In the Name of God

Journal of
Information Systems & Telecommunication

Vol. 3, No. 1, January-March 2015, Serial Number 9

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

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

Print ISSN: 2322-1437
Online ISSN: 2345-2773
Publication License: 91/13216

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

Indexed in:

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

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

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

Acknowledgement

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

- Abbasi Mehdi, Bu-Ali Sina University, Hamedan, Iran
- Abed Hodtani Ghosheh, Ferdowsi University of Mashhad, Mashhad, Iran
- Abdolvand Neda, Alzahra University, Tehran, Iran
- Ahadi Akhlaghi Iman, Sadjad University Of Technology, Mashhad, Iran
- Akbarizadeh Gholamreza, Shahid Chamran University of Ahvaz, Ahvaz, Iran
- Amiri Hadi, University of Maryland, Maryland, USA
- Anvaripour Mohammad, Academic Center for Education Culture and Research (ACECR), Tehran, Iran
- Azimzadeh Fatemeh, Academic Center for Education Culture and Research (ACECR), Tehran, Iran
- Daneshvar Sabalan, University of Tabriz, Tabriz, Iran
- Farsi Hasan, University of Birjand, Birjand, Iran
- Ghaffari Ali, Islamic Azad University, Tabriz Branch, Tabriz, Iran
- Ghanbari Mohammad, University of Essex, Colchester, UK
- Ghazi Maghrebi Saeid, Academic Center for Education Culture and Research (ACECR), Tehran, Iran
- Gholamian Mohammad Reza, Iran University of Science and Technology, Tehran, Iran
- Ghorashi Seyed Ali, Shahid Beheshti University, Tehran, Iran
- Habibi Bastami Ali, Khaje Nasir-edin Toosi University of Technology, Tehran, Iran
- Haddadi Farzan, Iran University of Science and Technology, Tehran, Iran
- Haji Mohammadi Zeinab, Abrar Institute of Higher Education, Tehran, Iran
- Hamidi Hojjatollah, Khaje Nasir-edin Toosi University of Technology, Tehran, Iran
- Hashemi Rafsanjani Hadi, Academic Center for Education Culture and Research (ACECR), Tehran, Iran
- Hosseini Monireh, Khaje Nasir-edin Toosi University of Technology, Tehran, Iran
- Hossein Khalaj Babak, Sharif University of Technology, Tehran, Iran
- Jamali Shahram, University of Mohaghegh Ardabili, Ardabil, Iran
- Jamshidi Azizollah, Shiraz University, Shiraz, Iran
- Kamandar Mehdi, Kerman Graduate University of Technology
- Kashef Seyed Sadra, Tarbiat Modares University, Tehran, Iran
- Kazerooni Morteza, Malek Ashtar University of Technology, Tehran, Iran
- Keshavarz Hengameh, University of Sistan & Baluchestan, Zahedan, Iran
- Khademi Morteza, Ferdowsi University of Mashhad, Mashhad, Iran
- Lotfi Abbasali, Academic Center for Education Culture and Research (ACECR), Tehran, Iran
- Mahdieh Omid, University of Zanjan, Zanjan, Iran
- Mashhadi Saeid, Sharif University of Technology, Tehran, Iran
- Modiri Nasser, Islamic Azad University, Zanjan Branch, Zanjan, Iran

- Moghaddasi Mohammad Naser, Islamic Azad University, Science and Research Branch, Tehran, Iran
- Mohammadi Shahriar, Khaje Nasir-edin Toosi University of Technology, Tehran, Iran
- Moradi Gholamreza, Amirkabir University of Technology, Tehran, Iran
- Najimi Maryam, Babol Noshirvani University of Technology, Babol, Iran
- Nayebi Mohammad Mehdi, Sharif University of Technology, Tehran, Iran
- Noroozi Yaser, Amirkabir University of Technology, Tehran, Iran
- Norouzi Ali, Istanbul University, Istanbul, Turkey
- Nourbakhsh Azamossadat, Islamic Azad university, Lahijan Branch, Lahijan, Iran
- Rafeh Reza, Arak University, Arak, Iran
- Rasi Habib, Shiraz University of Technology, Shiraz, Iran
- Rezvanian Alireza, Amirkabir University of Technology, Tehran, Iran
- Robat Mili Mohammad, Payame Noor University, Tehran, Iran
- Safarinejadian Behrooz, Shiraz University, Shiraz, Iran
- Sajedi Hedieh, University of Tehran, Tehran, Iran
- Shaker Gholam, Islamic Azad University, Science and Research Branch, Tehran, Iran
- Shokooh Saremi Mehrdad, Ferdowsi University of Mashhad, Mashhad, Iran
- Soleiman Meiguni Javad, Semnan University, Semnan, Iran
- Tavassoli Sude, Technique University of Kaiserslautern, Kaiserslautern, Germany
- Torabi Jahromi Amin, Nanyang Technological University, Singapore
- Zarrabi Vahid, Academic Center for Education Culture and Research (ACECR), Tehran, Iran

Table of Contents

A New Recursive Algorithm for Universal Coding of Integers

Mehdi Nangir*
Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran
mahdinangir@gmail.com

Hamid Behroozi
Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran
behroozi@sharif.edu

Mohammad Reza Aref
Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran
aref@sharif.edu

Received: 14/Jul/2013 Revised: 08/Sep/2014 Accepted: 13/Nov/2014

Abstract
In this paper, we aim to encode the set of all positive integers so that the codewords not only be uniquely decodable but also be an instantaneous set of binary sequences. Elias introduces three recursive algorithms for universal coding of positive integers where each codeword contains binary representation of the integer plus an attachment portion that gives some information about the first part [1]. On the other hand, Fibonacci coding which is based on Fibonacci numbers is also introduced by Apostolico and Fraenkel for coding of integers [2]. In this paper, we propose a new lossless recursive algorithm for universal coding of positive integers based on both recursive algorithms and Fibonacci coding scheme without using any knowledge about the source statistics [3].The coding schemes which don't use the source statistics is called universal coding, in these universal coding schemes we should use a universal decoding scheme in the receiver side of communication system. All of these encoding and decoding schemes assign binary streams to positive integers and conversely, without any need of use to probability masses over positive integers. We show that if we use Fibonacci coding in the first part of each codeword we can achieve shorter expected codeword length than Elias Omega code. In addition, our proposed algorithm has low complexity of encoding and decoding procedures.

Keywords: Universal Source Coding (data compression); Fibonacci Coding; Elias Coding Schemes; Integer Representation; Omega Coding; Redundancy.

1. Introduction

Researchers in the field of source coding and data compression have focused on offering efficient and fast running algorithms that have simple software and hardware implementation with expected codeword length close to entropy [4]. In this work, we aim to encode all the positive integers so that the set of codewords not only be uniquely decodable but also be an instantaneous set of binary sequences [5]. There are many situations in which the alphabet is a set of positive integers. For example, we might have a list of items and we wish to encode the position of an element in the list. Also, we may want to encode an image file by encoding the intensities.

We can classify algorithms in this field into two categories: In the first one, that we refer to as recursive algorithms, codewords have a prefix portion or a suffix portion of binary string that converts standard binary representation of positive integers to a uniquely decipherable presentation or even an instantaneous representation with minimum possible expected codeword length and low complexity encoding and decoding algorithms. In the second category, there are algorithms based on applying a complex mathematical method to compress equivalent binary representation of positive integers as much as possible. It is obvious that, with the same goal of getting shorter expected codeword length, algorithms in the second category have more complexity for encoding and decoding procedures compared with the algorithms in the first category.

Peter Elias offers three universal coding schemes to encode positive integers [1]. Universal coding means that there is no need to use the probability distribution or statistical properties of the source that we want to encode or the data that we want to compress. Elias named these three algorithms Gamma, Delta and Omega representations. Usually compared with Delta, Omega, ω, has shorter expected codeword length while Delta has shorter expected codeword length compared with Gamma. These three algorithms are in the first category; they attach a binary representation to standard binary presentation of the positive integer that provides the capability of being

* Corresponding Author

uniquely decipherable or instantaneous property to the codeword set [5]. Our proposed algorithm is also in the category of recursive schemes.

In Omega coding scheme for positive integers [1], first the binary representation of the integer is obtained. Then, we attach a prefix portion that shows the number of bits that was written in the previous step minus one in a binary format. We continue this procedure until the prefix portion be two-bit size part. At the end of the procedure, bit "0" is attached as a delimiter to indicate the end of the codeword. For example, the Omega codeword for 2012 is

$$\omega(2012) = 111010111110111000$$

Mathematically, Elias Omega codeword for any positive integer n_0 can be writtenas the following recursive structure [1], [6]

$$C_E(n_0) = [n_k]_2[n_{k-1}]_2 \dots [n_1]_2[n_0]_2 0 \qquad (1)$$

Where $[n]_2$ is the binary representation of n. Each n_k in (1) is obtained recursively by $n_k = \lfloor \log_2 n_{k-1} \rfloor$ where the recursive algorithm stops when the length of $[n_k]_2$ is two. The decoding procedure is simply the inverse of the encoding procedure: we read the first two bits of every codeword $C_E(n_0)$ to obtainn_k, and then we know that $n_k + 1$ represents the bit length of$[n_{k-1}]_2$. Thus the length of $[n_{k-1}]_2$ is recursively obtained from n_k. We continue this procedure until the first bit of the next portion is 0, and then the last portion is the number that is encoded. In fact, since the most significant bit (MSB) of every $[n_k]_2$ is "1", delimiter "0" can stop the recursion andn_0 can be easily found.

A simple coding scheme in the second category of algorithms is Fibonacci coding to encode positive integers [2], [7]. In this scheme, we write an integer number according to summation of Fibonacci numbers. The first 10 Fibonacci numbers, F_n for $n = 1,2,\cdots,10$, are shown in Table 1. To encode an integer$N$, first the largest Fibonacci number equal to or less than N is determined. If the number subtracted was the i'*th* Fibonacci number, F_i, a "1" will be placed as the i'*th* bit in the codeword. By subtracting this Fibonacci number from$N$, we repeat the previous steps for the reminder untila remainder of 0 is reached. Eventually, we add a "1" tothe rightmost digit in the codeword, which indicates the codeword is ended. The Fibonacci representation of an integer has an interesting property that it does not contain any adjacent 1's [8], so that receiving "11" string means the end of the codeword [2],[7]. For example, the Fibonacci codeword of 2012 is

$$F(2012) = 10100001000010011,$$

Since $2012 = F_1 + F_3 + F_8 + F_{13} + F_{16}$, in which F_i shows the i'*th* term in Fibonacci sequence. A disadvantage of the Fibonacci coding is that the complexity of coding and decoding algorithms increases by increasing the numbers to be coded or the length of stream to be decoded but this scheme is very efficient if larger numbers are more frequent than smaller ones [9].

Table 1: Fibonacci Numbers

$$F_1 = 1, F_2 = 2, \ F_n = F_{n-1} + F_{n-2} \ \ \forall n \geq 3$$

F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}
1	2	3	5	8	13	21	34	55	89

In this paper, we propose a new recursive algorithm to encode all of the positive integers. In this proposed coding scheme, every codeword has a prefix portion which provides the uniquely decipherable property to the codeword set. We observe that our proposed algorithm is an instantaneous set of binary codes with low complexity of encoding and decoding procedures (in both software and hardware implementations). Numerical simulations show that generally for most of probability distributions, the proposed coding has shorter expected codeword length that Elias Omega coding scheme in [1].

The paper is organized as follows. The problem definition is presented in Section 2. Our proposed universal coding for positive integers is introduced in Section 3. For comparison, the expected codeword length of our proposed algorithm is compared with that of the Elias ω code in Section 4, where two sufficient conditions on the discrete probability distribution of the input source are provided in order to get shorter expected codeword length than Omega coding. We present the decoding procedure in Section 6. Conclusions are finally drawn in Section 6.

2. Problem Definition

The general problem we encounter here is encoding the set of integer numbers without any use of probability distribution. In other words, the encoder and the decoder do not know the statistics of the input stream. This is called universal source coding or universal data compression [10]. The concept of universal coding of integers gets a lot of attention by researchers who work in source coding and data compression field [3].

Similar to [1], [6], [11]-[13] we have treated the universal coding of the positive integers that have decreasing probability distribution for positive integers, i.e.,

$$P(n) \geq P(n+1) \quad \forall n \in \mathbb{N}$$

where $P(n)$ is a probability distribution on the set of positive integers, $\mathbb{N} = \{1,2,3,\cdots\}$. We assume that the inputsymbols are emitted by a memoryless information source with a possibly infinite alphabet, i.e., data emitted from the source is an independent and identically distributed (i.i.d) stream. Since we assume that the source has infinite alphabet, we can apply a one-to-one mapping between the source alphabet and the set of positive integers.

We aim to provide an encoding algorithm that encodes all the positive integers independently. Thus, the encoder is a memoryless system. Note that there are some universal encoding schemes in the literature which use memory in the encoding procedures [14]-[16].

