting depth and matte using a color-filtered aperture," *ACM Transactions on Graphics*, vol. 27, no.5, p. 134., 2008.
- [11] C. Zhou, S. Lin, and S. K. Nayar, "Coded aperture pairs for depth from defocus and defocus deblurring," *International Journal of Computer Vision*, vol. 93, no. 1, pp. 53-72, 2011.
- [12] Y. Takeda, S. Hiura, and K. Sato, "Fusing depth from defocus and stereo with coded apertures," in *Computer Vision and Pattern Recognition (CVPR)*, 2013 IEEE Conference on, 2013, pp. 209-216.
- [13] A. Chakrabarti and T. Zickler, "Depth and deblurring from a spectrally-varying depth-of-field," in *Computer Vision—ECCV 2012*, ed: Springer, 2012, pp. 648-661.
- [14] A. Ashok and M. A. Neifeld, "Pseudorandom phase masks for superresolution imaging from subpixel shifting," *Applied optics*, vol. 46, no. 12, pp. 2256-2268, 2007.
- [15] S. K. Nayar, "Computational cameras: Approaches, benefits and limits," Technical Rep. 2011.
- [16] C. Zhou and S. K. Nayar, "Computational cameras: Convergence of optics and processing," *Image Processing, IEEE Transactions on*, vol. 20, no. 12, pp. 3322-3340, 2011.
- [17] E. Caroli, J. Stephen, G. Di Cocco, L. Natalucci, and A. Spizzichino, "Coded aperture imaging in X-and gamma-ray astronomy," *Space Science Reviews*, vol. 45, no.3, pp. 349-403, 1987.
- [18] S. R. Gottesman and E. Fenimore, "New family of binary arrays for coded aperture imaging," *Applied optics*, vol. 28, no. 20, pp. 4344-4352, 1989.
- [19] W. T. Welford, "Use of annular apertures to increase focal depth," *Journal of the Optical Society of America*, vol. 50, no. 8, pp. 749-752, 1960.
- [20] M. Mino and Y. Okano, "Improvement in the OTF of a defocused optical system through the use of shaded apertures," *Applied Optics*, vol. 10, no. 10, pp. 2219-2225, 1971.
- [21] P. C. Hansen, J. G. Nagy, and D. P. O'leary, *Deblurring images: matrices, spectra, and filtering*: Siam, 2006.
- [22] P. Campisi and K. Egiuzarian, *Blind image deconvolution: theory and applications*: CRC press, 2007.
- [23] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *Image Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 430-444, 2006.
- [24] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *Image Processing, IEEE Transactions on*, vol. 15, no. 11, pp. 3440-3451, 2006.
- [25] A. Lahoulou, A. Bouridane, E. Viennet, and M. Haddadi, "Full-reference image quality metrics performance evaluation over image quality databases," *Arabian Journal for Science and Engineering*, vol. 38, no. 9, pp. 2327-2356, 2013.
- [26] Y. Liu, J. Wang, S. Cho, A. Finkelstein, and S. Rusinkiewicz, "A no-reference metric for evaluating the quality of motion deblurring," *ACM Transaction on Graphics*, vol. 32, no. 6, p. 175, 2013.
- [27] O. Cossairt, M. Gupta, and S. K. Nayar, "When does computational imaging improve performance?," *Image Processing, IEEE Transactions on*, vol. 22, no. 2, pp. 447-458, 2013.
- [28] K. Mitra, O. S. Cossairt, and A. Veeraraghavan, "A Framework for Analysis of Computational Imaging Systems: Role of Signal Prior, Sensor Noise and Multiplexing," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 10, pp. 1909-1921, 2014.
- [29] Y. Weiss and W. T. Freeman, "What makes a good model of natural images?," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 2007, pp. 1-8.
- [30] Y. Gao, "Population size and sampling complexity in genetic algorithms," in *Proc. of the Bird of a Feather Workshops*, 2003, pp. 178-181.
- [31] <http://www.cs.washington.edu/research/imagedatabase/groundtruth/>

Mina Masoudifar is a Ph.D candidate in computer engineering at Ferdowsi University of Mashhad, Iran. She received the B.Sc and MS degree in computer engineering from Sharif University and Ferdowsi University of Mashhad, respectively. Her research interest is image restoration, computer vision and computational photography.

Hamid Reza Pourreza received the M.Sc degree in electrical engineering and the Ph.D degree in computer engineering from Amirkabir University of Technology, Tehran, Iran, in 1993 and 2002 respectively. He is an Associate Professor with the department of computer engineering, Ferdowsi University of Mashhad, Iran. His research interests include image processing, machine vision and computational imaging. He is senior member of IEEE and one of the founders and an active member of Eye Image Analysis Research Group (EIARG).

Design, Implementation and Evaluation of Multi-terminal Binary Decision Diagram based Binary Fuzzy Relations

Hamid Alavi Toussi

Department of Computer Science, Aarhus University, Aarhus, Denmark
hamid@cs.au.dk

Bahram Sadeghi Bigham*

Department of Computer Sciences, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan, Iran
b_sadeghi_b@iasbs.ac.ir

Received: 25/Jan/2015

Revised: 16/Mar/2015

Accepted: 02/Feb/2016

Abstract

Elimination of redundancies in the memory representation is necessary for fast and efficient analysis of large sets of fuzzy data. In this work, we use MTBDDs as the underlying data-structure to represent fuzzy sets and binary fuzzy relations. This leads to elimination of redundancies in the representation, less computations, and faster analyses. We also extended a BDD package (BuDDy) to support MTBDDs in general and fuzzy sets and relations in particular. Representation and manipulation of MTBDD based fuzzy sets and binary fuzzy relations are described in this paper. These include design and implementation of different fuzzy operations such as max, min and max-min composition. In particular, an efficient algorithm for computing max-min composition is presented. Effectiveness of our MTBDD based implementation is shown by applying it on fuzzy connectedness and image segmentation problem. Compared to a base implementation, the running time of the MTBDD based implementation was faster (in our test cases) by a factor ranging from 2 to 27. Also, when the MTBDD based data-structure was employed, the memory needed to represent the final results was improved by a factor ranging from 37.9 to 265.5. We also describe our base implementation which is based on matrices.

Keywords: Boolean Functions; BDD; MTBDD; Binary Fuzzy Relations; Fuzzy Connectedness; Image Segmentation.

1. Introduction

ROBDDs have been used in hardware community for model checking and circuit verification [1]. It has a variety of applications in other areas as well. For example, it is used in compiler community for efficient points-to analysis. In points-to analysis, it is either used as a compact representation of large sets (points-to sets in this case) [24,25,11] or, the whole analysis is encoded as Boolean functions (or relations) and performed by using Boolean operators (BDD is compact representation for a Boolean function) [2,3,4]. It is also used in image processing [5,6].

Efficient representation of fuzzy sets and relations can be of great importance for analysing large sets of fuzzy data. In this work, design and implementation of a MTBDD based data-structure for representing fuzzy sets and binary fuzzy relations is investigated. Our main idea is to use MTBDDs [7] as the underlying data-structure.

MTBDDs have been used to represent arrays and graphs [7,8]. Clark et al. discussed representation of 2-dimensional arrays and vectors by using MTBDDs. However, they did not provide any implementation. Also they used shadow nodes to simplify their algorithms. Our idea is incorporated into a modern BDD library (BuDDy [9]) without shadow nodes. Shadow nodes increase the size of MTBDDs and thus make the implementation less efficient. Instead, level attribute already presented in the BDD library, is used.

R. Iris Bahar used MTBDDs to perform matrix multiplication and also solve all pairs shortest paths problem [8]. D. Yu. Bugaychenko and I. P. Soloviev proposed MRBDD (Multi-root decision diagram) data-structure to represent integer functions. In their representation, a finite-valued function is represented by a list of k different ROBDDs (or k roots as they suggested) thus an assignment maps to a binary string of 0s and 1s of length k instead of just a 0 or 1. The resulted binary string should be decoded to a certain value. This value is the output of the function for the assignment. The list of ROBDDs which constitute the MRBDD share isomorphic sub-graphs (every sub-graph is also a ROBDD) [12].

In our implementation, because of the way that BuDDy allocates nodes, no two isomorphic ROBDD is ever allocated twice so our work does not just share sub-graphs among a set of ROBDDs belonging to a single MRBDD but among all ROBDDs in memory. This is also discussed in Section 2.2. Generally speaking, sharing is more pervasive in MRBDDs compared to MTBDDs since there are just two terminals instead of a set of terminals. However, implementation of operations on MTBDDs is straightforward because terminals are shown explicitly. This is not the case with MRBDD and for any operation, a correspondence between operation on output values and equivalent operation on binary encoding of the output values must be defined. Summation and multiplication on matrices which are represented by MRBDDs are explained in [12].

* Corresponding Author

We have evaluated our data-structure by solving fuzzy connectedness over different binary fuzzy relations. These relations are obtained from different images. Results are compared with a base implementation which uses 2-dimensional arrays to represent binary fuzzy relations. MTBDD based implementation was 18 – 27× faster when number of distinct membership values that can appear in relations is limited to 11 values. In all cases we have far better memory consumption when MTBDD based implementation is used. See Section 6 for more detail.

Our major contributions in this work are:

- Describing representation and manipulation of fuzzy sets and binary fuzzy relations based on MTBDDs
- Extending BuDDy library to support MTBDDs in general and fuzzy sets and binary fuzzy relations in particular
- Evaluation of our implementation by solving fuzzy connectedness problem

An introduction to ROBDDs, BuDDy and MTBDDs comes in Section 2. The way that BuDDy is extended is discussed in Section 3. Representation and manipulation of MTBDD based fuzzy sets and binary fuzzy relations are discussed in Sections 4 and 5. The empirical evaluation is given in Section 6 and finally, in Conclusion, we discuss possible future directions. A draft version of this work has been published in arXiv [23].

2. Background

2.1 Binary Decision Diagrams

BDD is a data structure for representing Boolean functions compactly. A completely unreduced BDD is shown in Figure 1 which represents the Boolean function $f = x1 .x2 .x3 + x1 .x2 .x3 + x1 .x2 .x3$. Behind some of the nodes, their associated functions are presented.

The representation which is shown in Figure 1 is canonical. However, it is completely inefficient since it takes $O(2^n)$ space to represent a Boolean function with n variables in memory. ROBDD (or BDD for short) tries to address this problem by eliminating redundancies in unreduced BDDs. For eliminating redundancies and having a canonical representation, following two constraints should be always satisfied in any ROBDD:

1. A ROBDD should be ordered, that is, variables should respect a given total order on any path in a ROBDD.
2. A ROBDD should be reduced which means that there are no two sub-graphs in a ROBDD that are isomorphic and also for any node in a ROBDD its low-child should be different from its high-child.