By considering integers of the source output into blocks, each one containing n integer numbers, wecan apply our algorithm to each block separately, similar to thescheme in [17]. From the decreasing probability distribution for integers, it is clear that the codeword length would be an ascending function of$n$. Our goal is to encode positive integers with low complexity and minimum redundancy. Note that for a probability distribution P, the normalized redundancy is defined in [18] as

$$R(P) \triangleq \frac{\mathbb{E}L(P) - H(P)}{H(P)}$$

where $\mathbb{E}L(P)$ is the expected codeword length over the probability distribution P, and $H(P)$ denotes the entropy of the source. Due to first theorem of Shannon [5], we know that$R(P) \geq 0$, so we get more efficient compression if the normalized redundancy approaches to zero. Although there are some schemes that achieve an expected codeword length close to the Shannon entropy over most of probability distributions [19], [20], the drawback of these algorithms is being lossy. In this paper, we describe a lossless scheme that encodes positive integers universally with expected codeword length close to the entropy of the source due to Elias Omega codeword length [1].

3. Proposed Coding Scheme

Here, we describe our proposed algorithm. As we mentioned before, our algorithm is a recursive scheme which has three steps:

1. First, we write the standard binary representation of the positive integer number that we intend to encode and remove its most significant bit. For instance, for integer 17, we write 0001.
2. Then, we count the number of bits that we obtained in the first step (clearly, for number n we have $\lfloor \log_2 n \rfloor$ bits). For the integer number 17, we have 4 bits.
3. Finally, we attach the Fibonacci code of $\lfloor \log_2 n \rfloor$ to the left side of the binary string we obtained in the first step.

As an example, the integer number 17 has 4 bits in the first step and hence the Fibonacci codeword for integer 4 is $F(4) = 1011$. We attach 1011 to the left side of 0001 and obtain the codeword 10110001.

However, there exists one problem here. In the above algorithm, the positive integer "1" does not have any codeword. In other words, this algorithm produces the null codeword for the number "1", so we have to assign a single bit to determine the number to be encoded is "1" or not. We add a single bit at the frontier of the codewords to determine whether the number that is encoded is "1" or not. At the decoder, if the first bit is 1 we know that "1" is encoded and if the first bit is 0 we conclude that the encoded number is not "1". So that the codeword of "1" is the single bit 1, i.e., we write$C(1) = 1$. As a result, the codeword for the positive integer number 17 is $C(17) = 010110001$ where we added single "0" bit to the beginning of the codeword "10110001".

In Table 2, we provide the codewords of some positive integers based on our proposed scheme.

Table 2: Codewords of some integers based on our proposed scheme

Integer number	Our codeword
1	1
9	00011001
23	010110111
53	00001110101
78	010011001110
1000	0100011111101000
2012	00100111111011100

4. Performance Comparison between our Proposed Scheme and Omega Scheme

For comparison, the codeword lengths of the proposed scheme for some integers versus Elias ω code, which is a verystrong andfamous method in data compression and algorithmic source coding, are provided in Table 3. Based on the codeword lengths in Table 3 and also using simulation results, we observe that the only integers for which our codeword lengths exceed Elias ω codeword lengths are {2,3,8,9,10,11,12,13,14,15} (among all of the positive integers) and for the rest of integers, $\{1, 4, 5, 6, 7, 16, 17, 18, \ldots\}$, our coding scheme performs superior to Elias ω coding scheme, i.e., our codewords have shorter lengths than Elias ω codewords. Thus, our proposed algorithm results in a shorter expected codeword length over the most of discrete probability distributions, compared with the Elias ω coding scheme.

An advantage of our proposed algorithm is that if there is no integer "1" in the input sequence, we can eliminate the first bit from all of the code words, since this bit is determining whether the encoded number is "1" or not.

We apply our proposed algorithm to encode all the integers between "1 to 128". In Figure 1, we presented codeword lengths of two coding schemes: our proposed coding scheme and Elias Omega coding scheme. We observe that except for some small integers, our codeword lengths are shorter than codeword lengths of Elias ω coding scheme. More precisely, the only integers that Omega codewords have shorter lengths are 2,3,8,9,10,11,12,13,14 and 15 among all the positive integers and for the rest of integers ({1,4,5,6,7,16,17,18,...}) our codewords have shorter lengths.

Table 3: Codeword length of our proposed scheme v.s Elias ω coding scheme

Integer number	Elias ω code	Our proposed code
1	1	1
2-3	3	4
4-7	6	6
8-15	7	8
16-31	11	9
32-63	12	11
64-127	13	12
128-255	14	13

Also note that because of constructing the prefix portion of the codeword from number of bits that is written in the first step of the encoding, the codeword lengths differ for integer numbers that are different powers of 2. In other words, for every positive integer$k$, all of the integers from 2^{k-1} to 2^k-1 have equal codeword length, which can be observed in Figure 1.

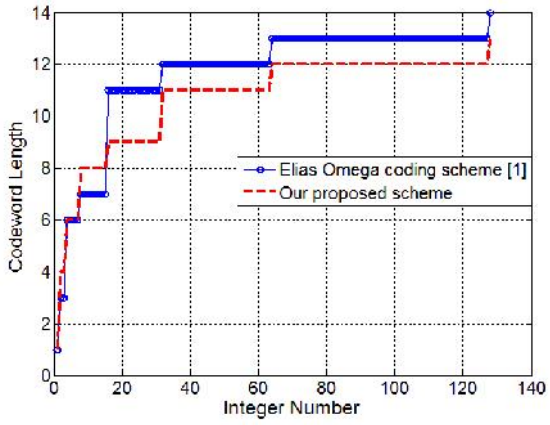


Fig. 1: Performance of our proposed algorithm compared with that of Elias Omega coding scheme.

In the next two Theorems, we provide strong sufficient conditions for discrete probability distributions over integers for which our proposed algorithm performs superior to the other algorithms such as Elias Omega coding scheme.

Theorem 1. A strong sufficient condition on the discrete probability distribution of a source to achieve shorter expected codeword length, based on our algorithm, than Omega coding scheme is

$$2(p_{16}+p_{17}+\cdots+p_{31}) \geq (p_2+p_3) + (p_8+p_9+\cdots+p_{15}) \qquad (2)$$

where p_i denotes the occurrence probability of number "i" in the output sequence of the discrete source.

Proof: From the condition provided in Theorem 1, if we add the following terms to both sides of (2):

$$p_1, 3(p_2+p_3), 6(p_4+\cdots+p_7), 7(p_8+\cdots+p_{15}), 9(p_{16}+\cdots+p_{31}),$$

we obtain

$$p_1+3(p_2+p_3)+6(p_4+\cdots+p_7)+7(p_8+\cdots+p_{15})+11(p_{16}+\cdots+p_{31}) \geq p_1+4(p_2+p_3)+6(p_4+\cdots+p_7)+8(p_8+\cdots+p_{15})+9(p_{16}+\cdots+p_{31}).$$

Let $L(\omega(n))$ and $L(C(n))$ denote the codeword lengths for the Elias Omega code and our proposed code, respectively. Since based on Table 3, we know that for the positive integers larger than "32", the codeword lengths of our proposed scheme are shorter than those of Elias ω code, we can write the following inequality

$$\begin{aligned} &p_1+3(p_2+p_3)+6(p_4+\cdots+p_7)+7(p_8+\cdots+p_{15})+11(p_{16}+\cdots+p_{31})+\sum_{j=32}^{\infty} p_j L(\omega(j)) \\ &> p_1+4(p_2+p_3)+6(p_4+\cdots+p_7)+8(p_8+\cdots+p_{15})+9(p_{16}+\cdots+p_{31})+\sum_{j=32}^{\infty} p_j L(C(j)) \end{aligned} \qquad (3)$$

If $\mathbb{E}[.]$ denotes the expectation of a random variable over the probability distribution of the source, then (3) results in

$$\mathbb{E}\left[L(\omega(n))\right] > E\left[L(C(n))\right].$$

Thus, we achieve a code whose expected codeword length is shorter than that of Elias coding scheme. This completes the proof.

For example, if we check the condition of Theorem 1, given in (2), for character probability distribution (relative frequencies of letters in the English language [21]) we observe that

$$2(p_{16}+p_{17}+\cdots+p_{31}) = 0.3895,$$

$$(p_2+p_3)+(p_8+p_9+\cdots+p_{15}) = 0.379,$$

which shows validity of the condition in Theorem 1 for a practical scenario.

Theorem 2. The second strong sufficient condition on the discrete probability distribution of a source to achieve shorter expected codeword length, based on our coding scheme, than Omega coding is

$$p_{16} + p_{17} + p_{18} + \cdots \geq (p_2 + p_3) + (p_8 + p_9 + \cdots + p_{15}) \quad (4)$$

where p_i denotes the occurrence probability of number "i" in the output sequence of the discrete source.

Proof: Using (4) and adding the following terms to both sides:

$$p_1, p_2 + p_3, p_4 + \cdots + p_7, p_8 + \cdots + p_{15},$$

we get

$$1 = p_1 + p_2 + p_3 + \cdots \geq p_1 + 2(p_2 + p_3) + p_4 + \cdots + p_7 + 2(p_8 + \cdots + p_{15}) \quad (5)$$

Since we know that our codewords have one bit more than Elias ωcoding scheme only for the following positive integers: {2,3,8,9,10,11,12,13,14,15}, and for the rest of integers, our scheme has shorter codeword length compared with the Omega coding scheme, we can write (5) as

$$\begin{aligned} p_1 + 3(p_2 + p_3) + 6(p_4 + \cdots + p_7) &+ 7(p_8 + \cdots + p_{15}) \\ &+ 11(p_{16} + \cdots + p_{31}) \\ &+ \sum_{j=32}^{\infty} p_j L(\omega(j)) \\ &> p_1 + 4(p_2 + p_3) \\ &+ 6(p_4 + \cdots + p_7) \\ &+ 8(p_8 + \cdots + p_{15}) \\ &+ 9(p_{16} + \cdots + p_{31}) \\ &+ \sum_{j=32}^{\infty} p_j L(C(j)) \end{aligned}$$

which results in

$$\mathbb{E}\left[L(\omega(n))\right] > E\left[L(C(n))\right].$$

Thus, the proof is complete.

5. Decoding Procedure

Decoding scheme of a universal source code also should be universal, which means that the decoder does not know the source statistics [22].

By assuming an ideal communication channel with no error, the decoding algorithm can be described as follows:

From the first bit of the codeword, the decoder knows that the coded number is "1" or not. In other words, if the MSB of the received codeword is 1, the integer number that is encoded is "1", otherwise the integer number that is encoded is not "1". In the latter case, the decoder reads the remaining part of the codeword until it receives the first "11" substring. This last part is the Fibonacci portion of the codeword. By decoding this Fibonacci portion and assuming that it is decoded to the integer number $\log_2 n$, the decoder then reads $\log_2 n$ bits after "11" sequence. This part with adding "1" as its MSB will be the binary presentation of the number that has been encoded.

For example, if we receive the following sequence: 110101100011, by applying the decoding scheme presented here we obtain the following sequence of integers: "1-1-17-1".

6. Conclusions

In this paper, we proposed a new recursive algorithm that achieves an expected codeword length which is shorter than that of other universal coding schemes such as Elias ω coding scheme. Note that Elias ω coding scheme has the shortest expected codeword length among all three coding schemes presented in [1]. In the proposed algorithm, we applied Fibonacci coding scheme in the prefix part of the codeword. More precisely, we first obtain the standard binary representation of the positive integer number and remove its most significant bit. Then, we count the number of bits that we obtained in the first step. We then attach the Fibonacci code of the number of bits obtained in the first step to the left side of the binary string we obtained in the first step. We know that the Fibonacci representation of an integer has the interesting property that it does not contain any adjacent 1's, so that receiving "11" string means the end of prefix part. Moreover, here we provided two sufficient conditions on the discrete probability distribution of the source to get shorter expected codeword length than that of Omega coding. Eventually, we applied our proposed algorithm to some first positive integers and observed its better performance compared with the codeword length for Elias ω coding scheme.

Acknowledgments

The authors would like to thank Dr. Farzan Haddadi for his valuable comments and helpful suggestions. This work has been supported in part by the Iran NSF under Grant No. 92/32575 and by the Iran Telecommunication Research Center (ITRC).

References

[1] P. Elias, "Universal codeword sets and representations of the integers," *IEEE Trans. Inf. Theory*, vol. 21, pp. 194–203, 1975.

[2] A. Apostolico and Aviezri S. Fraenkel, "Robust transmission of unbounded strings using Fibonacci representations," IEEE *Trans. Inf. Theory*, vol. 33, no. 2, pp. 238 – 245, Mar. 1987.

[3] P. Andreasen, Universal Source Coding, M.Sc. thesis, Math. Dept., Univ. of Copenhagen, July 2001.

[4] Z. Chen and M. H. Lee "On Fast Hybrid Source Coding Design," in Proc. of the 2007 International Symposium on Information Technology Convergence (ISITC 2007), pp. 143-147, Nov. 2007.

[5] T. M. Cover and J. A. Thomas, Elements of Information Theory. New York: 2nd Edition, John Wiley & Sons, 2006.

[6] T. Amemiya and H. Yamamoto, "A new class of the universal representation for the positive integers," *IEICE Trans. Fundamentals*, vol. E76-A, no. 3, pp. 447–452, Mar. 1993.

[7] S. Fraenkel and S. T. Klein, "Robust universal complete codes for transmission and compression," *Discrete Applied Mathematics*, vol. 64, 1996.

[8] D. E. Knuth, the Art of Computer Programming. vol. 1, 2nd Ed., Reading, MA, Addison-Wesley, 1973.

[9] S. Mohajer and A. Kakhbod, "Anti-uniform Huffman codes," *IET Communications*, vol. 5, no. 9, pp. 1213-1219, June 2011.

[10] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 783–795, Nov. 1973.

[11] V. I. Levenshtein, "On the redundancy and delay of decodable coding of natural numbers," *Probl. Cybern.*, vol. 20, 1968.

[12] R. Ahlswede, T. S. Han, and K. Kobayashi, "Universal coding of integers and unbounded search trees," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 669–682, Mar. 1997.

[13] G. I. Shamir, "Universal Source Coding for Monotonic and Fast Decaying Monotonic Distributions," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, Nov 2013.

[14] S. Jalali, S. Verdu, and T. Weissman, "A Universal Scheme for Wyner–Ziv Coding of Discrete Sources," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1737-1750, April 2010.

[15] A. Beirami, M. Sardari, and F. Fekri, "Results on the fundamental gain of memory-assisted universal source coding," in Proc. 2012 IEEE International Symposium on Information Theory (ISIT), July 2012, pp. 1092-1096.

[16] A. Beirami and F. Fekri, "Memory-Assisted Universal Source Coding," Data Compression Conference (DCC), April 2012.

[17] D. P. Foster, R. A. Stine, and A. J. Wyner, "Universal codes for finite sequences of integers drawn from a monotone distribution," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1713–1720, Jun. 2002.

[18] T. S. Han and K. Kobayashi, Mathematics of Information and Coding. American Mathematical Society, 2002.

[19] S. Jalali and T. Weissman, "Near optimal lossy source coding and compression-based denoising via Markov chain Monte Carlo," in Proc. 42nd Annual Conference on Information Sciences and Systems (CISS 2008), pp. 441-446, March 2008.

[20] S. Jalali and T. Weissman, "Lossy Source Coding via Markov Chain Monte Carlo," in Proc. IEEE International Zurich Seminar on Communications, pp. 80-83, March 2008.

[21] D. Salomon, Data Compression- The Complete Reference. Fourth Edition, Springer, 2007.

[22] E. Ordentlich, G. Seroussi, S. Verdu, and K. Viswanathan, "Universal Algorithms for Channel Decoding of Uncompressed Sources," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2243-2262, May 2008.

Mehdi Nangir received the B.Sc. degree in Electrical Engineering from Tabriz University and M.Sc. degree in Communication System Engineering from Sharif University of Technology in 2010 and 2012, respectively. He is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering at Khaje Nasir University of Technology, Tehran, Iran. His research interests include information theory, cryptography, data compression algorithms and source and channel coding.

Hamid Behroozi received the B.Sc. degree from University of Tehran, Tehran, Iran, the M.Sc. degree from Sharif University of Technology, Tehran, Iran, and the Ph.D. degree from Concordia University, Montreal, QC, Canada, all in Electrical Engineering in 2000, 2003, and 2007, respectively. From 2007 to 2010, he was a Postdoctoral Fellow in the Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada. He is currently an Assistant Professor in the Electrical Engineering Department, Sharif University of Technology, Tehran, Iran. His research interests include information theory, joint source-channel coding, and cooperative communications. Dr. Behroozi is the recipient of several academic awards including Ontario Postdoctoral Fellowship awarded by the Ontario Ministry of Research and Innovation (MRI), Quebec Doctoral Research Scholarship awarded by the Government of Quebec (FQRNT), Hydro Quebec Graduate Award, and Concordia University Graduate Fellowship.

Mohammad Reza Aref was born in city of Yazd in Iran in 1951. He received his B.Sc. in 1975 from University of Tehran, his M.Sc. and Ph.D. in 1976 and 1980, respectively, from Stanford University, all in Electrical Engineering. He returned to Iran in 1980 and was actively engaged in academic and political affairs. He was a Faculty member of Isfahan University of Technology from 1982 to 1995. He has been a Professor of Electrical Engineering at Sharif University of Technology since 1995 and has published more than 260 technical papers in communication and information theory and cryptography in international journals and conferences proceedings. His current research interests include areas of communication theory, information theory and cryptography with special emphasis on network information theory and security for multiuser wireless communications. At the same time, during his academic activities, he has been involved in different political positions. First Vice President of I. R. Iran, Vice President of I. R. Iran and Head of Management and Planning Organization, Minister of ICT of I. R. Iran and Chancellor of University of Tehran, are the most recent ones.

Statistical Analysis of Different Traffic Types Effect on QoS of Wireless Ad Hoc Networks

Mahmood Mollaei Gharehajlu *
Department of Electrical and Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran
m.mollaei@ee.kntu.ac.ir

Saadan Zokaei
Department of Electrical and Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran
szokaei@eetd.kntu.ac.ir

Yousef Darmani
Department of Electrical and Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran
darmani@eetd.kntu.ac.ir

Received: 08/Jul/2014 Revised: 20/Sep/2014 Accepted: 22/Oct/2014

Abstract
IEEE 802.11 based wireless ad hoc networks are highly appealing owing to their needless of infrastructures, ease and quick deployment and high availability. Vast variety of applications such as voice and video transmission over these types of networks need different network performances. In order to support quality of service for these applications, characterizing both packets arrival and available resources are essential. To address these issues we use Effective Bandwidth/Effective Capacity theory which expresses packet arrival and service model statistically. Effective Bandwidth asymptotically represents arrival traffic specifications using a single function. Also, Effective Capacity statistically describes service model of each node. Based on this theory, at first we modeled each node's service as an ON/OFF process. Then a new closed form of Effective Capacity is proposed which is a simple function and is dependent on a few parameters of the network. Afterward the performance of different traffic patterns such as constant bit rate, Poisson and Markov Modulated Poisson process are statistically evaluated in the case of both single and aggregate traffic modes. Using the proposed model we will show that traffic pattern affects QoS parameters even if all models have the same average packet arrival rate. We prove the accuracy of our model by a series of simulations which are run using NS2 simulator.

Keywords: Effective Bandwidth; Effective Capacity; Performance; CBR; Poisson; Markov Modulated Poisson Process.

1. Introduction

Recently, wireless ad hoc networks became popular, because of its low cost deployment, mobility support and high data rate. Based on these facts, different applications with different quality of service (QoS) requirements run over them. Lost sensitive applications such as web browsing and email and also time critical services, like voice and video traffics are examples of these applications. Providing an assured level of QoS for them need an accurate network performance evaluation.

IEEE 802.11 [1] is the accepted standard which is broadly used in wireless ad hoc networks. It has two major functions: Distributed Coordinate Function (DCF) and Point Coordinate Function (PCF). The former is the basic random access method that is used in both ad hoc and infrastructure wireless networks. PCF is an optional mode of IEEE 802.11 MAC layer that uses a centralized protocol.

In turn, DCF has two operation modes, Basic mode and RTS/CTS[1] mode. In DCF basic access scheme, the source node that has data frame to send, senses the channel and if it is idle (for more than DIFS[2] time), it transmits the frame. Otherwise, the transmission is postponed for a *back off* time which is a random interval uniformly distributed between $[0, cw]$ slot times where cw is the contention window size. A timer is set by that time and it decrements if the channel is idle and it freezes when the channel is busy. The source node commences to transmit if the *back off* timer becomes zero.

Performance evaluation of wireless ad hoc networks has been studied by researchers by simulation or analytical analysis. The majority of studies are under saturated circumstance which means that every node always has a packet to send. Under that condition, traffic characteristics such as inter arrival time and packet burstiness are not taken into account.

Bianchi [2] proposed a two dimensional Markov model under saturation condition. He computed the collision probability, and evaluated the throughput as a QoS parameter. Apart from his work, the average packet

[1] Request To Send/Clear To Send

[2] DCF Interframe Space

* Corresponding Author

delay is analyzed under saturation condition in [3], [4] and [5]. Since, the majority of internet applications exhibit ON/OFF characteristics [6], saturation condition is not a perfect model for this type of traffic. Assuming Poisson process as packet arrival model, [6-8] consider unsaturated condition by adding an idle state to Bianchi's Markov model.

Ref. [9] argues that internet traffic packet arrival properties and its distribution differ from Poisson process, so it is better modeled by self-similar processes. These processes are not easily applicable and tractable in Markov model due to its complexity.

Moreover, the Markov model only considers the average values of QoS parameters that might not be efficient for time critical multimedia applications [10]. In order to consider QoS boundaries, statistical performance analysis is proposed. Based on this approach, the most important QoS metrics will be the probabilities of exceeding a predefined delay and buffer size bound. Effective Bandwidth/Effective Capacity is an asymptotic statistical approach with sufficient accuracy for our purposes.

For different traffic models such as Constant Bit Rate (CBR), Poisson and Markov Modulated Poisson process effective bandwidth function is introduced. The Effective bandwidth theory characterizes traffic specification using a single function. It quantifies the service rate in order to have a specific queue overflow probability when random traffic is serving. Ref. [11] provides detailed information about it. On the contrary, Effective Capacity specifies queue overflow probability in the case of time varying service rate when constant bit rate traffic is serving. Wu and Negi [12] show that Effective Capacity is the dual of Effective Bandwidth. They proposed new Effective Capacity model for a wireless fading channel and introduced an accurate estimation algorithm to compute Effective Capacity. It should be noted that the proposed model in [12] does not consider multiple access, and all results are reported for a network consists of two nodes.

Assuming Markov Modulated Poisson Process (MMPP) as service model, [10] presented Effective Capacity for an IEEE 802.11 DCF shared channel. The proposed Effective Capacity is derived due to the duality between the Effective Bandwidth and the Effective Capacity. One of the disadvantages of the proposed model is that the derived Effective Capacity does not have a closed form function and it should be solved numerically. Also, traffic load affects the accuracy of the model.

Kafetzakis et al. [13] assumed an IEEE 802.11 station as an On/Off Semi-Markovian bursty server and derived the Effective Capacity which is suitable for highly loaded WLAN. However, for unsaturated condition, their approach needs to measure many extra parameters in order to examine Effective Capacity. Moreover, in order to apply their model in call admission control protocols, exchanging many extra signalings are needed.

In this paper, each node is assumed as an ON/OFF process model. During the ON time, the node has full access to the channel and transmits its packets with the maximum channel rate. This duration depends on the packet size and the channel data rate. However, during the OFF time, the node that has a data frame ready to be sent in its buffer waits until it senses the channel as idle. The OFF interval depends on the number of active nodes in the network, their traffic patterns, collision probability, minimum contention window etc. Based on these assumptions, the proposed Effective Bandwidth for ON/OFF process in [14] and its duality introduced in [12], we propose a novel Effective Capacity model that uses a few parameters, for IEEE 802.11 ad hoc networks. Unlike the proposed models in [10] and [13], the introduced Effective Capacity is closed form that depends on average service time of the node, channel capacity and packet size. In order to calculate the average service time of each node, a new Markov model for *back off* time under unsaturated condition is used.

Using the recommended Effective Capacity, the effect of different traffic models on delay are investigated and compared statistically. A stochastic bound is estimated for CBR, Poisson and MMPP in single and aggregate modes of operations. Average arrival rates of all traffics are assumed to be equal. Analytically, it is shown that as the traffic burstiness increases, the stochastic bound increases as well.

To validate our analytical results, extensive simulations are done using NS2-simulator [15]. The simulations result demonstrated the accuracy of the proposed model. Unlike the previous proposed models, our Effective Capacity model depends on a few parameters and it has a closed form. They are the advantages of our model. The proposed model can be used in distributed QoS provisioning and guaranteed based protocols such as call admission control and QoS aware routing algorithms.

The rest of the paper is organized as follows: Section II gives a brief overview of the Effective Bandwidth/Effective Capacity theory. Markov model for *back off* algorithm of IEEE 802.11 DCF mode is introduced in section III and the average service time is evaluated by that model. In section IV the Effective Capacity is proposed. The statistical delay bound for different traffic types are evaluated analytically and are validated by providing simulations in section V. Section VI concludes this paper.

2. Effective Bandwidth / Effective Capacity Theory

As mentioned, the source traffic is statistically modeled by Effective Bandwidth. Assume that A (t) is the amount of arrived traffic during [0, t). According to the Effective Bandwidth theory, it has been assumed that A (t) has stationary increments [11]. Suppose that log-moment generating function of A (t) is defined asymptotically as:

$$\Lambda(u) = \lim_{t\to\infty} \frac{1}{t} \log E\left[e^{uA(t)}\right] \tag{1}$$

And $\Lambda(u)$ exists for all $u \geq 0$. The Effective Bandwidth of A(t) for all $u \geq 0$ is given as [11-12] :

$$\alpha(u) = \frac{\Lambda(u)}{u} \qquad u \geq 0 \tag{2}$$

Now consider a server with a constant average service rate, μ, that serves A(t) as shown in Fig. 1.