Note that within every node in a ROBDD, level, a pointer to its low child, and a pointer to its high child are saved. Every node of a ROBDD which is also a ROBDD can be identified uniquely by a triple (level, low, high). Figure 2 shows the same Boolean function as in Figure 1 but in reduced form. We can also see this BDD and its associated Boolean function as a set which contains 011, 101 and 111 strings.

It is very common to use the term BDD to refer to ROBDD and we follow this practice in the rest of this paper.

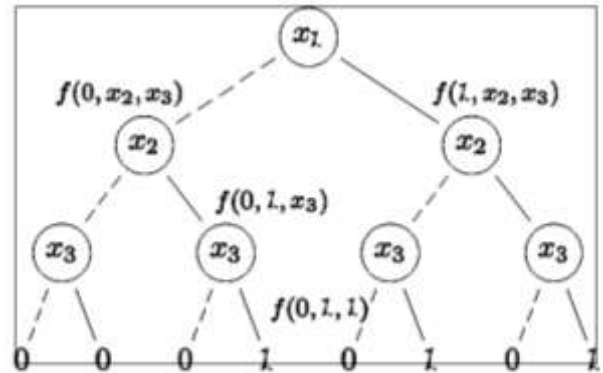


Fig. 1. A completely unreduced BDD which represents the Boolean function $f = \neg x1 .x2 .x3 + x1 .\neg x2 .x3 + x1 .x2 .x3$.

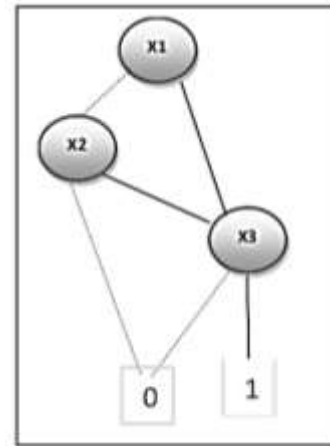


Fig. 2. The reduced version of the BDD previously shown in Figure 1.

2.2 BuDDy

BuDDy [9] is a library for creating and manipulating BDDs. It is written in C and also offers a C++ interface. Since we extended this library, it is useful to know some of its internals which affected our design.

In BuDDy all nodes (BDDs) are stored in an array which is named `bddnodes`. Every slot in this array has four fields namely level, low, high which are used to identify the BDD stored in the slot, and, the fourth field hash which is used to make searching the array more efficient by using hashing.

Every BDD can be uniquely identified by using its level, low and high attributes. In another word, we can associate a triple (level, low, high) with every BDD. At the core of BuDDy is a routine named `bdd_makenode` which is used for allocating BDDs. This routine only creates one entry for every distinct triple in `bddnodes` array and if it is asked to create a triple which is already inserted in `bddnodes`, it simply returns the index of the existing entry. This index represents the BDD in BuDDy. Also, if the triple sent to this routine contains the same value as its low and high, no new BDD will be allocated and the low will be returned.

In this way all the BDDs are always reduced and share sub-graphs that are isomorphic. Sub-graphs of any BDD are BDDs themselves and are allocated only once for any distinct triple. This brings some of the advantages of MRBDDs to our implementation. As described in [12], BDDs which constitute a MRBDD share isomorphic sub-graphs, but in BuDDy and as a consequence in our implementation any two BDDs share isomorphic sub-graphs.

2.3 Apply Operation

BDDs represent Boolean functions so one way to manipulate them is through Boolean operations (**or**, **and**, **xor**, etc). In general, there is a routine (`bdd_apply` in BuDDy) that takes two BDDs (which represent two Boolean functions) and makes a new BDD out of them by applying a Boolean operator. For further details see [13, 9].

2.4 Multi-Terminal Binary Decision Diagrams

A Multi-Terminal Binary Decision Diagram (or MTBDD for short) is a data-structure which has all the features of BDD and it allows more than two terminals. In Sections 3, 4 and 5, we explain how we have extended BuDDy to add support for MTBDDs as well as other functionalities which were needed.

3. Extending BuDDy to Support MTBDDs

In BuDDy, BDD type is defined as `int`. The integer representative of a BDD can be used to index into `bddnodes` array and retrieve its root. However, terminals are not required to be stored in `bddnodes` array explicitly since integers zero and one are reserved to show them. Integers greater than one are used to show non-terminals.

We chose not to define any new type to show MTBDD and simply used integer as their representative to comply with existing design. As a result, integers were also used to show terminals other than zero and one. However, the routine `bdd_makenode` can use any slot with index greater than one in `bddnodes` array to store a BDD (a non-terminal) and returns the slot's index as the BDD's representative. Thus using integers greater than one for showing terminals could introduce new complexities in this routine. To overcome this problem, negative integers were used to show terminals other than zero and one (-1 cannot be used to show a terminal since it indicates an uninitialized slot in `bddnodes`). In this way, all the existing routines continue to work (except `bdd_apply` in cases that it encounters terminals other than zero and one).

The BuDDy library was extended in a generic way so it can be used in similar scenarios. Major routines which are added to the library are `mtbdd_apply` and `mtbdd_maxmin_compose` (`mmc` for short). The former was added to handle maximum and minimum operators for MTBDDs, and the latter is simply a new functionality which was added to do max-min compositions of two binary fuzzy relations. See Subsection 5.1 for further detail.

Floating points were not used to show the membership values since imprecision in floating-points was not acceptable for us and we would like to have fully deterministic results. A C struct which has a field of type `Integer` is used to represent membership values. For example, an instance of this struct with an integer set to 25 shows 0.025 when precision of three digits is used. It may be worth noting that the precision should be known in advance to interpret a membership value. We used three different precisions in our benchmarks. Precision of three digits which can show 1001 different membership values, precision of two digits which can show 101 different membership values and precision of a single digit which can show 11 different membership values (Note that 1 is considered to be a membership value).

4. MTBDDs as Fuzzy Sets

In the MTBDD representation of a fuzzy set, there are as many terminals as there are different membership values in the fuzzy set (including zero and one). Different paths (including those which are reduced or are not represented explicitly) show different members of the fuzzy set. The terminal each path ends at, shows its membership value. In Figure 3, membership value 0.3 is represented by -3 so the MTBDD shows the fuzzy set $\{0.3/0000, 0.3/0001, 0.3/0010, 0.3/0011, 1/0100, 1/0110, 1/1000, 1/1010\}$. Strings 0000, 0001, 0010, ... correspond to numbers 0, 1, 2, ... respectively.

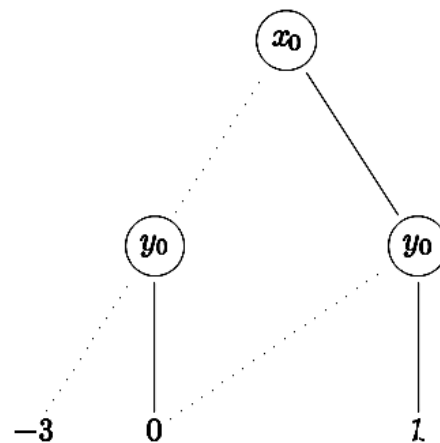


Fig. 3. A MTBDD which represents the binary fuzzy relation $\{1/(1, 1), 0.3/(0, 0)\}$

4.1 Intersection and Union Operations

Maximum and minimum operators were used as fuzzy set intersection and fuzzy set union respectively [14]. The general apply routine which was mentioned in Section 2 takes two BDDs as its operands and another parameter as its operator, then, it applies the operator to the operands. A slightly modified apply routine (`mtbdd_apply`) which handles MTBDDs and maximum/minimum operators was added to the BuDDy library. Our implementation can be easily extended to include other t-norm operators.

5. MTBDDs as Binary Fuzzy Relations

Any binary fuzzy relation has two domains, two disjoint sets of BDD variables are used. Every set of BDD variables is mapped to one of the domains. For example, if a domain has eight objects, three variables would be needed to show all its members (i.e. $2^3 = 8$).

In Figure 4, there are two domains and each one has two objects so all objects can be encoded by using one variable for each domain. Variable $x0$ is used to encode the objects in the first domain and variable $y0$ is used to encode objects in the second domain. Suppose that non-terminal -3 is mapped to 0.3. In this way (0, 1) and (1, 0) are associated with 0 membership value. (1, 1) is associated with 1 membership value and (0, 0) is associated with 0.3 membership value. You can also see it as the 2-dimensional array:

0.3	0
0	1

Since we implemented an algorithm similar to 4-way block-multiplication to compute the max-min composition of two binary fuzzy relations which are represented as MTBDDs, it is desirable to partition each relation into four blocks and access each one in constant time. In order to make this possible, interleaved variable ordering was used [7]. This means that if variables x_i constitute domain x and variables y_i constitute domain y , the ordering of variables would be $x0, y0, x1, y1, x2, y2, \dots$. See Figure 5 for an example. Binary fuzzy relations can be seen as square matrices of size $n \times n$. In a binary fuzzy relation A that $n \neq 2^k$, an identity matrix with smallest possible size is attached to A to meet this requirement. This technique is also used in [7]. This is working for max-min composition since minimum of zero and any other membership value is zero (similar to matrix multiplication and multiplication of zero by other element).

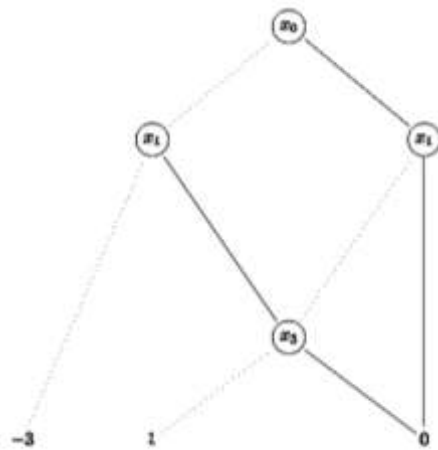


Fig. 4. A MTBDD representing a fuzzy set

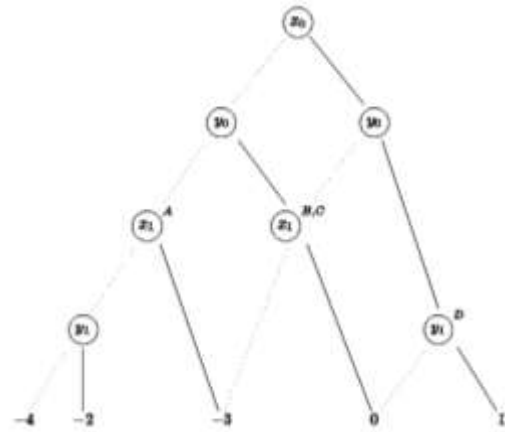


Fig. 5. A binary fuzzy relation represented as a MTBDD. Nodes A, B, C and D show four partitions of this MTBDD. Both B and C correspond to the same node.

5.1 Max-min Composition

In general max-min composition of two binary fuzzy relations $R_1 (D \times D)$ and $R_2 (D \times D)$ is defined as follows:

$$R_3(a, b) = \max_{c \in D} \min(R_1(a, c), R_2(c, b))$$

The max-min composition procedure is similar to block matrix multiplication. A binary fuzzy relation can be viewed as a 2-dimensional matrix. This matrix is partitioned into four sub-matrices (blocks) in the procedure (Figure 5).

max-min composition was implemented as a recursive procedure which is shown in Figure 6. During the recursion, at each call, parameters of the call (MTBDD a and MTBDD b) should be interpreted based on the depth of the recursion which is passed as the third parameter (call_level). This is because, the partitioning does not create four new MTBDDs but returns four sub-graphs of the original MTBDD as its partitions so a hypothetical level (root_level) is assumed. The mentioned hypothetical level indicates the smallest possible level of the resulted partitions (Figure 6). For example, consider the MTBDD shown in Figure 5, four partitions would be created after partitioning namely A, B, C and D. The hypothetical root level for these partitions is two which corresponds to the variable $x1$. This technique (introducing and using a hypothetical root level) avoids creation of new matrices (MTBDDs) and makes the max-min composition procedure more efficient. In order to compute the max-min composition of MTBDDs a and b, we have to call $mmc(a, b, 0)$.

6. Evaluation

To evaluate our representation, we extended the BuDDy library to represent and manipulate binary fuzzy relations by using MTBDDs. Also, we implemented binary fuzzy relations based on two dimensional arrays. The array implementation was used as a baseline (base implementation). Images in our input set are obtained from UIUC image database [15] and The Berkeley Segmentation Dataset and Benchmark [16].

In our experiment, the fuzzy-connectedness problem is solved for different images in the input set. Results of these experiments and further details come next.

6.1 Evaluation Results and Further Details

Results are given in tables 2, 3 and 4. Input relations for our experiments, are affinity relations, which are created from various images. Affinity relation is a symmetric and reflexive fuzzy relation which assigns a membership value to a pair of pixels based on their local properties [17]. We initialized this relation only for pairs of pixels which are neighbour in a given image and membership value for any other pair of pixels in the image is set to zero. This leads to sparsity of affinity relations. The following similarity measure which was used in [18] has been employed to initialize affinity relations. δ is the largest diff (diff is computed for every pair of pixels) and $c_r, c_g, c_b, d_r, d_g, d_b$ show color intensities associated with c and d pixels respectively:

$$\begin{aligned} \text{simil}(c, d) &= 1 - \sqrt{\text{diff}} / \delta \text{ where } \text{diff} \\ &= (c_r - d_r)^2 + (c_g - d_g)^2 + (c_b - d_b)^2 \end{aligned}$$

Images used for creating affinity relations are shown in Table 1. The first one is from UIUC image database [15] and the next three images are obtained from The Berkeley Segmentation Dataset and Benchmark [16]. The last image is a synthetic image from reference [18]. All experiments are run on a machine with 2.8 GHz Intel CPU and 4 GB of RAM running Fedora 14.

In the rest of this Section, problem of fuzzy-connectedness is investigated. Input to this algorithm is an affinity relation which is extracted from an image, and final output is a relation that assigns a membership value to every pair of pixels in the image. The output can be used to create different clusters of pixels [17]. The final goal of fuzzy connectedness is to calculate FC relation which is a reflexive, symmetric and transitive relation. It is basically max-min transitive closure of the initial affinity relation. The FC relation is a full binary fuzzy relation. It assigns a membership value to every pair (c, d) . This value is the maximum strength of all possible paths from c to d . The strength of a path is the smallest membership value along the path. FC relations are obtained by computing max-min transitive closure of affinity relations.

We implemented two different versions to compute the transitive closure. Our base implementation used two dimensional arrays to represent binary fuzzy relations and, it employed Floyd-warshall algorithm as shown in Figure 7. n is the number of pixels in the image. c stores the affinity relation initially and represents fuzzy connectedness (FC) relation at the end. Our second implementation used MTBDDs to represent binary fuzzy relations and, it computed FC relation by using Repeated Squaring algorithm as shown in Figure 8. Affinity relation is the input to this algorithm and at the end, res would be a MTBDD that represent the fuzzy connectedness relation (FC). Because of the MTBDD special structure we could

not use the Floyd-warshall algorithm in conjunction with this data-structure efficiently.

Table 2 shows running-times of our base and MTBDD based programs which compute max-min transitive closure of affinity relations obtained from our test images. Three versions are shown in the table. Column base shows the base implementation and the other two columns under mtbdd show the MTBDD based implementation with one and two digits of precision. The MTBDD based implementation is significantly faster than the base implementation when precision of one digit is used (it is $18 - 27\times$ faster). Compared to base implementation, using 2-digits precision improved running time in all cases, but one, which was our smallest image (40×27). In this particular case all running times were under one minute. In other cases MTBDD based implementation with two digits precision is faster by a factor ranging from 2 to 7.

When images are larger and their pixels are more homogeneous, MTBDD based implementation becomes faster relative to the base implementation. Note that running-time in base implementation only depends on size of input image (i.e. number of pixels). In contrast, MTBDD based implementation's running-time depend both on size of image and values stored in every pixel of the image. For example, running-times for image3 and image4 are the same when base implementation is used but it takes less time for MTBDD based implementation to compute the transitive closure when it takes image4 as input. This is because image4 leads to MTBDDs which are more compact (this is described in more detail in the next paragraph).

As described in Section 5, different paths in MTBDD representation of a relation show different pairs in the relation. More commonalities among paths lead to more reductions and, a more compact MTBDD representation of the relation. Computations on a smaller MTBDD take less time. Two different images even with the same size result in different MTBDDs with different sizes. An image which results in a MTBDD with more commonalities in its paths occupies less memory and results in fewer computations in the MTBDD based solver (Sparsity in input relation and also images with homogeneous pixels leads to more compact MTBDDs).

In Table 3, number of entries in FC relations, number of terminals and number of nodes in MTBDD representation of these relations are shown. Algorithms which were used to compute FC relations are show in Figure 7 (base version) and Figure 8 (MTBDD version).

Number of nodes in MTBDD base implementation is extremely lower than number of entries in base implementation which leads to a far improved memory consumption when MTBDDs are used to represent binary fuzzy relations. Every array entry in base implementation is three bytes (two bytes for a short integer and one byte for a flag) and the size of every bddnode is 20 bytes [9]. In Table 4 size of array and MTBDD based representation of FC relations are shown (in KB). The column size (r) indicates number of entries in the array representation of the relation (base implementation), column size (array) shows the amount of memory allocated for representing

arrays in the base implementation and, the other two columns indicate the amount of memory allocated for representing MTBDDs in the MTBDD based implementation (precision of one and two digits). Considering this table, MTBDD based representation takes $37.9 - 265.5 \times$ less memory than the array representation depending on the input image.

It is also noteworthy that shape and number of nodes in BDDs (and MTBDDs as well) also depend on variable ordering beside data they are representing since different

variable ordering leads to different paths with different degree of sharing. However, a fixed variable ordering is used in our implementation.

Number of terminals is also shown in Table 3. This gives the number of distinct (hard) clusterings that can be obtained from the resulted relation. When number of terminals is limited to 11 (1-digit precision), it is five or six and in the other case, when number of terminals is limited to 101 (2-digit precision), it is usually around 25.

```

procedure mmc(a, b, callLevel)

  if both a and b are terminals then return min(a, b)
  if (r = mmc_cache[a, b, callLevel]) != NULL then return r
  if (r = mmc_cache[b, a, callLevel]) != NULL then return r

  rootLevel = callLevel + 2

  partition a into sa[0], sa[1], sa[2] and sa[3] based on rootLevel
  partition b into sb[0], sb[1], sb[2] and sb[3] based on rootLevel

  t1 = mmc(sa[0], sb[0], callLevel + 1)
  t2 = mmc(sa[1], sb[2], callLevel + 1)
  l = mtbdd_apply(t1, t2, mtbddopFuzzymax)
  t1 = mmc(sa[0], sb[1], callLevel + 1)
  t2 = mmc(sa[1], sb[3], callLevel + 1)
  h = mtbdd_apply(t1, t2, mtbddopFuzzymax)
  L = bdd_nakenode(rootLevel + 1, l, h)

  t1 = mmc(sa[2], sb[0], callLevel + 1)
  t2 = mmc(sa[3], sb[2], callLevel + 1)
  l = mtbdd_apply(t1, t2, mtbddopFuzzymax)
  t1 = mmc(sa[2], sb[1], callLevel + 1)
  t2 = mmc(sa[3], sb[3], callLevel + 1)
  h = mtbdd_apply(t1, t2, mtbddopFuzzymax)
  H = bdd_nakenode(rootLevel + 1, l, h)

  r = bdd_nakenode(rootLevel, L, H)
  mmc_cache[a, b, callLevel] = r
  return r

end procedure

```

Fig. 6. Max-min composition (mmc) routine in pseudo code. a and b are MTBDDs and callLevel indicates the depth of the recursion.

```

for k = 1 to n do
  for i = 1 to n do
    for j = 1 to n do
      c[i, j] = max(c[i, j], min(c[i, k], c[k, j]))

```

Fig. 7. Using Floyd-warshall algorithm to compute max-min transitive closure of affinity relation.

```

res = affinity
repeat
  old = res
  res = mmc(res, res, 0)
until res == old

```

Fig. 8. Using MTBDDs and Repeated Squaring algorithm to compute max-min transitive closure of affinity relation.

Table 1. Images which are used for creating affinity relations.






Image	Image name	Size
	<i>Image0</i>	80x65
	<i>Image1</i>	40x27
	<i>Image2</i>	60x40
	<i>Image3</i>	90x60
	<i>Image4</i>	90x60

Table 2. Running times of Fuzzy Connectedness problem solver for both base and MTBDD based implementations (in seconds).