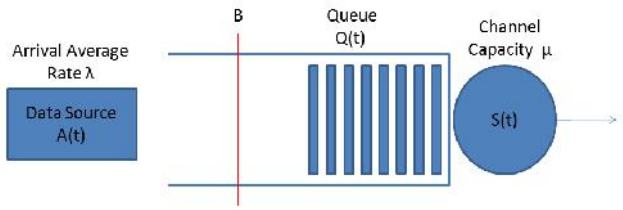


Fig. 1 A server with A(t) as arrival process and S(t) as service process

Q(t) is the queue length at t and S(t) is the number of served bits in [0,t). Based on the proposed theorem, the queue length bound violation probability is shown by [16]:

$$\sup_t \Pr\{Q(t) > B\} \approx e^{-\theta_B(\delta)B} \qquad B \to \infty \tag{3}$$

For smaller value of B, formula (4) is more accurate [12]:

$$\sup_t \Pr\{Q(t) > B\} \approx \eta(\delta) e^{-\theta_B(\delta)B} \qquad B \to \infty \tag{4}$$

In both equations, $\theta_B(\delta)$ and $\eta(\delta)$ are depended to server rate. It should be noted that $\eta(\delta)$ is the probability of having a non-empty queue. $\theta_B(\delta)$ is called QoS exponent value [12], and is the solution of:

$$\alpha(\theta) = \delta \tag{5}$$

Assuming $D(t)$ as packet delay at t, the probability of exceeding a predefined delay value, $D_{\max}$, is given by:

$$\sup_t \Pr\{D(t) > D_{\max}\} \approx \eta(\delta) e^{-\theta(\delta) D_{\max}} \tag{6}$$

Where $\theta(\delta) = \theta_B(\delta) * \delta$.

$\alpha(0)$ and $\alpha(\infty)$ are the average and maximum arrival rates of A(t). Conceptually, (6) shows that to have an expected delay violation probability equal to ε, the minimum value of serving rate of the server in Fig. 1 should be δ which is the solution of $\varepsilon = \eta(\delta) e^{-\theta(\delta) D_{\max}}$ [12].

Now, suppose that the arrival rate is constant and equal to α. Also, S (t) is the sum of served bits during [0,t). By these assumptions, the Effective Capacity is given by:

$$\alpha_c(u) = \frac{-\Lambda_c(-u)}{u} \tag{7}$$

where $\Lambda_c(-u)$ is computed by [12]:

$$\Lambda_c(-u) = \lim_{t\to\infty} \frac{1}{t} \log E\left[e^{-uS(t)}\right] \tag{8}$$

Based on this definition delay violation probability is:

$$\sup_t \Pr\{D(t) > D_{\max}\} \approx \eta_c(\alpha) e^{-\theta_c(\alpha) D_{\max}} \tag{9}$$

where $\theta_c(\alpha)$ is the solution of $\alpha_c(\theta_c) = \alpha$.

Assuming the queue model in Fig. 1 with $\alpha(u)$ as the Effective Bandwidth of arrival packets and $\alpha_c(u)$ as the Effective Capacity of server, the delay violation probability is computed by (9) where α is the solution of $\alpha_c(\alpha) = \alpha(\alpha)$.

It should be noted that by considering these results, the violation probability of D(t) from $D_{\max}$ could be calculated for different arrival models.

3. *Back off* Markov Model and IEEE 802.11 Service Model

3-1- DCF overview

The DCF mode of IEEE 802.11 protocol is well suited for wireless ad hoc networks. The DCF is operated in both basic and RTS/CTS access modes. In the basic access method, any node who wants to send a data frame, listens to determine channel status. The node transmits the frame, if it founds the channel idle for a DIFS interval. Otherwise, the node defers the transmission for a random interval which is called *back off* time. In the *back off* state, the node set a timer by the *back off* time and it decrements if the channel senses idle for each slot time duration. When the node senses the channel busy, the timer freezes. The node starts its transmission when the timer expires. Receiver node sends an acknowledgment (ACK) frame if it detects error free frame. The node will arrange to retransmit the frame, if it does not receive an ACK frame after ACK-Timeout period.

The *back off* time is a random waiting time which is chosen from a uniformly distributed random variable in the interval [0, CW -1] slot time where CW is contention window size. $CW_{\min}$ and $CW_{\max}$ are the minimum and maximum sizes of CW , respectively. At the first transmission stage, CW is set to its minimal value, namely, $CW_{\min}$. After each unsuccessful transmission, contention window size is doubled and retransmission process is rearranged. CW value increases until $CW_{\max}$ reaches to the maximum retry limit stage, m' , of retransmission. After that, CW remains unchanged for $m - m'$ stages where m the maximum retransmission step is. The value of m and m' are defined in IEEE 802.11 standard [1]. To sum up, if we define w_i as the contention window size in the i th step:

$$w_i = \begin{cases} 2^i CW_{\min} & i \leq m' \\ 2^{m'} CW_{\min} & i > m' \end{cases} \tag{10}$$

In RTS/CTS mechanism in order to initiate a transmission process, the transmitter node sends an RTS frame if it senses the channel idle. The receiver node responds it by sending a CTS frame.

Both RTS and CTS frames contain an estimated value of transmission time. Therefore, every node that hears RTS or CTS frame, stop sending any requests during the data frame transmission. This mechanism is also called virtual carrier sensing and it partly solved hidden and exposed terminal problems. Consequently, it improves the performance of IEEE 802.11 MAC protocol.

3-2- *Back off* model

In order to compute each node's service time statistics, *back off* time should be modeled accurately. Therefore, we use our analytical model that is proposed in [18]. We introduced the model briefly in this subsection. We consider a wireless ad hoc network consists of N static nodes that are in the transmission range of each other. Assuming Poisson process as packet arrival model, all nodes are under unsaturated conditions. The collision probability is shown by P and is assumed fixed as [2]. Fig. 2 shows a two dimensional Markov model that is evolved from [2], [3] and [8]. Based on the figure, *back off* states are depicted by (*s(t),b(t)*) pair where *s(t)* and *b(t)* are *back off* step and *back off* timer value in each step, respectively. During the analysis, (*i,k*) is a pair of integers which expresses the value of (*s(t),b(t)*).

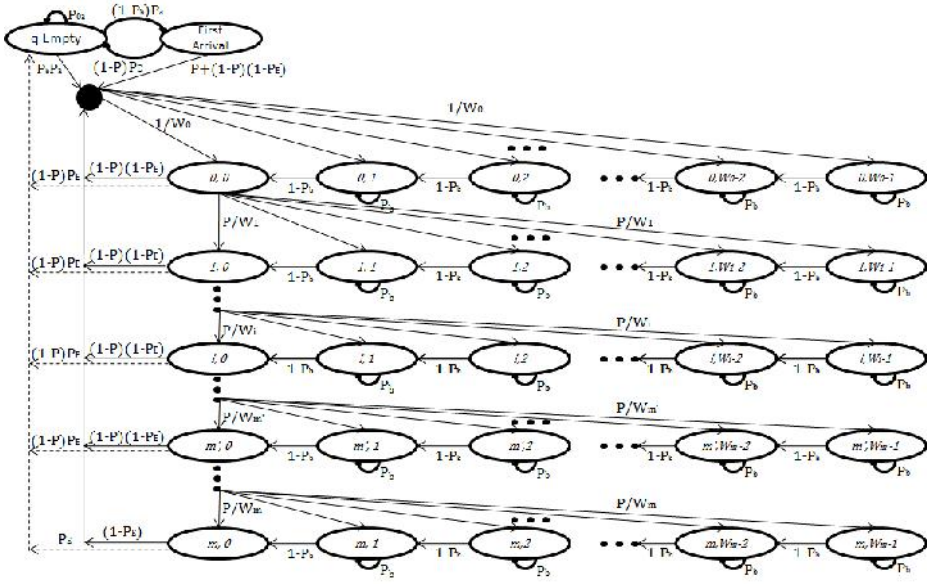


Fig. 2 *back off* Markov model

As Fig. 2 shows, i starts from zero at the first transmission attempt. Due to each unsuccessful transmission, i is incremented in each stage. m is the maximum value of i that is the maximum transmission limit. In order to consider unsaturated condition in our model, a new idle state, *qEmpty* , is introduced in Markov model. After each successful transmission, nodes enter to *qEmpty* state with the probability of P_E when their buffers are empty. λ and μ are the average arrival rate and average service time of each node, respectively. Also, we express P_{0a} as the probability of having no arrived packet and P_a as the probability of having at least one arrived packet during one slot time interval. When a node is in its *qEmpty* state, if a packet arrives, the node enters to its *FirstArrival* state.

Let $b_{i,k}$ be the probability of being in (*i,k*) state. In the steady state, $b_{i,k}$ can be found as:

$$b_{i,k} = \begin{cases} \dfrac{(1-(1-P_b)^{wi-k})}{w_i(1-P_b)P_b} * \left\{ \begin{array}{l} (1-P) * \rho * \\ \sum_{i=0}^{m-1} b_{i,0} + b_{m,0} * \rho \\ + b_{FirstArriv\,al} * (P \\ + \rho * (1-P)) \end{array} \right\} & i = 0 \\ \dfrac{(1-(1-P_b)^{wi-k})}{w_i(1-P_b)P_b} b_{i,0} & 0 < i \le m \end{cases} \quad (11)$$

In this equation, $\rho = \frac{\lambda}{\mu}$ is the server busy probability and P_b is the channel busy probability. Moreover, let consider $b_{q\,Empty}$ and $b_{First\ Arrival}$ as the probability of being in *qEmpty* and *FirstArrival* states, respectively. To satisfy the probability normalization condition:

$$\sum_{i=0}^{m} \sum_{k=0}^{wi-1} b_{i,k} + b_{qempty} + b_{FirstArrival} = 1 \quad (12)$$

Solving (11) and (12), the $b_{0,0}$ can be calculated. Considering $b_{0,0}$ in (13), transmission probability which is expressed by ^{2} , that is dependent on collision probability is given by:

$$^2 = \sum_{i=0}^{m} b_{i,0} + b_{FirstArrival} =$$
$$\left(\frac{1-P^m}{1-P} + \frac{(1-\,^\circ)*(1-P_b)}{1-(1-\,^\circ)*(1-P_b)*(1-P)} \right) * b_{0,0} \quad (13)$$

Also, collision probability can be stated as:

$$P = 1-(1-^2)^{N-1} \quad (14)$$

Collision and transmission probabilities are obtained by applying nonlinear solution in both (13) and (14).

3-3- Service time analysis

To analyze service time, we suppose each node has ON and OFF states. During the ON state, a node sends data frame in channel with full data rate. In OFF state, the node is in its *back off* state and it does not send any frame. Therefore, the average service time is computed by considering the duration of two successful transmissions. Fig. 3 shows this duration for an indexed node which is denoted by node A. As it is depicted in this figure, the time interval between *k*th and (*k+1*)th successful transmission of node A may contain V steps of transmission attempts. Let X_{Si} be a random variable time that node A spends in its *i*th *back off* step. During this interval, various events might be occurred such as:

- Idle time slot,
- Successful and unsuccessful transmission of all stations except A that makes channel busy,
- Unsuccessful transmission of station A.

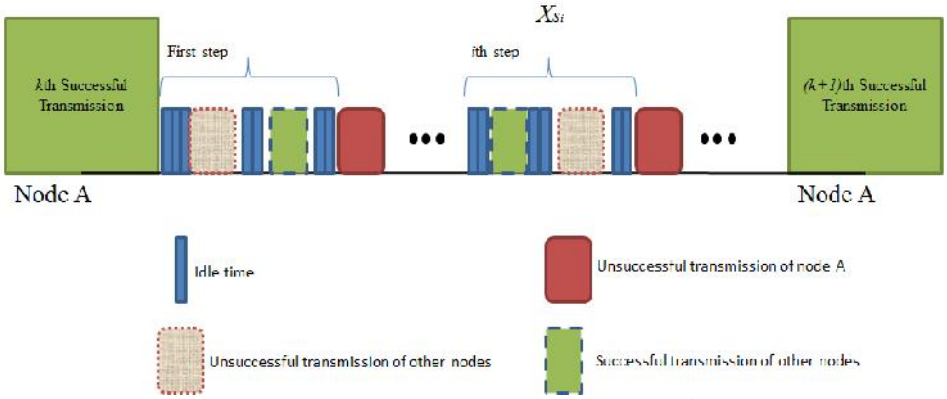


Fig. 3 Possible events during two successful transmission of node A

Therefore, X_{Si} is the spent time in the *i*th stage and can be defined as follow:

$$X_{Si} = N_s * \overline{T_s} \quad (15)$$

where $\overline{T_s}$ is the average slot time [2] that can be computed as:

$$\overline{T_s} = (1-P_{tr})*\pm + P_{tr} * P_s * T_{su}$$
$$+ P_{tr} * (1-P_s) * T_c \quad (16)$$

where P_{tr} is the probability of having at least one transmitting node and P_S is the successful transmission probability of each node. T_{su} is the channel busy period during a successful data packet transmission time and T_c is the busy period of the channel in the case of collision. Moreover, N_s is a random variable that is uniformly distributed with in [0, w_i -1] as the time slot number in step *i*. The average of X_{Si} is given by:

$$E[X_{Si}] = \frac{1}{\ll_{Si}} = \frac{W_i - 1}{2} * \overline{T_s} \quad (17)$$

i is a random variable with geometric distribution with parameter $q = 1-P$. By calculating the statistical average of $\frac{1}{\ll_{Si}}$, the average service time yields to:

$$\frac{1}{\ll_S} = \sum_{i=0}^{m} \frac{1}{\ll_{Si}} * P^i * (1-P) \quad (18)$$

4. Proposed Effective Capacity Model

As introduced in section II, to compute statistical delay bound by equation (9), an estimation of Effective Capacity is required. Also it has been mentioned that, there is a duality between Effective Bandwidth and Effective Capacity which is proven in [12]. The Effective Bandwidth of the most famous traffic models are calculated and proposed in [11,14].

To have an estimation of Effective Capacity, at first we model service time. We represented IEEE 802.11 node's service time as an ON/OFF model. Supposing service time as an exponential distribution and due to [7-8] assumption which is approximately precise, we can model each node service time as a Markov Modulated Fluid model (MMF) [14]. Fig. 4 illustrates a two state MMF for an IEEE 802.11 wireless node.

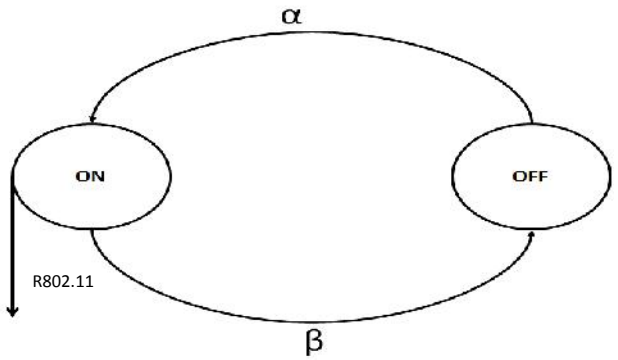


Fig.4 A two state Markov Modulated Fluid model

Fig. 3 clarifies the proposed Markov modulated fluid model in Fig. 4. As the Fig. 4 shows, there are two ON and OFF states. The state of the model alternates between ON and OFF. The average period of being in ON state is $\frac{1}{} = \frac{P_{size} * 8}{R_{802.11}}$ being in OFF state is $\frac{1}{\sim}$ which is the average service time of a node that is calculated by (18). Also the rate of fluid model is 0 and $R_{802.11}$ as the Markov process is in state OFF and ON, respectively. Moreover, based on the model the probability of being in ON and OFF state is $\frac{\sim}{\sim +}$ and $\frac{}{\sim +}$.

To sum up, $\sim$ is the changing rate of OFF to ON state.

By these assumptions, the recommended Effective Bandwidth for MMF [14] is given by (19)

$$\sim^{B}_{MMF}(u) = \frac{R_{802.11} * u - (\sim + \ )}{2 * u} + \frac{\sqrt{(R_{802.11} * u - (\sim + \ ))^2 + 4 * \sim * R_{802.11} * u}}{2 * u} \quad (19)$$

By applying the duality to equation (19), our Effective Capacity model can be derived as:

$$\sim^{c}_{802.11}(u) = \frac{(\sim + \ ) + R_{802.11} * u}{2 * u} - \frac{\sqrt{(R_{802.11} * u + (\sim + \ ))^2 - 4 * \sim * R_{802.11} * u}}{2 * u} \quad (20)$$

As (20) shows, and unlike [10,13] 's Effective Capacity the proposed Effective Capacity is depend on a few parameters namely $R_{802.11}$, P_{size} , and $\sim$. If both $R_{802.11}$ and P_{size} are constant, the equation is related to $\sim$. As explained, $\sim$ is dependent to average service time. Consequently, our proposed Effective Capacity is related to a parameter which could be measured by considering each packet service time in MAC layer. This is the best acheivement of our proposed Effective Capacity that could be used in most QoS guarantees approaches such as statistical call admission controls and statistical QoS aware routing protocols.

5. Simulation Results

In this section we support our Effective Capacity model by extensive simulations using NS2-simulator. In addition, the effect of different traffic models on delay bound are evaluated and compared with each other. In our simulations, a single hop WLAN with different number of nodes is considered where all of the nodes are in their transmission range. All nodes in all scenarios are randomly and uniformly distributed in 150*150 m^2 areas and their transmission range is 250 meters. Each node uses IEEE 802.11 as its MAC layer protocol with data transmission rate ($R_{802.11}$) equal to 2 Mbps. In all simulations, regarding to the number of active nodes (data frame transmitters) two scenarios are considered. In the first scenario two nodes, and in the second one, eight nodes send packets. Moreover, main and background flows are two types of traffics that are considered in each simulation. The background traffic is the same in all simulations and is assumed Poisson traffic. However, the main traffic is one of the CBR, Poisson, MMPP or the aggregated of them. Average arrival rate in both traffics, namely background and main traffics are assumed 32 Kbps and all the packet sizes are fixed and assumed 500 bytes. The reported results are the average of 10 time simulations where each simulation lasts for 500 seconds. Network parameters are summarized in Table 1.

Table 1: Network parameters

Network Parameters	
Number of active nodes	2 and 8
Network area	150*150 m^2
MAC layer protocol	IEEE 802.11
Maximum node speed	0 m/s
Drop policy	DropTail
Antenna type	Omni-directional

5-1- CBR traffic model

The CBR model transports traffic at a constant bit rate. Fig. 5 demonstrates the delay bound violation probability when the main traffic is assumed with that model versus $D_{\max}$. The figure compares simulation and analytical results. Analytical curves are obtained from equation (9) where $^{-c}(\mathrm{a})$ is the solution of equation (20) and a_{CBR} intersection, $\sim^{c}_{802.11}(^{-c}) = \mathrm{a}_{CBR}$. a_{CBR} is the average CBR packet arrival rate that is assumed as 32 Kbps. The curves show in both 2 and 8 active node cases, the probability of exceeding $D_{\max}$, decreases exponentially as $D_{\max}$ increases. The decreasing trend in both simulation and analytical curves are the same, especially when the value of $D_{\max}$ is less than 100 *ms*. In addition, the figure shows that the probability reduction in 2 nodes is more than in 8 nodes. That is because when the number of active nodes grows, it raises the queue delay and the collision probability which leads to increase the average packet delay. Consequently, delay violation probability is affected and increased.

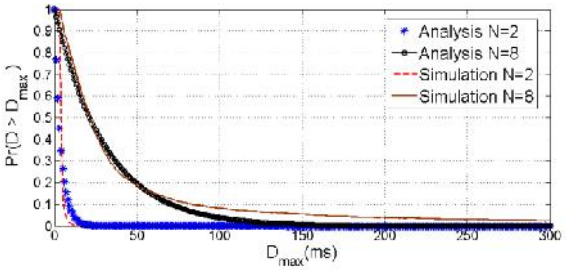


Fig.5 Analytical and simulation results for delay violation probability for CBR arrival model

5-2- Poisson traffic model

To evaluate the effect of Poisson traffic on delay violation probability, we use $\sim^{B}_{Poisson}(u)$ as [13]:

$$\sim^{B}_{Poisson}(u) = \frac{\mathrm{a}_{Poisson}(e^{u*P_{size}} - 1)}{u} \qquad (21)$$

where $\mathrm{a}_{Poisson}$ and P_{Size} are the Poisson average arrival rate and packet size, respectively. As explained, to compute delay exceeding probability by equation (9), $^{-c}(\mathrm{a})$ and λ should be obtained. To do this, $\sim^{B}_{Poisson}(u) = \sim^{c}_{802.11}(u)$ should be solved respect to u. The intersection of these equations is plotted in Fig. 6. Therefore, QoS exponent, $^{-c}(\mathrm{a})$ and λ will be computed analytically.

Fig. 7 compares analytical and simulation results when the main traffic arrival model is Poisson process. Like the CBR case, delay violation probability decreases exponentially. Moreover, as the figures depicted in both cases, analytical and simulation results are matched well which clarify the accuracy of our proposed Effective Capacity model.

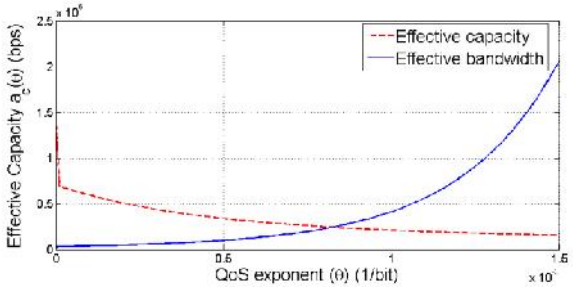


Fig.6 The intersection of proposed Effective Capacity and Effective Bandwidth of Poisson process

.

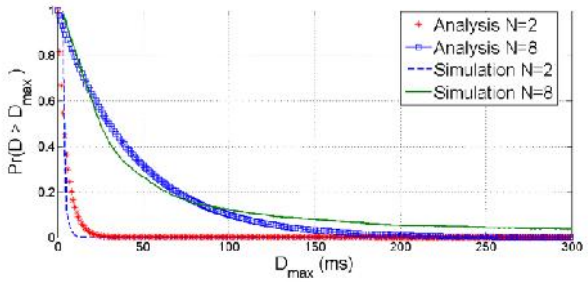


Fig.7 Analytical and simulation results for delay violation probability for Poisson arrival model

5-3- MMPP traffic model

MMPP process is a suited model for bursty traffic sources [17]. Based on this model, the arrival rate of packets changes as the Markov chain states are varied. In this paper we use an MMPP with two ON and OFF states. The Effective Bandwidth of this traffic model is given by [13]:

$$\sim^{B}_{MMPP}(u) = \frac{(e^{u*P_{size}} - 1)*\mathrm{a}_{MMPP} - (\sim + \quad)}{2*u} + \frac{\sqrt{((e^{u*P_{size}} - 1)*\mathrm{a}_{MMPP} - (\sim + \quad))^2 + 4*\sim*(e^{u*P_{size}} - 1)*\mathrm{a}_{MMPP}}}{2*u} \qquad (22)$$

where $\frac{1}{\sim}$ and $\frac{1}{\quad}$ are the average ON and OFF durations. P_{Size} is the packet size and a_{MMPP} is the average arrival rate of the packets during ON period. Based on these assumptions, average traffic rate is:

$$\sim^{B}_{MMPP}(0) = \frac{\mathrm{a}_{MMPP} * P_{Size} * \sim}{(\sim + \quad)}$$

Assuming traffic rate 32 Kbps, packet size of 500 bytes, and both $\frac{1}{\sim}$ and $\frac{1}{\quad}$ equal to 1 second, a_{MMPP} is obtained 15.38. Based on the Effective Bandwidth/Effective Capacity theory, to compute QoS exponent, equations (22) and (19) must be equal. Fig. 8 represents delay exceeding probability in both 2 and 8 active nodes versus $D_{\max}$. As the figure shows, simulation and analytical results are almost matched. However, they are not very similar as in the case of CBR and Poisson arrival model.

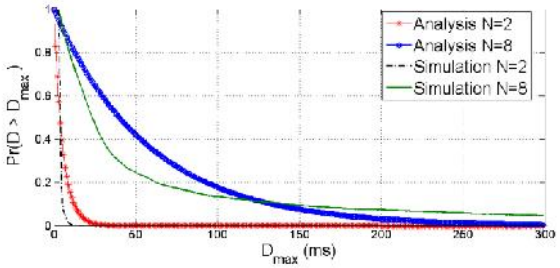


Fig.8 Analytical and simulation results for delay violation probability for MMPP arrival model

5-4- Aggregated traffic

In most computer networks, each node serves both single and aggregated flow at the same time. Therefore, providing assured level of QoS is also important in the case of aggregated traffic. It has been shown that the Effective Bandwidth of aggregated traffic is sum of the Effective Bandwidth of each single traffic flow. In other words, the aggregated Effective Bandwidth of CBR, Poisson and MMPP is given by:

$$\sim^{B}_{Aggregate}(u) = \sim^{B}_{MMPP}(u) + \sim^{B}_{Poisson}(u) + \sim^{B}_{CBR}(u) \qquad (23)$$

To evaluate upper delay bound for aggregated traffic, we use equation (9) where QoS exponent is obtained by solving $\sim^{B}_{Aggregate}(u) = \sim^{C}_{802.11}(u)$.

Fig. 9 compares analytical and simulation results when aggregated traffic arrives. The average input traffic is about 100 kbps. As the graph reveals, simulation and analytical results have the same trend that verifies our proposed model for aggregated traffic. However, there is an inconsistency between simulation and analytical results around 100 ms. Actually, this behavior has two major reasons. As explained in section 2, the delay violation probability is upper bounded by exponential function. The QoS exponent directly affects the value of this function. An accurate estimation of QoS exponent will eventuate upper bound precisely. The QoS exponent is estimated by $\sim^{c}(\theta) = \sim(\theta)$ where both $\sim^{c}(\theta)$ and $\sim(\theta)$ are obtained approximately. Therefore, having an exact evaluation of Effective Capacity and Effective Bandwidth leads us to have precise delay upper bound.

Also, as (6) reveals, an exponential function upper bounds the delay violation probability. It has been shown in [19] that this upper bound is not necessarily an exponential function. [19] Shows that this probability is given generally by:

$$\sup_{t} \Pr\{D(t) > D_{\max}\} \approx f(D_{\max}) \tag{24}$$

where f is a general function that could be calculated by summing many exponential functions. Therefore to have an accurate estimation of the upper bound, that general function must be considered.

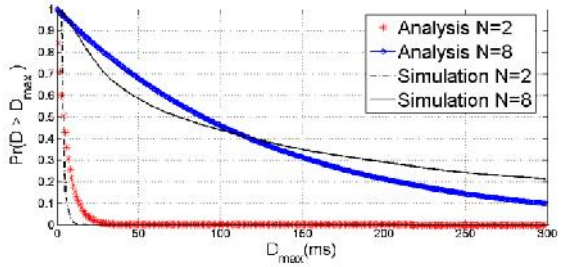


Fig.9 Analytical and simulation results for delay violation probability for aggregated traffic

Finally, Fig.10 is depicted to have comparison between delay violation probabilities of different traffic types where all traffic types have the same arrival rates. As the figure shows, the probability of exceeding from a predefined delay for MMPP is more than the CBR and Poisson process. For example, when $D_{\max}$ is 50 *ms*, delay exceeding probability for MMPP is about 0.6, while it is about 0.4 and 0.3 for Poisson and CBR traffic model, respectively. This comparison reveals that despite the same average arrival rates for different traffic models, the delay violation probabilities are different and strongly dependent to traffic model. MMPP is a bursty traffic model, and in spite of having the same average compare to CBR and Poisson, its delay probability is higher than the others. It can be concluded that the average arrival rate does not completely express a traffic pattern. Therefore, to provide QoS for a specific traffic type, more statistical information such as its arrival model is essential rather than relying on its average.