Image	Base	mtbdd 1	mtbdd 2
<i>Image0</i>	3574	134.61	2236
<i>Image1</i>	32.01	1.78	41.90
<i>Image2</i>	350.72	13.88	285.48
<i>Image3</i>	4000	214.64	1898.21
<i>Image4</i>	same as image3	148.95	533.52

Table 3. Number of entries in FC relations and number of nodes which are allocated to represent the relation in its MTBDD representation

Image	Size(r)	mtbdd terminals 1	mtbdd terminals 2	Nodes in mttbdd 1	Nodes in mttbdd 2
<i>Image0</i>	27040000	6	28	25683	216461
<i>Image1</i>	1166400	6	25	2212	30753
<i>Image2</i>	5760000	5	27	4007	87306
<i>Image3</i>	29160000	5	26	16469	581770
<i>Image4</i>	same as image3	5	10	34995	57439

References

- [1] E. Clarke, O. Grumberg, and D. Long. "Symbolic Model Checking for Sequential Circuit Verification." In IEEE Transactions on Computer Aided Design, 1994.
- [2] Marc Berndt, Ondřej Lhoták, Feng Qian, Laurie Hendren, and Navindra Umanee. "Points-to analysis using BDDs." In Proceedings of the ACM SIGPLAN 2003 Conference on Programming Language Design and Implementation, 2003, pp 103–114.
- [3] John Whaley and Monica Lam. "Clonning-based context-sensitive Pointer Alias analysis using Binary Decision Diagrams." In Proceeding of PLDI, 2004.
- [4] Ondřej Lhoták, Stephen Curial, and Jos'e Nelson Amaral. "Using XBDDs and ZBDDs in points-to analysis." Software, Practice and Experience, vol 39, Issue 2, pp 63–188, 2009.
- [5] Watis Leelapatra, Kanchana Kanchanasut, and Chidchanok Lursinsap. "Displacement BDD and geometric transformations of binary decision diagram encoded images." Pattern Recognition Letters, vol 29, Issue 4, pp 438–456, March 2008.
- [6] Mike Starkey, Randy Bryant, and Y Bryant. "Using ordered binary-decision diagrams for compressing images and image sequences." Technical report, CMU-CS, 1995.
- [7] Edmund M. and Fujita Clarke, M., McGeer, P C., McMillan, K., Yang, J C-Y, and X Zhao. "Multi-Terminal Binary

Table 4. Size of array and MTBDD representation in KB

Image	size(r)	size(array)	size(mttbdd 1)	size(mttbdd 2)
<i>Image0</i>	27040000	81120	513	4329
<i>Image1</i>	1166400	3499	44	615
<i>Image2</i>	5760000	17280	80	1746
<i>Image3</i>	29160000	87480	329	11635
<i>Image4</i>	29160000	87480	699	1148

7. Conclusion

In this work, we designed and implemented a MTBDD based data-structure to represent fuzzy sets and relations. Also, the BuDDy library was extended to support MTBDDs, and it was employed to implement our idea. Promising results were obtained in evaluation of our method. In particular, considering the fuzzy connectedness problem and compared to the base implementation, when MTBDD based implementation was used, the running-time was improved by a factor ranging from 2 to 27, and, when the memory needed to represent final results was improved by a factor ranging from 37.9 to 265.5.

In the future we would like to apply our data-structure to other problems in fuzzy systems which involve manipulating binary fuzzy relations and fuzzy sets, specially, problems with very large sets of fuzzy data such as the use of fuzzy sets in data mining, approximate reasoning and information retrieval based on fuzzy logic [19, 20, 21, 22].

Extending our current implementation to a framework for research in fuzzy systems is another direction we would like to follow. In particular, researchers would be able to add t-norms operators of their interest, and, design and run new experiments on top of our framework.

- Decision Diagrams: An Efficient Data-Structure for Matrix Representation.” *Formal Methods in System Design*, 1997.
- [8] R. I. Bahar, E. A. Frohm, C. M. Gaona, E. Macii, A. Pardo, and F. Somenzi. “Algebraic Decision Diagrams and Their Applications. *Formal Methods in System Design*,” 10, 1997.
- [9] Jorn Lind-Nielsen. *BuDdy* library. <http://sourceforge.net/projects/buddy/>, 2002.
- [10] D. Bugaychenko. “On application of multi-rooted binary decision diagrams to probabilistic model checking. In *Verification, Model Checking, and Abstract Interpretation*,” pp 104–118. Springer, 2012.
- [11] Vaclav Dvorak. “Branching program-based programmable logic for embedded systems.” In *Proceedings of ICONS 2012*, pp 109–115. International Academy, Research, and Industry Association, 2012.
- [12] D. Yu. Bugaychenko and I. P. Soloviev. “Application of multiroot decision diagrams for integer functions.” *MATHEMATICS*, vol 43, Issue 2, pp 92–97, 2010.
- [13] Randal E. Bryant. “Graph Based Algorithm for Boolean function manipulation.” In *IEEE Transactions on Computers*, 1985.
- [14] L. A. Zadeh. “Fuzzy Sets.” *Information and Control*, pp 338–353, 1965.
- [15] Shivani Agarwal. “UIUC Image Database for Car Detection.” <http://cogcomp.cs.illinois.edu/Data/Car/>, April 2002. Accessed: 2012-08-20.
- [16] D. Martin, C. Fowlkes, D. Tal, and J. Malik. “A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics.” In *Proceeding of 8th International Conference on Computer Vision*, vol 2, pp 416–423, July 2001.
- [17] Jayaram K. Udupa and Punam K. Saha. *Fuzzy Connectedness and Image Segmentation*. In *Proceeding of the IEEE*, vol 91, pp 1649-1669, 2003.
- [18] Pedro F. Felzenszwalb. “Efficient Graph-Based Image Segmentation.” *Journal of Computer Vision*, vol 59, Issue 2, pp 167-181, 2004.
- [19] M. Delgado, N. Mann, M. Martn-Bautista, D. Snchez, and M. Vila. “Mining fuzzy association rules: An overview.” *Soft Computing for Information Processing and Analysis*, vol 164, pp 351–373, 2005.
- [20] Ulrich Bodenhofer, Eyke Hllermeier, Frank Klawonn, and Rudolf Kruse. “Special issue on soft computing for information mining.” *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, vol 11, pp 397–399, 2007.
- [21] Amel Borgi and Herman Akdag. “Knowledge based supervised fuzzy-classification: An application to image processing.” *Annals of Mathematics and Artificial Intelligence*, vol 32, Issue 1, pp 67–86, 2001.
- [22] Ariel Gmez, Carlos Len, Jorge Roper, Alejandro Carrasco, and Joaquin Luque. Sabio. “Soft agent for extended information retrieval.” *Applied Artificial Intelligence*, vol 27, pp 249–277, 2013.
- [23] Hamid A. Toussi and Bahram Sadeghi Bigham, Design, Implementation and Evaluation of MTBDD based Fuzzy Sets and Binary Fuzzy Relations, preprint arXiv:1403.1279 [cs.DS], [Online]. Available: <http://arxiv.org/abs/1403.1279>
- [24] Hamid A. Toussi and Abbas Rasoolzadegan, “Flow-sensitive points-to analysis for Java programs using BDDs,” In *Proceeding of 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp.380-386, 29-30 Oct. 2014.
- [25] Ben Hardekopf and Calvin Lin. “Semi-sparse flow-sensitive pointer analysis,” In *ACM SIGPLAN Notices*, vol. 44, no. 1, pp. 226-238. ACM, 2009.

Hamid Alavi Toussi obtained the M.Sc in Computer Science in 2011 from University of Sistan and Baluchestan, Zahedan, Iran. He also holds a B.Sc in Computer Engineering (obtained in 2009 from Islamic Azad University of Mashhad, Iran). Currently, he is a Ph.D student in Computer Science at Aarhus University (Computer Science department), working on program analysis for web applications.

Bahram Sadeghi Bigham is an Assistant Professor in Computer Sciences at the Institute for Advanced Studies in Basic Sciences (IASBS). His research interests are in the areas of Medical Applications of AI, Computational Methods, Data Mining, and Robotics. Prior to arriving at IASBS, Dr. Sadeghi worked as a Postdoctoral Fellow at the University of Cardiff in the School of Computer Science. In June 2008, He completed his Ph.D at Amirkabir University of Technology (Tehran Polytechnic), where he also completed a M.Sc in 2000. His B.Sc is from University of Birjand in Mathematics.

Unsupervised Segmentation of Retinal Blood Vessels Using the Human Visual System Line Detection Model

Mohsen Zardadi*

Department of Electrical and Computer Engineering, Birjand University, Birjand, Iran
zardadi@birjand.ac.ir

Nasser Mehrshad

Department of Electrical and Computer Engineering, Birjand University, Birjand, Iran
nmehrshad@birjand.ac.ir

Seyyed Mohammad Razavi

Department of Electrical and Computer Engineering, Birjand University, Birjand, Iran
smrazvi@birjand.ac.ir

Received: 10/Sep/2015

Revised: 30/Apr/2016

Accepted: 16/May/2016

Abstract

Retinal image assessment has been employed by the medical community for diagnosing vascular and non-vascular pathology. Computer based analysis of blood vessels in retinal images will help ophthalmologists monitor larger populations for vessel abnormalities. Automatic segmentation of blood vessels from retinal images is the initial step of the computer based assessment for blood vessel anomalies. In this paper, a fast unsupervised method for automatic detection of blood vessels in retinal images is presented. In order to eliminate optic disc and background noise in the fundus images, a simple preprocessing technique is introduced. First, a newly devised method, based on a simple cell model of the human visual system (HVS) enhances the blood vessels in various directions. Then, an activity function is presented on simple cell responses. Next, an adaptive threshold is used as an unsupervised classifier and classifies each pixel as a vessel pixel or a non-vessel pixel to obtain a vessel binary image. Lastly, morphological post-processing is applied to eliminate exudates which are detected as blood vessels. The method was tested on two publicly available databases, DRIVE and STARE, which are frequently used for this purpose. The results demonstrate that the performance of the proposed algorithm is comparable with state-of-the-art techniques.

Keywords: Retinal Vessel Segmentation; Simple cell Model; DRIVE Database; STARE Database.

1. Introduction

Retinal blood vessel segmentation provides information for diagnosis, treatment, and evaluation of various cardiovascular and ophthalmologic diseases such as hypertension, diabetes and arteriosclerosis [1]. Various features of retinal blood vessels such as length, width, and tortuosity guide ophthalmologists to diagnose and/or monitor pathologies of different eye anomalies [2-4]. Automatic segmentation of retinal blood vessels is the first step in the development of a computer-assisted diagnostic system. A large number of methods and algorithms that have been published are related to retinal blood vessel segmentation [5]. Each of these methods have their own merits and shortcomings. The algorithms in this field can be classified into techniques based on match filtering, pattern recognition, morphological processing, multiscale analysis and vessel tracking.

As with the processing of most medical images, the signal noise, drift in image intensity, and lack of image contrast cause significant challenges in the extraction of blood vessels. The vessels can be expected to be connected and form a binary tree-like structure. However, the shape, size, and local grey level of blood vessels can vary and background features may have similar attributes to vessels.

The vessel intensity profiles approximate to a Gaussian shape, or a mixture of Gaussians. Therefore, Gabor filters, which are a multiplication of Gaussian and cosine functions, may be a good approximation of the vessel intensity profiles. Gabor filters are also utilized to model simple cells in the primary visual cortex. The simple cells in the human visual system respond vigorously to an edge or a line of a given orientation and position. It can be expected that a computational model of a simple cell by a Gabor filter may extract blood vessels effectively.

In this paper, we offer an unsupervised approach for retinal blood vessel segmentation in fundus images. Our method was inspired by operations of the human visual system in perception of edges and lines at different directions. This paper is organized as follows: a review of other published vessel segmentation solutions in section two, a presentation of the proposed method in section three, results and comparisons with other existing methods in section four, and finally, the main conclusions of this work in section five.