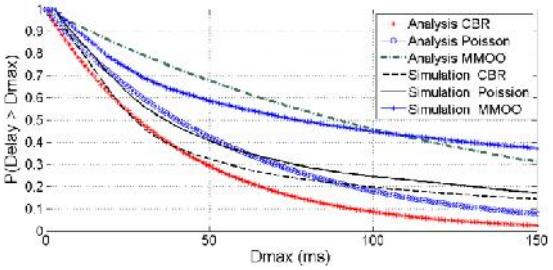


Fig.10 Analytical and simulation results for delay violation probability for CBR, Poisson and MMPP traffic models

6. Conclusion

In this paper the effect of different traffic sources on delay is investigated statistically. To address this issue we used Effective Bandwidth/Effective Capacity theory. To investigate IEEE 802.11 wireless node's service model a novel Effective Capacity is introduced. The effect of CBR, Poisson and MMPP model in both single and aggregated modes are considered. In all reported results, both analytical and simulation results were well matched which prove the accuracy of introduced Effective Capacity model. Also, the results show that despite the same arrival rate for all investigated traffic models, delay bounds are different and depend on traffic pattern types. Due to the burstiness nature of this traffic type, MMPP suffers from the worst delay bound among the other schemes. We conclude that considering the average traffic to provision a QoS for a specific type of traffic is not sufficient, and the traffic model is also essential to be considered.

References

[1] ANSI/IEEE, Std. 802.11 Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer Specifications, IEEE, (1999)

[2] Bianchi, G., "Performance analysis of the IEEE 802.11 distributed coordination function," IEEE JSAC, vol. 18, no. 3, pp. 535-547, Mar. 2000

[3] Ziouva, E. and Antonakopoulos, T., "CSMA/CA performance under high trafficc conditions: throughput and delay analysis," Computer Communications, 25, 313-321. (2002)

[4] Zhai, H., Kwon, Y. and Fang, Y., "Performance analysis of IEEE 802.11 MAC protocols in wireless LAN," Wireless Communications and Mobile Computing, 4, 917-931.(2004)

[5] Tadayon, N., Askari, E., Aissa, S. and Khabazian, M., "A Novel Analytical Model for Service Delay in {IEEE} 802.11 Networks," Systems Journal, IEEE, 6, 627-634.(2012)

[6] Malone, D., Duffy, K. and Leith, D., "Modeling the 802.11 Distributed Coordination Function in Nonsaturated Heterogeneous Conditions," IEEE/ACM Transactions on Networking, 15, 159-172.(2007)

[7] Lee, W., Wang, C. and Sohraby, K., "On use of traditional M/G/1 model for IEEE 802.11 DCF in unsaturated traffic conditions," in Wireless Communications and Networking Conference, 2006. WCNC 2006. IEEE, vol. 4, 2006, pp. 1933–1937.

[8] Dong, L.F., Shu, Y. T.,Shen, H. M. and Ma, M. D., "Packet Delay Analysis of Analysis on IEEE 802.11 DCF Under Finite Load Traffic in Multihop Ad hoc Networks," Science in China Series F: Information Sciences, Springer-Verlag, 51, 408-416.(2008)

[9] Paxson, V. and Floyd, S., "Wide-Area Traffic: The Failure of Poisson Modeling," IEEE/ACM Transactions on Networking, vol. 3 no. 3, pp. 226-244, June 1995

[10] Abdrabou, A. and Zhuang, W., "Stochastic delay guarantees and statistical call admission control for IEEE 802.11 single-hop ad hoc networks," *Wireless Communications, IEEE Transactions on* , vol.7, no.10, pp.3972,3981, October 2008

[11] Kelly, F. P., Zachary, S. and Zeidins, I., Eds., "Notes on effective bandwidth," vol. 4. Oxford University Press, 1996.

[12] Wu, D., and Negi, R., "Effective capacity: a wireless link model for support of quality of service," *Wireless Communications, IEEE Transactions on* , vol.2, no.4, pp.630,643, July 2003

[13] Kafetzakis, E., Kontovasilis, K., and Stavrakakis, I., "A novel Effective Capacity-based framework for providing statistical QoS guarantees in IEEE 802.11 WLANs", Computer Communications, vol. 35, Issue 2, pp. 249,262, January 2012,

[14] Courcoubetis, C. and Weber, R., "Effective Bandwidths for Stationary Sources," *Probability in Eng. and Info. Sci.,* vol. 9, no. 2, pp.285 -294 1995

[15] NS, "ns Network Simulator," URL: http://www.isi.edu/nsnam/ns/

[16] Chang, Ch., Thomas, J.A., "Effective bandwidth in high-speed digital networks," Selected Areas in Communications, IEEE Journal on , vol.13, no.6, pp.1091,1100, Aug 1995

[17] Dattatreya, G. R.," Performance Analysis of Queuing and Computer Networks, " Chapman and Hall/CRC, 2008

[18] Mollaei, M., Darmani, Y., Zokaei, S. "Delay Analysis of IEEE 802.11 Based Ad-Hoc Network under Unsaturated Condition" 2014, DOI: 10.1007/s11277-014-1940-7.

[19] Jiang, Y., Liu, Y., "Stochastic Network Calculus," Springer, 2008

Mahmood Mollaei Gharehajlu received the B.Sc. degree in Electrical Engineering from Tabriz University, Tabriz, Iran, in 2004 and the M.Sc. degree in Electrical Engineering from K. N. Toosi University of Technology, Tehran, in 2008. He is currently working toward his Ph.D. degree at Department of Electrical and Computer Engineering, K. N. Toosi University of Technology, Tehran. His current research interests include performance analysis of ad hoc networks and quality of service provisioning for multi hop ad hoc wireless networks.

Saadan Zokaei received the Master's degree in Electrical Engineering from the University of Tehran, Tehran, Iran, and the Ph.D. degree in Electrical Engineering from the Department of Communication and Information Technology, University of Tokyo, Tokyo, Japan, in 1994.
He is currently an Associate Professor with the Department of Electrical and Computer Engineering, K. N. Toosi University of Technology, Tehran. His research interests include information security, wireless networks, and next-generation networks.

Yousef Darmani received his B.Sc. In Electronics from the Science and Technology University, Tehran, Iran in 1987 and his M.Sc. in Digital Electronics from the Sharif University, Tehran Iran in 1991. Then he joined the Electrical Engineering Department of K. N. Toosi University of Technology. He received his Ph.D. in Computer Networking from the University of Adelaide, Adelaide, Australia in 2004. Currently he works as an assistant professor in K. N. Toosi University of Technology. His research interests are VOIP, real time communication over the Internet, Wireless and Ad-hoc networks and their protocols and computer hardware.

Fusion of Learning Automata to Optimize Multi-constraint Problem

Sara Motamed*
Department of Computer Engineering, Fuman Branch, Islamic Azad University, Fuman, Iran
Samotamed@yahoo.com

Ali Ahmadi
Department of Computer Engineering, K.N. Toosi University of Technology, Tehran, Iran
ahmadi@eetd.kntu.ac.ir

Received: 23/Sep/2013 Revised: 11/Oct/2014 Accepted: 12/Nov/2014

Abstract
This paper aims to introduce an effective classification method of learning for partitioning the data in statistical spaces. The work is based on using multi-constraint partitioning on the stochastic learning automata. Stochastic learning automata with fixed or variable structures are a reinforcement learning method. Having no information about optimized operation, such models try to find an answer to a problem. Converging speed in such algorithms in solving different problems and their route to the answer is so that they produce a proper condition if the answer is obtained. However, despite all tricks to prevent the algorithm involvement with local optimal, the algorithms do not perform well for problems with a lot of spread local optimal points and give no good answer. In this paper, the fusion of stochastic learning automata algorithms has been used to solve given problems and provide a centralized control mechanism. Looking at the results, is found that the recommended algorithm for partitioning constraints and finding optimization problems are suitable in terms of time and speed, and given a large number of samples, yield a learning rate of 97.92%. In addition, the test results clearly indicate increased accuracy and significant efficiency of recommended systems compared with single model systems based on different methods of learning automata.

Keywords: Stochastic Automata with Fixed and Variable Structures; Discrete Generalized Pursuit Automata; Fusion Method; Parallel Processing**.**

1. Introduction

Learning automata is one of the important models of learning and the goal of this method is to determine the optimal actions in unpredictable situation. The most important application of learning automata is to estimate parameters [3] while it is used in identification of pattern and play theory [4, 5, 6, 7]. All researches up to late 1980 have been reviewed and discussed at this book (Narendra and Thathachar) [1, 10]. Numerous examples and applications of learning automata have been presented in [9]. Another important feature of learner systems is their potential to improve efficiency with time. According to [2], the objective of the learner system is to improve a function which is not totally identified. Therefore, an approach to the problem is to decrease the objectives of the learner system which is defined on a set of parameters which aim at finding the optimum parameters set; reinforcement learning is a common method in finding the optimum objective. Main benefit of reinforcement learning compared with other learner methods are that it requires no information from the environment (than reinforcing signal) [1]. Stochastic learning automata are one of the reinforcement learning models which try to increase their efficiency and are divided into fixed structure stochastic automata (FSSA) and variable structure stochastic automata (VSSA) [8]. Alonso and Mondragon introduced a framework of limitations according to Markov decision making and optimization methods [16]. Similarly [17], reinforcement learning algorithms have been used for multi-factor systems and elements. In the later section, the paper presents an optimum method by the use of a learning method by using constraints based on Markov hypothesis. The reinforcement learning is defined in an environment using Markov model, by multi-objective functions and long term rewarding [18]. To prove their claim they applied this method to statistical plays with reward average limitations and obtained significant results. Also they explained the limited reinforcement learning method by stochastic and true rewards with a new view for optimization problems. The objective of this paper is to analyze the object by parallel computations and learning automata by heuristic algorithms. Cardei considered a special algorithm of learning automata which is called pursuit algorithm [7]. This algorithm pursues the current optimal action and if this action is not the one with the minimum penalty probability, this algorithm pursues a wrong action. This algorithm presents with two models of continuous and discrete. The continuous pursuit algorithm that used a reward and penalty learning paradigm, denoted CP_{RP}. Later [19] introduced the first discretized

* Corresponding Author

Pursuit estimator algorithm by presenting a discretized version, denoted DP_{RI} that uses a Reward-Inaction learning paradigm. The discretized generalized pursuit algorithm (DGPA) is another algorithm that generalizes the concepts of the Pursuit algorithm. In this algorithm all the actions have higher estimates than the current chosen action. This paper is used this DGPA to predict the best actions in unstable situations. In section 2, an effective model of partitioning is introduced by multi-constrained problems and also in sections 3 reviews the stochastic learning automata method. In the next parts of this section, fixed and variable structures and discrete generalized pursuit automata (DGPA) are considered. The fusion of DGPA is defined in section 4, to remove the problem of convergence in a very big space as a recommended method. Experimental results and conclusion are given in sections 5 and 6.

2. Partitioning by Multi-constraint

The main goal of this algorithm is to allocate a set of elements to elected classes and similar groups which are hopefully placed in the same class. This paper introduced an optimal model of classification in special situations. Also this paper is explained the proper model of partitioning to solve the multi constraint problems. Object allocation is a problem of partitioning a set of P with |P| elements in N classes. These elements have a certain capacity of limitations in each class [12]. This means that each class has limited capacity considering constraints for connecting joints. In the problem, it is supposed the relationship between P elements is weighted in a certain way and there is no connection of the object with itself. The aim in placing the processes on nodes is to find the shortest completion time in parallel applications. It is obvious that if ($|P| \leq |N|$), each node would host in more than one process. Partitioning the process set would be possible where ($|P|>|N|$) and is grown in a combinatory manner and computations of processes connection are defined by Narendra and Thathachar [12]. To express limitations and relations this part should apply conventions for external and internal relations. $X_{i,n}$ is a typical alternative of $X_{i,n} \in \{0,1\}$ and its P_i process is allocated to N_n node so the quantity is one, otherwise it is zero. It assumed that the external relation is formed from a node with the supposition that P_i has been allocated to this certain node [12]. Then $\sum_{j=1}^{|P|} (1 - X_{i,n}) w_{i,j}$ the external connection of P_i processes with all other processes of P_j would be formed. If each P_j is allocated to N_n node, then $X_{j,n} = 1$ and this process has no participation with the above sum. To find the external relation of the node, the sum of similarities should be added and then multiplied and rearranged. In the external relation limit calculated by Eq. 1 [12]:

$$\sum_{i=1}^{|P|}\sum_{j=1}^{|P|}(X_{i,n} - X_{j,n}X_{j,n})W_{i,j} \leq 1 \quad n = 1,...,|N| \qquad (1)$$

The above equation shows the set of processes and their connections from a subset to another. The only guiding quantity is a $w_{i,j}$ relation which indicates the P_i process on the node to P_j distant process which is not on the node. Therefore, the internal connection may be obtained by a similar formula, but by adding the connection for going from the distant process to the process on $w_{j,i}$, the result changes to Eq. 2 [12]:

$$\sum_{i=1}^{|P|}\sum_{j=1}^{|P|}(X_{i,n} - X_{j,n}X_{j,n})W_{j,i} \leq 1 \quad n = 1,...,|N| \qquad (2)$$

Set of constraints 3 limits total computation time for the processes allocated to each node with the node normalized capacity. Set of constraints 4 assures that a process only be allocated to a single node [12].

$$\sum_{i=1}^{|P|} X_{i,n}T_i \leq 1 \qquad n = 1,...,|N| \qquad (3)$$

$$\sum_{i=1}^{|N|} X_{i,n} \leq 1 \qquad n = 1,...,|P| \qquad (4)$$

Along a cycle, automaton picks a behavior and according to performance receives a response from the environment. The response may be a penalty or award [10]. Automaton obtains this response and knowledge form former behaviors and so selects its next move. The goal of the learning automaton is an optimum measure beyond the set of permitted behaviors. Automaton adjusts itself with the environment through learning optimum operation selection. Modeled learning sample is found in systems with insufficient knowledge about the environment and their startup by learning automaton applications. In FSSA, there is the property that output and transfer functions are not changing with time. The problem is based on the fact that static map of a subclass is obtained from learning automaton solutions and is used for solving object partitioning problems. For pairing and computation of w_i external connection for all processes, the node is selected with most violation from constraint 3. For instance P_A, which has been randomly selected among the processes on the node according to experimental distribution form their τ_i average weights, is allocated to this node. Then P_A process randomly selects another P_B process according to W_A distribution probability. A set of $\langle P_A , P_B \rangle$ processes are considered as a pair where pairing is said to be successful. If two processes belong to the same node, then these two processes receive a reward, unless the pairs are unsuccessful and are both penalized [12].

3. Stochastic Learning Automata

Learning automata are one of the important models of learning on unknown random environment. Each automaton has a finite set of input and certain probability of reward or penalty from environment. So the most important application of learning automata is to estimate parameters, while it is used in classification subjects, play theory and identification of pattern [10]. Learning automata are classified into the two groups of fixed and variable structure learning automata. In stochastic automata one action is selected randomly. Then the response of environment to this action is calculated by

probabilities. Now the new action is selected again and according to updated action probabilities, and the processors is repeated.

A stochastic automaton is defined by the sextuple {x, φ, α, p, A, G} and each parameter is:

- X is a input set.
- $\{\varphi_1, \varphi_2, \ldots, \varphi_r\}$ is a finite set of internal state.
- $\{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ is a finite set of output or response set.
- $p = \{p1, p2, \ldots, pr\}$ represents the action probability set.
- A is an algorithm which generates $p(n + 1)$ from $p(n)$.
-G is the output function.

For execution of training, the feedback signal from the environment, which triggers the updating of the action probabilities by the automaton, can be given by specifying an appropriate "error" function.

3-1- Fixed structure stochastic automata (FSSA)

Along a cycle, automata would pick a behavior and according to performance receives a response from the environment in a manner that the response may be a penalty or reward [10]. Automata obtain this response and knowledge from former behavior for expression of the next measure. The goal of the learning automaton is an optimum measure beyond the set of permitted behaviors. Automaton adjusts itself with the environment through learning optimum operation selection. Modeled learning sample by learning automaton applications is found in systems with insufficient knowledge about the environment and startup. In FSSA there is the property that output and transfer functions would change with time. The problem is that stable mapping a subclass of the learning automaton solutions and is used for solving the object partitioning problems. For pairing and calculation of external connection of w_i for all processes, first a node with most violation from region 3 is selected. A process allocated to this node, for instance P_A which has randomly selected among the processes on the node according to experimental distribution form their τ_i average weights. Then P_A process would randomly select another P_B process according to the possibility of W_A distribution. Set of $\langle \mathbf{P_A}, \mathbf{P_B} \rangle$ process is considered as a pair and called successful pairing. If the two processes belong the same node then these two processes would receive a reward unless the pairs are unsuccessful and both penalized [12].

3-2- Variable structure stochastic automata (VSSA)

VSSA is a replacement for FSSA and its transfer and output matrixes are changed in time [10, 13, 14 and 15]. They are defined as a possible operation vector P(K) where $P_i(K)$ is the possibility of i^{th} operation in the set of A operation, which is selected in K time from available |A| operations. As $\sum_i P_i(K) = 1$, for all Ks, updating the law for possibility vector would be continuous or discontinuous. The quickest learning automaton convergence belongs to the VSSA family. Adaptability of the family with automata to solve the specific problems may hopefully improve the speed of obtaining a solution. Thathachar and Sastry [16] introduced what is known as estimating algorithms. The main feature of these algorithms is that they maintain estimates of possible rewards for each operation and use them in possibility updating equations. In the first stage of functional, automaton selects an operation and the environment produces a response to this operation. According to the estimating algorithm answer, estimation of possible rewards for that operation is updated. Changes in possibility vector is P(K) operation based on $\hat{d}(k)$ and the vector being executed is estimated from the rewards possibility which is updated according to feedback from the environment. A model of random automata learning is DGPA, referred to in section 3-3.

3-3- Discrete generalized pursuit automata (DGPA)

Pursuit algorithm is a special type of estimator algorithm and it converged in statistic space rapidly. This model of pursuit algorithm works based on reward inaction learning paradigm and it updates the action of probability vectors if the environment rewards the chosen action. One of the problems in standard learning algorithms is their relatively slow convergence in selection of optimum operation in static environments for the removal of which various solutions have been introduced. One of the first solutions is disconnection of possibility space [10] in which possibility of operation selection can pick only certain quantities in the range of [0, 1]. On this basis, most standard algorithms are discrete. One of the existing problems in new models is premature convergence of learning algorithms with non-optimized operations. The root of these problems is in limiting their probability space. Thatachar and Sastry opened a new rout in their research by introducing estimator algorithms in line with their endeavors to improve learning algorithms convergence. The most important feature of such algorithms is in maintaining a continuous estimation from the possibility of receiving the reward for each operation and using it in updating automata equations. In other words, in the first stage of operation cycle, automaton selects an operation and then environment produces a response for it. According to this response, the estimating algorithm updates the reward possibility estimation for that operation. Pursuit automata are a group of estimating algorithms. As it is clear from its name, such algorithms are identified based on the fact that operation possibility vector encourages the operation which is currently considered as the best operation according to estimations. This is made with increasing the possibility of an operation wherein the reward possibility estimation is highest than other operations. Pursuit automata are divided in two continuous and discrete classes [11, 12]. The difference between these two algorithms is in updating the law for operation possibilities. According to the results, partitioning on the basis of pursuit automata with variable structure is effective for only small classes [10]. Therefore, the above problems cannot possibly be used as an ideal method for solving such problems. Our goal in this paper was to solve these problems by using the fusion of DGPAs. This model is called fusion of DGPAs (FDGPA) and is defined in section 4 to solve multi-constraint problems.

4. Fusion of DGPA (FDGPA)

An important problem in automata learning is their learning rate or their convergence speed equivalence. This is highly important for learning as it mostly changes slowly in the environment and learning processes should be completed in the environment before main changes appear, unless learning is ineffective. One idea is to use parallel operations because this method is considered as increased convergence speed. But if we have parallel learning automaton operations, each automaton would produce its operation and according to the signal, the environment would produce reinforcement in time. As there are multi responses, convergence speed would be raised. FDGPA is considered to be parallel instead of single learning automaton. Fig. 1 indicates our recommended algorithm. In Fig. 1, partitioning algorithm is applied to the input data for allocation of |P| elements to |N| classes with the constraints and it goes on to the next step. In that stage, data are divided to n classes so that they are given to processors for DGPAs which are applied to a parallel manner. As seen in Fig. 1, n parameters are the total number of operations in n={1,…,n} and P is the number of processors while in this case, each processor is an automaton. In our parallel model indexes are fixed and Current Primes are changing each moment while input vectors are divided by each processor in a certain number. Processors output is the automaton input to which DGPA learning is applied and the output for this step is the environment input while a cycle is repeated for the permitted number. In fusion section the set of operations is given by α and operation possibility vector p(K) is common by all n automata.

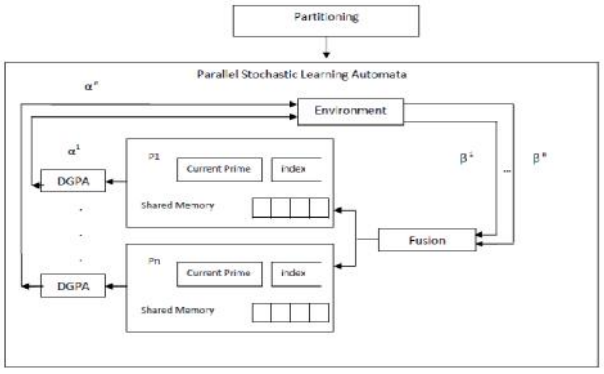


Fig.1. FDGPA model

For instance, each DGPA is based on the common operation possibility vector which is selected from (α^i (k)). This vector obtains its own reinforcement signal ((β^i (k)). Possibility vector for updated common operation is obtained based on all selected operations and obtained reinforcement learning. Fusion method uses all information obtained by the possibility of updating. The supposition is that β^j (k) $\square$ [0,1] is obtained for all i, k and the fusion vector is computed by the below equations [11]. To calculate the total response to α^j in Eq. 5:

$$q_i(k) = \sum_{j=1}^{n} {}^{j}(k) I\{\sim^{j}(k) = \sim_i\} \qquad (5)$$

In Eq. 6 is calculated the response to obtain the sum total:

$$q(k) = \sum_{j=1}^{n} {}^{j}(k) = \sum_{j=1}^{n} q_i(k) \qquad (6)$$

Output from fusion step is obtained from Eq.7:

$$p_i(k+1) = p_i(k) + \tilde{\lambda} \left(q_i(k) - q(k) p_i(k) \right)$$
$$i = 1, \ldots, r \qquad (7)$$

Where $\lambda \epsilon (0, 1]$ is learning parameter and $\tilde{\lambda} = \lambda / n$ is its normalized value. According to computations in eq.7, p(k) has been updated only once and does not require to be updated in each learning. Updated value p(k+1) is shared by all automata for selection of the next operation. This algorithm is fit for n sizes and for speeding up the rate of convergence.

5. Experimental Results

This paper studies the problem of partitioning a P set of **|P|** elements (or objectives) in **|N|** multiple discordant levels with the objective of having similar cluster elements. In this same level, objects can be connected in a multi-constraint (possible or discordant) method. This method has been formed by static mapping from allocation of a set of processes in parallel application to the set of computing nodes. In order to compare our fusion method with FSSA, multi-constraint partitioning is run on processes. We first presented a FSSA algorithm to solve multi-constraint problems. Solution is applicable but requires some centralized collaborations. A solution should be free from a centralized control mechanism. For this reason VSSA was presented. But this method is not ideal for such problems for sets with scattered data and is not applicable for big sets. There are two different models of pursuit learning algorithm which are called discrete and continuous. The differences between two models of pursuit algorithm are the updating rules for the actions probabilities.

For this reason is used the DGPAs fusion method. The software used for simulating the data is Matlab. Table 1 and 2 indicate learning rate of our recommended method in 200 iterations. Results indicate that fusion of stochastic learning automata algorithms give better results than single method stochastic learning automata algorithms. Plus by increasing the number of processors the rate of learning has increased.

Table 1: The results of single FSSA and fused FSSAs.

Number of nodes	***Number of processes***	***Fusion of DGPAs with two parallel processors***	***Fusion of DGPAs with three parallel processors***
2	4 6 8	100 99.90 99.12	100 100 99.50
4	8 12 16	98.66 98.51 98.43	99.14 99.10 99.03
6	12 18 24	98.40 98.37 98.05	99.06 98.48 98.12
16	64	97.03	97.92

Table1 indicates learning rate in single FSSA and fusion FSSAs and it can be understood that fusion of FSSAs give better result than single FSSA.

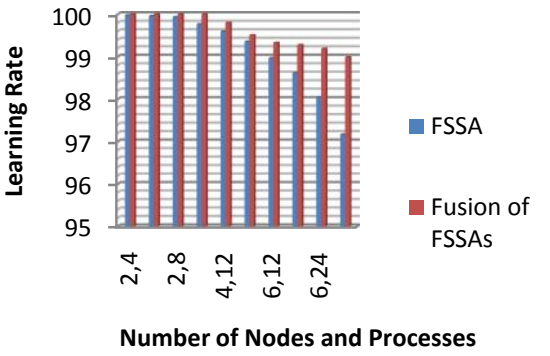


Fig.2. Comparison of single FSSA and fused FSSAs results on different nodes and processes

FSSAs. This figure indicates that fusion of FSSAs give better result than single FSSA.

Table 2: The results of fused DGPAs with two and three parallel processors.