2. Related Work

Retinal blood vessel segmentation methods can be divided into two broad categories: unsupervised and

* Corresponding Author

supervised methods. Due to the fast algorithms of unsupervised methods, they are generally preferred over supervised methods for medical decision support systems and real time applications. As the proposed method is also an unsupervised method, it will be more focused on in the review section of this paper.

2.1 Unsupervised Methods

Unsupervised methods perform vessel segmentation without any prior labeling information to decide whether a pixel belongs to a vessel, or not. A match filtering technique is one of the common unsupervised approaches in retinal vessel segmentation [6]. Matched filtering convolves the image with multiple matched filters for detection of blood vessels. The matched filter was first introduced by Chaudhuri et al. [7]; they proposed a two-dimensional linear kernel with a Gaussian profile to detect the blood vessels. The kernel is rotated in 12 directions and the maximum response is selected for each pixel. This method was improved by Hoover et al. [7] who combined local and region-based properties of the vessels. They employed a thresholding technique with iterative probing for each pixel. Gang et al. [8] also extended Chaudhuri et al.'s method by an amplitude-modified second order Gaussian filter. Gang et al. optimized the matched Gaussian filter by the vessel width. To summarize, all of these matched filtering methods suffer from low signal-to-noise ratio (SNR).

There are some unsupervised segmentation methods which improved accuracy and/or speed of the segmentation. Cinsdikici and Aydin in [9] present a combination of ant colony optimization algorithm and a hybrid model of the matched filter. They employ some preprocessing techniques and combine matched filter results with an ant colony algorithm to extract blood vessels. High computational cost and set parameters are the main drawbacks of this approach; however Cinsdikici and Aydin increased the accuracy of match filters.

In order to address high computational costs, Amin and Yan proposed a high speed vessels detection algorithm which extracted blood vessels in 10 seconds for each retinal image [10]. They used a phase congruency to enhance the vessel intensity in the algorithm. The phase congruency of an image is computed by a Log-Gabor wavelet. Afterwards, blood vessels are segmented by using a threshold probing technique on phase congruency response images. In the Amin and Yan method, several phase congruency images should be obtained from a single retinal image for different parameters and the best one is determined based on the ROC curve. The phase congruency image that produces the highest area under the ROC curve is considered an optimum image; and related parameters are recorded. As a result of high computational costs, their method suffers from a complex parameter adjustment procedure.

Zana and Klein [11] presented a different approach to extract the vessels. They used mathematical morphology and curvature evaluation in a noisy environment. Their

method is based on four steps: noise reduction, linear pattern with Gaussian-like profile improvement, cross-curvature evaluation and linear filtering. Like other morphological processes, Zana and Klein's method depends heavily on the length of linear structure elements, and causes difficulties with highly tortuous vessels.

Fraz et al. [12] combine vessel centerlines with bit plane slicing to identify the vessel patterns in the retina. Fraz et al. used an orientation map to segment blood vessels. The orientation map of vessels is generated by using a multidirectional top-hat operator with a linear oriented structuring element which emphasizes the vessels in the particular direction. Then, significant information is obtained from the greyscale image using bit plane slicing. Finally, the vessels map is acquired by a combination of the vessel centerlines and the orientation maps. In this method, the vessel centerlines, vessel shape, and orientation maps play crucial roles in the segmentation of the vascular tree. Therefore, because of the prominent light reflection of some vessels like arterioles, and a mismatch between the highest local intensity across the blood vessels and the vessel center, this approach is less suited for all blood vessels in the retinal images. Mendonca et al. [13] also propose a vessel centerline detection in combination with multiscale morphological reconstruction. The vessel centerlines are generated by applying Difference of Offset Gaussian (DoOG) filter. Vessel highlighting is acquired by a modified top hat operator with variable size circular structuring elements. They employ multiscale morphological reconstruction to obtain binary image.

Lam and Yan, on the other hand, propose a vessel segmentation algorithm based on the divergence of vector fields [14]. They locate vessel-like objects by using Laplacian operator and pruning noisy objects based on the centerlines. Consequently, the vessel-like patterns which are far away from the centerlines are removed. Although Lam and Yan's method is almost robust, it is a time-consuming approach. It requires around 25 minutes to produce the vessels for a single retinal image [14]. Recently, Lam et al. have presented a vessel profile model to detect blood vessels [15]. This algorithm is based on regularization-based multi-concavity modeling. A differentiable concavity measure on perceptive space is designed to extract blood vessels in retinal images with bright lesions. A line-shape concavity measure is also presented to distinguish dark lesions from the vessels. The vessels are obtained by a lifting technique. A main disadvantage of Lam et al.'s method is high computational cost. It takes, on average, 13 minutes to extract blood vessels.

Al-Diri et al., alternatively, present an active contour model for segmenting retinal vessels [16]. They detect blood vessels by growing a 'Ribbon of Twins' active contour model which employs two pairs of contours to capture each vessel edge. Initially, a tramline filter is used to locate an initial set of potential vessel segment centerline pixels. Then, the segment growing algorithm converts the tramline image into a set of segments.

Finally, a junction resolution method extends the discrete segments and resolves various crossings and junctions. Al-Diri et al. gained good results, but the method suffers from high computational complexity.

2.2 Supervised Methods

Due to the limited success of unsupervised methods in achieving an acceptable output, some researchers have focused on development of supervised algorithms [17-22]. In supervised methods, observer data (gold standard) is firstly used for the classification of vessels based on given features. Then, a feature vector is required for each pixel. Lastly, a classifier is needed to classify each pixel as a vessel pixel or a non-vessel pixel such as k-nearest neighbors (KNN), feed-forward NN and Bayesian classifier.

All the supervised methods use a variety of approaches to calculate the feature vector and also to classify each pixel. Niemeijer et al. [19] introduce a long feature vector (31-Dimension) for each pixel consisting of the Gaussian filter and its derivatives at five scales. They also compare three classifiers and they show the performance of k-nearest neighbors (KNN) classifier is superior for all experiments. Staal et al. [20] also employed the same classifier, however they utilize ridge profiles to compute features for each pixel. On the other hand, some methods use a short feature vector. For instance, Soares et al. [21] utilize a six-D feature vector and Marin et al. [22] introduce a 7-D feature.

Although supervised methods are robust for many retinal images, they are costly in terms of processing time. They also need ground truth data that are already classified, which may not be available in real life applications. Therefore, supervised approaches are not as common as unsupervised algorithms for medical decision support systems and real time applications. As a result, we focused on unsupervised methods in this research. In this paper, we propose a new unsupervised algorithm based on a directional line sensitivity model of the human visual system to detect blood vessels in the retinal fundus images in almost real time.

3. The Proposed Method

Fundus images contain Red, Green, and Blue (RGB) images of the retina. The green channel and grey-level images provide the best vessel-background contrast of the RGB-representation, while the blue channel offers poor dynamic range and the red one is the brightest colour channel and has low contrast. Therefore, grey-level and green channel images are the best choices for image segmentation. In this research, we employed grey-level retinal images. A flow chart of our method is shown in Fig. 1 which consists of four main steps. These steps will be described in detail in the following sections.

3.1 Preprocessing

Retinal fundus images are not uniform images. The contrast of the retinal fundus images tends to be bright in the centre and diminish at the side, hence preprocessing is

essential for minimizing this effect and to have a more uniform image. We introduce a simple technique to minimize the drift in image intensity and lack of image contrast, and generate a uniform image by using a median filter.

The principal function of median filters is to force points with distinct intensity to be more similar to their neighbors. Median filters are quite popular because they provide excellent noise-reduction (impulse noise) capabilities, with less blurring than linear smoothing filter of similar size. Median filters replace the value of a pixel by the median of the grey levels in the neighborhood of that pixel. Therefore, a median filter can be used to achieve the estimation of the background image and location of optic disc. The optic disc is a round area in the fundus images where retinal nerve fibers collect to form the optic nerve. To achieve this goal, the median filter must be large enough in size to remove blood vessels as a noise. The experimental results demonstrate that in 576×720 fundus images, employing a 25×25 median filter will remove blood vessels from the grey-level image and the background and optic disc will appear.

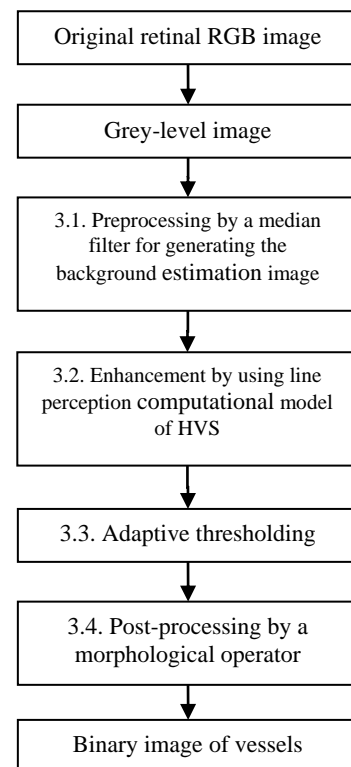


Fig. 1. A flow chart of the proposed method

If $i(x,y)$ is defined as a luminance distribution of original RGB input image and $g(x,y)$ is defined as a grey-level image of $i(x,y)$, then the background, $b(x,y)$, of a grey-level image is obtained by median filter with size of window 25×25 . Fig. 2 (a) depicts an input original RGB image from a DRIVE database as (x,y) , and Fig. 2 (b and c) illustrates the grey level of $i(x,y)$ and the output of median filter (i.e. $b(x,y)$), respectively. This clearly shows that the presented filter with the appropriate size is a good approximation of the background in retinal fundus images. Once the background is computed, a uniform

image is acquired by subtracting the background image from the grey-level image:

$$U(x, y) = g(x, y) - b(x, y) \quad (1)$$

This equation generates a uniform image. Fig. 2(d) shows the $U(x, y)$.

In order to enhance the blood vessel intensity, we need an image with vessel pixels containing brighter intensity than background pixels. This can be generated by using negative transformation:

$$s(x, y) = 255 - U(x, y) \quad (2)$$

Finally, to avoid processing the black border and corners, a binary mask is applied to $s(x, y)$. These two last steps are shown in Fig. 2 (e and f). Through the proposed preprocessing approach, a uniform and normalized image of the retina with vessel pixels brighter than non-vessel pixels is generated. In the next section, a novel vessel enhancement technique to increase vessel's intensity is presented.

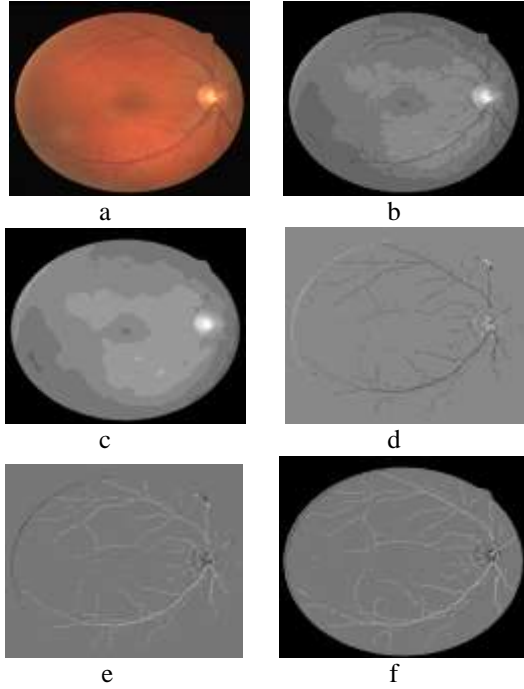


Fig. 2 Preprocessing steps (a) original RGB retinal image (b) grey-level image (c) result of the median filtering (d) result of subtracting background image from grey-level image (e) negative image (f) result of the preprocessing step.

3.2 Vessels Enhancement by Computational Model of a Simple Cell

Due to the poor local contrast of blood vessels, intensity of vessel pixels must be enhanced. The proposed method was motivated by neurons which respond to line and edge in the primary visual cortex. It is not feasible to build a computational model of HVS for image processing applications directly from the physiology of the HVS, due to its tremendous complexity. However, Computational models with different aspects of HVS were developed, aiming at observations from

psychovisual experiments or sequential processing of the visual information in different layers of the HVS [22-27].

Hubel and Wiesel [28] identified two main classes of neuron which they called simple and complex cells. They proved that the majority of neurons in primary visual cortex reveal orientation selectivity [24-26]. Typically, such a neuron would react to a line or an edge of a given orientation in a given area of the visual field, called its receptive field (RF). Generally, simple cells are neurons which respond to an edge or growing/declining line, while neurons which do not react are called complex cells.

The computational models were extended based on simulation of the cell operation [23]. Simple cells are typically modeled by linear spatial summation followed by half-wave rectification. A family of two-dimensional Gabor functions was proposed as a model of the receptive field of simple cells [23,29]. A Gabor filter is a linear and local filter, and its kernel is multiplication of Gaussian and cosine functions. For an input image with luminance distribution of $s(x, y)$, a simple cell's response compute by convolution [27]:

$$S_{\sigma, \lambda, \theta_k, \varphi}(x, y) = h_{\sigma, \lambda, \theta_k, \varphi}(x, y) * s(x, y) \quad (3)$$

$$h_{\sigma, \lambda, \theta_k, \varphi}(x, y) = \cos\left(\frac{2\pi}{\lambda} \tilde{x} + \varphi\right) e^{-\frac{\tilde{x}^2 + \gamma^2 \tilde{y}^2}{2\sigma^2}} \quad (4)$$

$$\begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} \cos \theta_k & \sin \theta_k \\ -\sin \theta_k & \cos \theta_k \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\theta_k = \frac{(k-1)\pi}{N_\theta} \quad \text{for } k = 1, 2, \dots, N_\theta \quad (5)$$

Where θ_k is the preferred orientation of a simple cell's response, $1/\lambda$ is spatial frequency, and N_θ is the number of total preferred orientations. Ellipticity of the receptive field and its symmetry with respect to the origin are controlled by constant parameter γ and angle parameter $\varphi \in (-\pi, \pi]$, respectively. The width of the receptive field of the simple cell is defined by a σ parameter. The ratio σ/λ determines the special frequency bandwidth, therefore it defines the number of parallel excitatory and inhibitory regions of the receptive field. In this research, we fix the ratio $\lambda = \sigma/0.56$ to have half-response bandwidth of one octave [27]. We also fix the parameter $\varphi = \pi$ to generate a symmetric $S_{\sigma, \lambda, \theta_k, \varphi}(x, y)$. In order to enhance vessel intensity, we employed the computational model of the simple cell (4) as a filter kernel to convolve with the preprocessed image. A block diagram of the proposed approach for improving blood vessels intensities based on the HVS line detection model is shown in Fig. 3.

The vessel enhancement process causes a side effect on non-vessel pixels which are similar to blood vessels. Non-vessel pixels are enhanced as much as blood vessels. These false vessel-like objects are related to some illnesses or various conditions in image acquisition. In order to suppress this side effect, a pruning function is proposed. This function is also motivated by operation of simple cells in the HVS. Simple cells in the HVS react to

an input signal when the output is greater than a particular threshold. Therefore, in each preferred orientation, we consider an adaptive threshold value on the cell responses, $S_{\sigma,\lambda,\theta_k,\varphi}(x,y)$. The pruning function is referred to as an activation function. We define the adaptive activation function $\mu_\alpha(x,y)$ as follows:

$$\mu_\alpha(x,y) = \begin{cases} 1, & \text{if } S_{\sigma,\lambda,\theta_k,\pi}(x,y) \geq \alpha \times \max(S_{\sigma,\lambda,\theta_k,\pi}) \\ 0, & \text{if } S_{\sigma,\lambda,\theta_k,\pi}(x,y) < \alpha \times \max(S_{\sigma,\lambda,\theta_k,\pi}) \end{cases} \quad (6)$$

Constant parameter of α ($0 < \alpha \leq 1$) controls the activation function. As a rule, there is a trade-off between salience of the thin vessels and signal to noise ratio. Although thin vessels are enhanced by setting high value for α , the false vessel-like objects are also highlighted.

The activation function $\mu_\alpha(x,y)$ must be applied to the cell response $S_{\sigma,\lambda,\theta_k,\pi}(x,y)$. As a result, the enhancement regions of $s(x,y)$ is computed by $E_{\sigma,\theta_k,\alpha}(x,y)$ as follows:

$$E_{\sigma,\theta_k,\alpha}(x,y) = \mu_\alpha(x,y) \times S_{\sigma,\lambda,\theta_k,\pi}(x,y) \quad \text{for } \lambda = \sigma/0.56 \quad (7)$$

Experimental results demonstrated a significant role of the proposed adaptive activation function on Gabor filter responses. It improved the accuracy of the proposed method.

Finally, vessel enhancement is completed by combining the cell responses at various directions into a single output. After the kernel is rotated in N_θ directions, the maximum response is selected for each pixel. The output image is considered as a salient image.

$$SI_{\sigma,\alpha}(x,y) = \max(E_{\sigma,\theta_k,\alpha}(x,y)) \quad \text{for } k = 1, 2, \dots, N_\theta \quad (8)$$

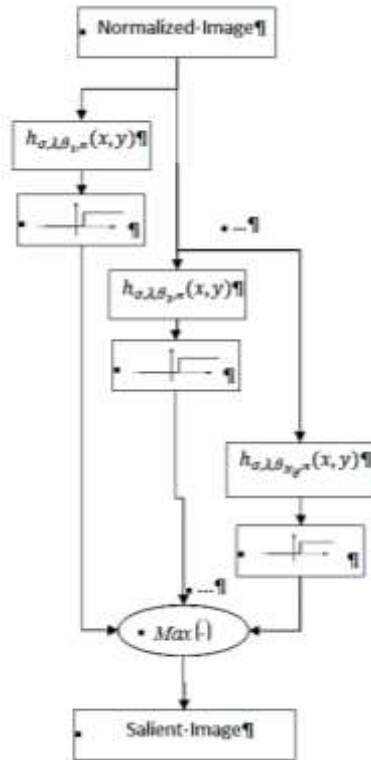


Fig. 3. A procedure of the proposed method for the blood vessels enhancement

3.3 Blood Vessels Segmentation

The output of our method must be a binary image with high value '1' for vessel pixels and low value '0' for non-vessel pixels. Due to the vessel enhancement step, the intensity of vessel pixels is considerably higher than non-vessel pixels, and they would be segmented by a simple threshold. Therefore, we generate the segmentation map by using a simple threshold. In other words, the adaptive threshold acts as a classifier and classifies each pixel as vessel or non-vessel to obtain the vessel binary image. The classifier is modified by intensity average as follows:

$$BM_{\sigma,\alpha,\beta}(x,y) = \begin{cases} 1, & \text{if } SI_{\sigma,\alpha}(x,y) \geq \beta \times \text{mean}(SI_{\sigma,\alpha}) \\ 0, & \text{if } SI_{\sigma,\alpha}(x,y) < \beta \times \text{mean}(SI_{\sigma,\alpha}) \end{cases} \quad (9)$$

Where $BM_{\sigma,\alpha,\beta}(x,y)$ is the binary map of the blood vessels and β is a constant parameter which controls a ratio of vessel pixels and non-vessel pixels. We fixed β to 3/2 for the images.

3.4 Post-Processing

The last phase in our method is a local morphological process on the binary map to overcome the problems arising from over segmentation. Over-segmentation occurs because of lesions or noise in the original image. We can improve the output binary map by removing the lesion and noise areas. This can be done by a local morphological process. Generally, over-segmentation areas are smaller than the thinnest vessel. Hence, all connected objects which are smaller than the thinnest vessel should be removed. Practical experience demonstrated that the thinnest vessel has about 200 pixels in DRIVE [30] and STARE [31] databases. Consequently, in the binary maps, the objects whose areas are less than 200 pixels should be removed.

4. Experimental Results and Evaluation Metrics

4.1 Databases

We utilized the images included in the well-known DRIVE and STARE databases to assess the performance of the proposed method. The DRIVE database comprises 40 eye-fundus colour images. The image set is divided into a test and training sets and each one contains 20 retinal images. The test set is employed for measurement of performance of the vessel segmentation algorithms. The DRIVE database also provides two manual segmentations on each image of the test set which made by two different human observers. The manually segmented images by the 1st human observer are used as a gold-standard image (ground truth). In the STARE database, there is just one image set. It contains 20 images; ten of these contain pathology. It includes two manual segmentations by Hoover and Kouznetsova. The performance is computed with the segmentations of the 1st observer as a gold-standard image.

4.2 Metrics

Our algorithm is evaluated in terms of sensitivity (Se), specificity (Sp), positive predictive value (Ppv), negative predictive value (Npv), and accuracy (Acc). Se and Sp metrics are the ratio of well-classified vessel and non-vessel pixels, respectively. Ppv and Npv are the ratio of pixels classified as vessel pixels and the ratio of pixels classified as background pixels which are both correctly classified. Finally, Acc is a global measurement, and provides the ratio of total well-classified pixels. These metrics are defined as:

$$Se = \frac{TP}{TP+FN} \quad (10)$$

$$Sp = \frac{TN}{TN+FP} \quad (11)$$

$$Ppv = \frac{TP}{TP+FP} \quad (12)$$

$$Npv = \frac{TN}{TN+FN} \quad (13)$$

$$Acc = \frac{TP+TN}{TP+FN+TN+FP} \quad (14)$$

Where TP is the number of pixels correctly classified as vessel pixels; TN is the number of pixels correctly classified as non-vessel pixels; FN is the pixels belonging to a vessel, but is recognized as background pixels. FP is the pixels incorrectly classified as vessel pixels.

The proposal method was implemented on a Windows-7 operating system running on an Intel Pentium 2.7 GHz processor with 4 G RAM. In the implementation of the proposal method, N_θ is set at 8, or the angle resolution is 22.5° which is able to be aligned with vessels in different directions. The constant value of α is fixed at 0.16 for DRIVE database and 0.14 for STARE database. The width of the receptive field of simple cells (σ parameter) is set at 0.6 and the ellipticity of the receptive field (γ parameter) is set at 0.4. They found by iteration to match the filter properly with vessels. As all parameters in Gabor filter are constant, there is no need for ROC curve. The required mask images for the preprocessing step are available in the DRIVE database. For the STARE database, we have generated the mask images as the STARE database did not provide them.

4.3 Results

The performance results of the DRIVE and STARE databases are shown in Table 1 and Table 2. The last rows of the tables show average Se , Sp , Ppv , Npv , and Acc values in each database. The maximum and minimum values are in bold. It can be perceived that the average sensitivity value on DRIVE images is higher than STARE images from the tables. The minimum values of Se are also 0.6650 and 0.4858 on the DRIVE and STARE databases respectively, i.e. our method is more appropriate for the DRIVE than the STARE regarding the ratio of well-classified vessel pixels.

In terms of accuracy, the average values are 0.9403 and 0.9445 for the DRIVE and STARE databases, respectively.

The accuracy of the proposed method on both databases is almost the same. Although the minimum accuracy of the STARE database (0.9150) is a weakness of our method, the average accuracy values are comparable with other results in the literature for both databases.

Two examples of the proposed segmentations from both databases along with gold standard segmentations are given in Fig. 4 and 5. In terms of quality, the proposed method is comparable with the related gold standards.

Table 1. Performance results on DRIVE database images

image	Se	Sp	Ppv	Npv	Acc
1	0.8090	0.9533	0.6889	0.9750	0.9369
2	0.7734	0.9718	0.7967	0.9677	0.9470
3	0.7319	0.9671	0.7579	0.9624	0.9380
4	0.7408	0.9672	0.7456	0.9664	0.9412
5	0.6993	0.9768	0.8015	0.9604	0.9440
6	0.6723	0.9757	0.7942	0.9552	0.9385
7	0.6929	0.9697	0.7471	0.9607	0.9380
8	0.7512	0.9444	0.6223	0.9689	0.9234
9	0.6486	0.9802	0.7919	0.9600	0.9456
10	0.7108	0.9706	0.7419	0.9658	0.9430
11	0.6921	0.9674	0.7357	0.9599	0.9355
12	0.7412	0.9661	0.7263	0.9685	0.9417
13	0.6650	0.9734	0.7785	0.9538	0.9354
14	0.7879	0.9566	0.6741	0.9754	0.9394
15	0.8141	0.9470	0.6077	0.9806	0.9349
16	0.7250	0.9730	0.7719	0.9657	0.9453
17	0.7137	0.9678	0.7220	0.9665	0.9412
18	0.7579	0.9688	0.7270	0.9733	0.9480
19	0.8185	0.9559	0.6856	0.9782	0.9414
20	0.8219	0.9610	0.6823	0.9814	0.9481
Average	0.7384	0.9657	0.7300	0.9673	0.9403

Table 2. Performance results on STARE database images

image	Se	Sp	Ppv	Npv	Acc
1	0.7309	0.9391	0.5587	0.9707	0.9192
2	0.6792	0.9448	0.5216	0.9708	0.9232
3	0.7574	0.9542	0.5649	0.9804	0.9399
4	0.5176	0.9795	0.7039	0.9556	0.9396
5	0.681	0.9704	0.7314	0.9626	0.9398
6	0.8404	0.9493	0.6030	0.9848	0.9401
7	0.8485	0.9641	0.7185	0.9833	0.9528
8	0.8620	0.9625	0.6971	0.9858	0.9533
9	0.8059	0.9685	0.7335	0.9789	0.9528
10	0.7907	0.9286	0.5479	0.9759	0.9150
11	0.7827	0.9747	0.7605	0.9776	0.9568
12	0.8312	0.9779	0.7976	0.9823	0.9640
13	0.7841	0.9660	0.7340	0.9739	0.9465
14	0.7945	0.9627	0.7201	0.9749	0.9446
15	0.7280	0.9720	0.7522	0.9683	0.9465
16	0.7714	0.9426	0.6413	0.9687	0.9224
17	0.7935	0.9619	0.7219	0.9739	0.9433
18	0.4858	0.9964	0.8985	0.9674	0.9652
19	0.5595	0.9907	0.7712	0.9758	0.9679
20	0.6319	0.9846	0.7756	0.9696	0.9573
Average	0.7338	0.9645	0.6977	0.9741	0.9445

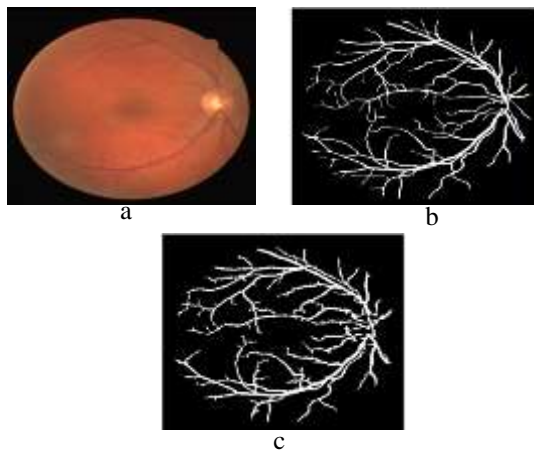


Fig. 4. Sample result of the proposed algorithm (a) the image number 20 from DRIVE database (b) the gold standard (c) result of the binary image of the proposed method

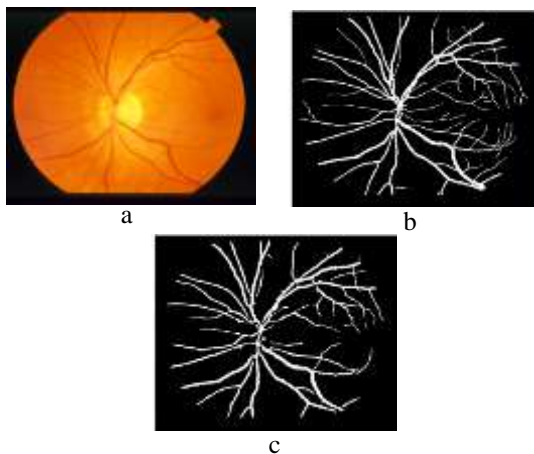


Fig. 5. Sample result of the proposed algorithm (a) the image number 12 from STARE database (b) the gold standard (c) result of the binary image of the proposed method.

4.4 Evaluation

In order to evaluate the performance of the proposed method, we compare our simulation results with the state-of-the-art results and hand-labeled ground truth segmentations in Table 3. The performance of the proposed method is evaluated based on two criteria: the value of accuracy (Acc) to measure the ability of hard-classification, and the ‘sensitivity’ (Se) to measure the ratio of well-classified vessel pixels. The results of Zana [11] and Jiang [32] were taken from the DRIVE database website [30]. A comparative analysis shows that the proposed method achieved better performance metrics than most of the unsupervised methods.

In general, supervised methods outperform unsupervised methods. However, even though our method is unsupervised, sensitivity values of the proposed method for STARE and DRIVE databases are almost higher than the supervised methods. Although Fraz et al [6], Al-Diri et al. [16] and You [33] seem to present better results than our approach, they suffer from high computational complexity and execution time when finding the vessels.

The performance of our method in terms of accuracy is almost superior when compared to unsupervised approaches. Clearly, sensitivity values of unsupervised methods are inferior or not accessible; however, a few methods provide higher accuracy than the proposed approach. In Lam et al. method [15], which presented the highest accuracy among all other methods, the sensitivities were not reported and more importantly it takes 13 minutes to generate the binary image. Table 4 shows the elapsed time comparison among our method with some state-of-the-art algorithms. A comparative analysis shows that Amin et al. [10] method has presented the fastest algorithm among the state-of-the-art algorithms with 0.9081 and 0.9191 accuracy values on STARE and DRIVE databases, respectively. The proposed method is implemented on 2.7GHz machine, which is $((2.7-2.66)/2.66) \times 100 = 1.5\%$ increase in machine speed compared to the machine (2.66 GHz) that Amin et al. used. As a result, their method which took 10 seconds on 2.66 GHz machine, will take $10/(1 + 0.015) \approx 9.85$ seconds on a 2.7 GHz machine. Therefore, the proposed method improves the execution time by about 50% and also presents more accuracy. Consequently, in terms of algorithm speed, the proposed method outperforms all other blood vessel segmentation methods.

5. Conclusion

Automatic segmentation of the retinal blood vessel is the first step in developing a computer-assisted diagnostic system. Extracting blood vessel in the fundus image is a challenging problem. We presented an effective and fast retinal vessel segmentation technique based on the simple cell operation of a primary visual cortex by using a Gabor filter. The proposed method is a fast and simple unsupervised method which does not require any training. The performance of this method was shown by sensitivity, specificity, positive predictive value, negative predictive value and accuracy measurements on DRIVE and STARE databases. Our method consists of four steps: preprocessing, blood vessel enhancement, adaptive thresholding, and post-processing.

Preprocessing is a mandatory step for almost all medical images because of the signal noise, drift in image intensity, and lack of the image contrast. First, we utilized a median filter for preprocessing to generate a uniform image from fundus images. Although, we are not pioneers in using a median filter, we utilized the median filter for fundus images to generate a uniform image. Then, blood vessel enhancement was accomplished based on the simple cell operation in the primary visual cortex. We utilized a directional Gabor filter at eight directions to increase blood vessel intensity. All parameters for the Gabor filter were fixed and did not require to be changed for each image. Next, adaptive thresholding was used to generate a binary image. Adaptive thresholding was utilized as a simple classifier to classify each pixel as a vessel pixel or non-vessel pixel. Finally, a local morphological process was used on the binary image to overcome the problems arising from lesions or noise.

There are solely two variable parameters, α and β , in our method to compromise the accuracy and the sensitivity. Although α and β were fixed in our implementation, they can be changed according to the requirements of its application.

The proposed method outperforms almost all other unsupervised state-of-the-art methods in terms of accuracy, sensitivity and speed. The proposed method is able to acquire the blood vessels in retinal images at about five seconds with an average accuracy of 0.9445 and 0.9403 for the STARE and DRIVE databases, respectively. However, due to a trade-off between the detection of narrow vessels and noise in our approach, several spots are falsely segmented as vessels.

We aim to improve the proposed method by applying a multi-scale Gabor filter instead of a single-scale filter in the future work. The proposed method may be also modified by applying some standard classifiers like the K-Nearest Neighbors algorithm instead of adaptive thresholding to increase the accuracy.

Table 3. Performance results compared to other methods on the STARE and DRIVE databases

Segmentation Methods	Year	STARE database		DRIVE database		Method type
		se	Acc	Se	Acc	
2 nd human observer		0.8951	0.9348	0.7796	0.9470	Hand labeled
Niemeijer [19]	2004	N.A	N.A	N.A	0.9416	Supervised Methods
Soares [21]	2006	0.7207	0.9480	0.7332	0.9466	
Staal [20]	2004	N.A	0.9516	N.A	0.9441	
Marin [22]	2011	0.6944	0.9526	0.7067	0.9452	
Fraz [6]	2012	0.7548	0.9534	0.7406	0.9480	
Hoover [7]	2000	0.6747	0.9264	N.A	N.A	Unsupervised Methods
Zana [11]	2001	N.A	N.A	0.6971	0.9377	
Jiang [32]	2003	N.A	0.9009	N.A	0.9212	
Mendonca [13]	2006	0.6996	0.9440	0.7344	0.9452	

References

- [1] B. Bowling, Kanski's Clinical Ophthalmology: A Systematic Approach, Eighth ed. Sydney, Australia: Elsevier Health Science, 2015.
- [2] M. Esmaeili, H. Rabbani, A. Dehnavi, and A. Dehghani, "Automatic detection of exudates and optic disk in retinal images using curvelet transform," IET image processing, vol. 6, pp. 1005-1013, 2012.
- [3] S. W. Franklin and S. E. Rajan, "Diagnosis of diabetic retinopathy by employing image processing technique to detect exudates in retinal images," IET Image Processing, vol. 8, pp. 1-9, 2014.
- [4] M. J. Fowler, "Microvascular and macrovascular complications of diabetes," Clinical Diabetes, vol. 26, pp. 77-82, 2008.
- [5] J. Anitha, C. K. S. Vijila, and D. J. Hemanth, "An Overview of Computational Intelligence Techniques for Retinal Disease Identification Applications," International Journal of Reviews in Computing, vol. 5, pp. 29-46, 2009.
- [6] M. M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A. R. Rudnicka, C. G. Owen, et al., "An ensemble classification-based approach applied to retinal blood vessel segmentation," Biomedical Engineering, IEEE Transactions on, vol. 59, pp. 2538-2548, 2012.
- [7] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," Medical Imaging, IEEE Transactions on, vol. 19, pp. 203-210, 2000.
- [8] L. Gang, O. Chutatape, and S. M. Krishnan, "Detection and measurement of retinal vessels in fundus images using amplitude modified second-order Gaussian filter," Biomedical Engineering, IEEE Transactions on, vol. 49, pp. 168-172, 2002.
- [9] M. G. Cinsdikici and D. Aydın, "Detection of blood vessels in ophthalmoscope images using MF/ant (matched filter/ant colony) algorithm," Computer methods and programs in biomedicine, vol. 96, pp. 85-95, 2009.

Segmentation Methods	Year	STARE database		DRIVE database		Method type
		se	Acc	Se	Acc	
Lam [14]	2008	N.A	0.9474	N.A	N.A	
Al-Diri [16]	2009	0.7521	N.A	0.7282	N.A	
Cinsdikici [9]	2009	N.A	N.A	N.A	0.9293	
Lam [15]	2010	N.A	0.9567	N.A	0.9472	
Fraz [12]	2011	0.7311	0.9442	0.7152	0.9430	
You [33]	2011	0.7260	0.9497	0.7410	0.9434	
Amin [10]	2011	0.7261	0.9081	0.6608	0.9191	
Proposed Method	2014	0.7338	0.9445	0.7384	0.9403	

N.A: Not Available

Table 4. Comparison of the execution times based on STARE and DRIVE databases

Vessel detection method	Computer configuration	STARE	DRIVE	Execution time
		Acc	Acc	
Manual segmentation	Second observer	0.9348	0.9470	120 min
Lam [14]	MATLAB, Intel Pentium 2.66 GHz, 512 MB RAM	0.9474	N.A	25 min
Lam [15]	MATLAB, Duo CPU 1.83 GHz, 2 GB RAM	0.9567	0.9472	13 min
Soares [21]	MATLAB, AMD Athlon 2.2 GHz, 1 G RAM	0.9480	0.9466	3 min 10 s
Staal [20]	N.A, Intel Pentium 1 GHz, 1 G RAM	0.9516	0.9441	15 min
Al-Diri [16]	MATLAB, 1.2 GHz Pentium system	N.A	N.A	11 min
Amin [10]	MATLAB, Intel Pentium 2.66 GHz, 512 MB RAM	0.9081	0.9191	10 s
Cinsdikici [9]	N.A		0.9293	35 s
The Proposed Method	Intel Pentium 2.7 GHz, 4 G RAM	0.9445	0.9403	5 s

N.A: Not Available

- [10] M. A. Amin and H. Yan, "High speed detection of retinal blood vessels in fundus image using phase congruency," *Soft Computing*, vol. 15, pp. 1217-1230, 2011.
- [11] F. Zana and J.-C. Klein, "Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation," *Image Processing, IEEE Transactions on*, vol. 10, pp. 1010-1019, 2001.
- [12] M. Fraz, S. Barman, P. Remagnino, A. Hoppe, A. Basit, B. Uyyanonvara, et al., "An approach to localize the retinal blood vessels using bit planes and centerline detection," *Computer methods and programs in biomedicine*, 2011.
- [13] A. M. Mendonca and A. Campilho, "Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction," *Medical Imaging, IEEE Transactions on*, vol. 25, pp. 1200-1213, 2006.
- [14] B. S. Lam and H. Yan, "A novel vessel segmentation algorithm for pathological retina images based on the divergence of vector fields," *Medical Imaging, IEEE Transactions on*, vol. 27, pp. 237-246, 2008.
- [15] B. S. Lam, Y. Gao, and A.-C. Liew, "General retinal vessel segmentation using regularization-based multiconcavity modeling," *Medical Imaging, IEEE Transactions on*, vol. 29, pp. 1369-1381, 2010.
- [16] B. Al-Diri, A. Hunter, and D. Steel, "An active contour model for segmenting and measuring retinal vessels," *Medical Imaging, IEEE Transactions on*, vol. 28, pp. 1488-1497, 2009.
- [17] G. Gardner, D. Keating, T. Williamson, and A. Elliott, "Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool," *British journal of Ophthalmology*, vol. 80, pp. 940-944, 1996.
- [18] C. Sinthanayothin, J. F. Boyce, H. L. Cook, and T. H. Williamson, "Automated localisation of the optic disc, fovea, and retinal blood vessels from digital colour fundus images," *British Journal of Ophthalmology*, vol. 83, pp. 902-910, 1999.
- [19] M. Niemeijer, J. Staal, B. van Ginneken, M. Loog, and M. D. Abramoff, "Comparative study of retinal vessel segmentation methods on a new publicly available database," in *Medical Imaging 2004*, pp. 648-656.
- [20] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *Medical Imaging, IEEE Transactions on*, vol. 23, pp. 501-509, 2004.
- [21] J. V. Soares, J. J. Leandro, R. M. Cesar, H. F. Jelinek, and M. J. Cree, "Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification," *Medical Imaging, IEEE Transactions on*, vol. 25, pp. 1214-1222, 2006.
- [22] D. Marín, A. Aquino, M. E. Gegúndez-Arias, and J. M. Bravo, "A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features," *Medical Imaging, IEEE Transactions on*, vol. 30, pp. 146-158, 2011.
- [23] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex," *Journal of Neurophysiology*, vol. 58, pp. 1233-1258, 1987.
- [24] D. G. Albrecht and W. S. Geisler, "Motion selectivity and the contrast-response function of simple cells in the visual cortex," *Visual neuroscience*, vol. 7, pp. 531-546, 1991.
- [25] D. C. Somers, S. B. Nelson, and M. Sur, "An emergent model of orientation selectivity in cat visual cortical simple cells," *The Journal of neuroscience*, vol. 15, pp. 5448-5465, 1995.
- [26] D. Ferster, S. Chung, and H. Wheat, "Orientation selectivity of thalamic input to simple cells of cat visual cortex," *Nature*, vol. 380, pp. 249-252, 1996.
- [27] C. Grigorescu, N. Petkov, and M. A. Westenberg, "Contour detection based on nonclassical receptive field inhibition," *Image Processing, IEEE Transactions on*, vol. 12, pp. 729-739, 2003.
- [28] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, p. 106, 1962.
- [29] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Optical Society of America, Journal, A: Optics and Image Science*, vol. 2, pp. 1160-1169, 1985.
- [30] Research Section, Digital Retinal Image for Vessel Extraction (DRIVE) Database. Utrecht, The Netherlands, Univ. Med. Center Utrecht, Image Sci. Inst. [Online]. Available: <http://www.isi.uu.nl/Research/Databases/DRIVE/> [Feb. 1, 2016]
- [31] STARE database, STARE Project Website. Clemson, SC, Clemson Univ. [Online]. Available: <http://www.ces.clemson.edu/~ahoover/stare/> [Feb. 1, 2016]
- [32] X. Jiang and D. Mojon, "Adaptive local thresholding by verification-based multithreshold probing with application to vessel detection in retinal images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, pp. 131-137, 2003.
- [33] X. You, Q. Peng, Y. Yuan, Y.-m. Cheung, and J. Lei, "Segmentation of retinal blood vessels using the radial projection and semi-supervised approach," *Pattern Recognition*, vol. 44, pp. 2314-2324, 2011.

Mohsen Zardadi received his M.Sc degree in 2006 from Ferdowsi University of Mashhad, Mashhad, Iran in Electrical Engineering. Since 2012 he has been pursuing his Ph.D degree in the Department of Electrical and Computer Engineering of Birjand University, Birjand, Iran. His research interests include computer vision and image processing.

Nasser Mehrshad received the B.Sc degree from Ferdowsi University of Mashhad, Mashhad, Iran, in 1995 and the M.Sc and Ph.D degrees from Tarbiat Modares University, Tehran, Iran, in 1998 and 2005, respectively, both in Biomedical Engineering. He is currently an Associate Professor in the Department of Electrical and Computer Engineering, the University of Birjand, Birjand, Iran. His research interests include Computer Vision, Digital Signal and Image Processing, Biometrics and Biomedical Data Mining.

Seyyed Mohammad Razavi received the B.Sc degree in Electrical Engineering from Amirkabir University of Technology, Tehran, Iran, in 1994 and the M.Sc degree in Electrical Engineering from the Tarbiat Modares University, Tehran, Iran, in 1996, and the Ph.D degree in Electrical Engineering from the Tarbiat Modares University, Tehran, Iran, in 2006. Now, he is an Associate Professor in the Department of Electrical and Computer Engineering, the University of Birjand, Birjand, Iran. His research interests include Computer Vision, Pattern Recognition and Artificial Intelligence Algorithms.