Number of nodes	*Number of processes*	*FSSA*	*Fusion of FSSAs*
2	4 6 8	99.98 99.96 99.94	100 100 100
4	8 12 16	99.77 99.60 99.36	100 99.80 99.51
6	12 18 24	98.97 98.62 98.05	99.33 99.28 99.19
16	64	97.17	98.99

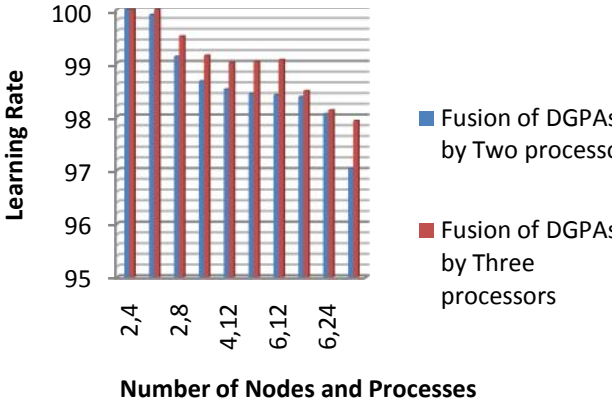


Fig.3. Comparison of the results from fused DGPAs with different processors

Fig. 3. Indicate the learning rate in fused algorithms of DGPAs with two and three parallel processors. As deduced, any more the number of processors, learning rate would be higher.

Fig. 4, 5 and 6 indicate partitioning of P elements in N class by applying proposed algorithm.

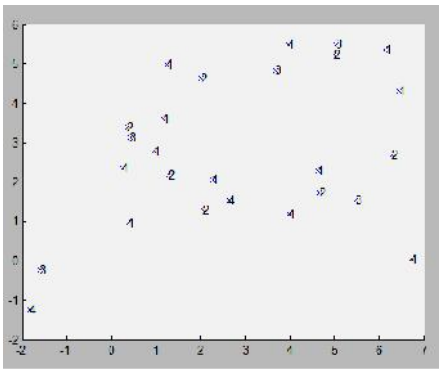


Fig.4. Partitioning the data with FDGPA for 4 nodes and 8 processors

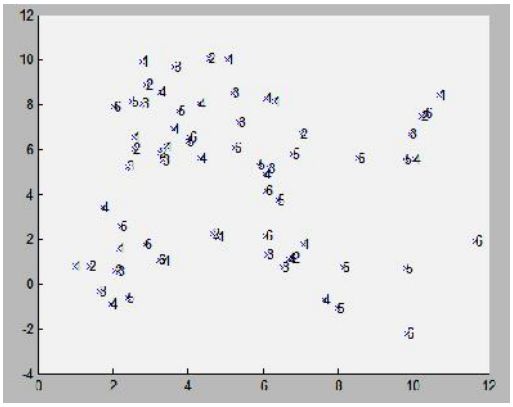


Fig.5. Partitioning the data with FDGPA for 6 nodes and 12 processors

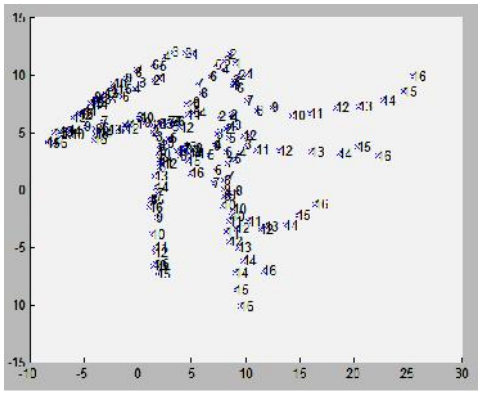


Fig.6. Partitioning the data with FDGPA for 16 nodes and 64 processors

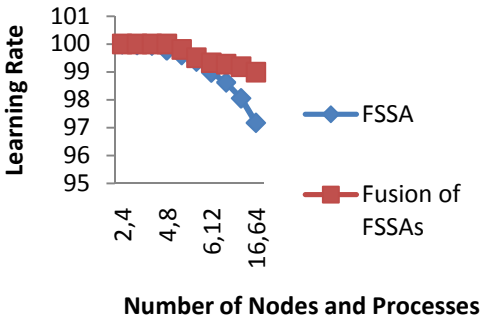


Fig.7. The results of fused and single model algorithms

This section presents a comparison of the performance of the different model of learning automata for using the partitioning the data by multi constraint rules. Looking at Fig. 7, it found that the fused of automata's algorithms have a higher learning rate than single model automata's algorithm. Furthermore, by adding the number of processors in FDGPA, our model's performance rises.

6. Conclusion

Over the last two decades, many new families of learning automata have emerged, with the class of estimator algorithms being among the fastest ones. All of those algorithms are divided into the two main groups of variable stochastic automata and fixed structure automata. One of the big advantages of variable structure automata has been the efficiency of this model in dynamic environment. So this paper is used a special model of variable structure automata is called DGPA algorithm. In contrast to the GPA algorithm, the DGPA always increases the probability of all the actions with higher estimates.

Then this paper is introduced a new model of variable stochastic learning automata by name of FDGPA. This model is used the fusion of DGPA to solve the partitioning of multi-constraint problems.

Review of the results in table 1 and 2 indicate that when there are 16 nodes and 64 processes fused learning rate would be 98.99 in fixed structure automata while the fused learning rate in stochastic learning automata with variable structure would be 97.92% using three processors with 200 iterations.

Overall, the proposed algorithm proved to be faster than the other algorithms in environments with more than two actions. When convergence speed is very low in stochastic automata learning algorithms with fixed and variable structures, and the number of samples increases, stochastic automata learning algorithm with variable structure would not be efficient.

Therefore, it can be claimed that the fused model outperforms the single model. Also, the number of processors increases in our parallel model as the learning rate climes. This method can be used in structural identification of pattern recognitions as a powerful classifier in further studies.

References

[1] K. S. Narendra, and M. A. L. Thathachar," Learning automata: An introduction", Prentice Hall, 1989.

[2] K.S. Narendra, and M. A. L Thathachar, "Learning automata a survey", IEEE Transactions on Systems, Man and Cybernetics, vol. 4, no. 4, 1974.

[3] E. Mance, and S. H. Stephanie, "Reinforcement learning: A tutorial", Wright Laboratory, 1996.

[4] M. R. Meybodi, and M. Taherkhani, "Application of Cellular Learning Automata to Modeling of Rumor Diffusion", Proceedings of 9th Conference on Electrical Engineering, 2001, pp. 102-110.

[5] H. Beigy, and M. R. Meybodi, "A Mathematical Framework for Cellular Learning Automata", Advances on Complex Systems, Vol. 7, No. 3, 2004, pp. 1-25.

[6] V. Raghunathan, C. Schurgers, V. Park, and M. B. Srivastava, "Energy-Aware Wireless Microsensor Networks", IEEE Signal Processing Magazine, Vol 19, 2002, pp 40-50.

[7] M. Cardei, and J. Wu, "Coverage inWireless Sensor Networks", Department of Computer Science and Engineering Florida Atlantic University Boca Raton, FL 33431, 2004.

[8] B.J. Oommen, and M. Agache, "A Comparison of Continuos and Discretized Pursuit Learning Scheme", IEEE, Carleton University, Ottawa, 1988.

[9] J. Winter, Y. Xu, and W. C. Lee, "Energy Efficient Processing of K Nearest Neighbor Queries in Locationaware Sensor Networks", Second International Conference on Mobile and Ubiquitous Systems, 2005.

[10] M.A.L. Thatachar, P.S. Sastry, "Varieties of Learning Automata: An Overview", IEEE transaction on system, man, and cybernetics- part B, 2002.

[11] G. Horn, B.J. Oommen, "Solving Multiconstraint Assignment Problems Using Learning Automata", IEEE transaction on system, man, and cybernetics- part B, 2002.

[12] K. S. Narendra, and M. A. L. Thathachar, "Learning Automata", Englewood Cliffs. NJ: Prentice-Hall, 1989.

[13] V. I. Varshavskii, and I. P. Vorontsova, "On the behaviour of stochastic automata with a variable structure", Remote Control, vol. 24, 1963, pp. 327–333.

[14] M. A. L. Thathachar, and P. S. Sastry, "A new approach to designing reinforcement schemes for learning automata", IEEE Trans, vol. SMC-15, no. 1, 1985, pp. 168–175.

[15] N. Abe, P. Melville, C.K. Reddy, C. Pendus , and D.L. Jensen, "Integrationg Data Modeling and Dynamic Optimization Using Constrained Reinforcement Learning", Mathematical Science Department, 2005.

[16] E. Alonso, E. Mondragon, "Associative Reinforcement Learning A Proposal to Build Truly Adaptive Agent and Multi-agent System", International Conference on Agents and Artificial Intelligence, 2013.

[17] E. S. Mannor, N. Shimkin, "A Geometric Approach to Multi-Criterion reinforcement Learning", Journal of Machine Learning Research 5, 2004.

[18] E. Euchibe, K. Doya, "Constrained Reinforcement Learning from Intrinsic and Extrrinsic Rewards". Okinava Institue Science and Technology, Japan, 2009.

[19] B.J. Oommen, and J. K. Lanctôt, "Discretized Pursuit Learning Automata", IEEE Trans. Syst. Man. Cybern., vol. 20, No.4, 1990, pp.931-938.

Sara Motamed received the B.Sc. from Kerman Azad University, Kerman, Iran and M.Sc. degree from Qazvin Azad University, Qazvin, Iran. Since 2012, she started her PhD in the field of artificial intelligence in Science and Research University, Tehran, Iran. She is interested in speech, image recognition and cognitive science. Now, she is lecturer at the faculty of computer engineering in Fuman Azad University, Fuman, Iran.

Ali Ahmadi received his B.Sc. in Electrical Engineering from Amirkabir University, Tehran, Iran and M.Sc. and Ph.D. in Artificial Intelligence and Soft Computing from Osaka Prefecture University, Japan in 2001 and 2004, respectively. He worked as a researcher in Hiroshima University, Japan during 2004–2007. He is currently assistant professor at K.N. Toosi University of Technology. His research interests are interactive learning models, virtual reality and artificial life, semantic data mining and information fusion, image search engines, and Software–hardware system co-design.

Extracting Rules from Imbalanced Data: The Case of Credit Scoring

Seyed Mahdi Sadatrasoul*
Department of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran
Sadatrasoul@iust.ac.ir

Mohammad Reza Gholamian
Department of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran
Gholamian@iust.ac.ir

Kamran Shahanaghi
Department of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran
Shahanaghi@iust.ac.ir

Received: 25/Mar/2014 Revised: 08/Aug/2014 Accepted: 15/Sep/2014

Abstract
Credit scoring is an important topic, and banks collect different data from their loan applicant to make an appropriate and correct decision. Rule bases are of more attention in credit decision making because of their ability to explicitly distinguish between good and bad applicants. The credit scoring datasets are usually imbalanced. This is mainly because the number of good applicants in a portfolio of loan is usually much higher than the number of loans that default. This paper use previous applied rule bases in credit scoring, including RIPPER, OneR, Decision table, PART and C4.5 to study the reliability and results of sampling on its own dataset.
A real database of one of an Iranian export development bank is used and, imbalanced data issues are investigated by randomly Oversampling the minority class of defaulters, and three times under sampling of majority of non-defaulters class. The performance criterion chosen to measure the reliability of rule extractors is the area under the receiver operating characteristic curve (AUC), accuracy and number of rules. Friedman's statistic is used to test for significance differences between techniques and datasets. The results from study show that PART is better and good and bad samples of data affect its results less.

Keywords: Credit Scoring; Banking Industry; Rule Extraction; Imbalanced Data; Sampling

1. Introduction

In today's competitive economy, credit scoring is widely used in banking industry. Every day, individual's and company's records of past borrowing and repaying actions are gathered and analyzed by information systems. Banks use this information to determine the individual's and company's profit. Application (credit) scoring is one of the main issues in the process of lending[1]. Credit scoring is used to answer one key question – what is the probability of default within a fixed period, usually one year. Credit scoring use banks historical loans data to classify customer as good or bad.

There are many techniques suggested to perform classification in the credit scoring problems including statistical and intelligent techniques. Logistic regression is the most favorite statistical and traditional method used to assess the credit scores[2]. Harrell applied Linear discriminant analysis and he shown that it is as efficient as logistic regression[3]. There are also many intelligent techniques applied to the problem including neural networks, Bayesian networks, support vector machines, case based reasoning, decision trees, and etc. Some studies have shown that neural networks, SVM, decision trees and other intelligent techniques, are superior to statistical techniques [4-6].

In recent years hybrid techniques are also proposed and they are the main focus of many researchers. Hybrid techniques usually use different algorithms strengths to improve the other algorithms weaknesses. In some hybrid techniques both statistical and intelligent techniques are used together. There are so many miscellaneous hybridization algorithms used in the literature. Lee et al used a hybrid neural discriminant technique with BP neural network and discriminant analysis, the hybrid model showed better accuracy than the BP neural network and discriminant analysis individually[7]. In another study Lee and Chen introduced a two-staged hybrid procedure with artificial neural networks and multivariate adaptive regression[8]. Tsai and Chen divide hybrid approaches into four main categories, they also organized 4 experiments with different combinations of clustering algorithms and classifiers; among their experiments logistic regression and neural network hybrid shown the best accuracy[9]. Huang, Chen and Wang studied using

* Corresponding Author

Meta heuristic techniques in order to tune intelligent techniques parameters, An application of support vector machines, genetic algorithms and F-score is studied and showed better results than using the pure SVM model[10]. In the last decade, using Ensemble techniques increased in the area and in some cases they give better accuracy rate[11, 12].West, Dellena and Qian used Neural network ensemble strategies including cross validation, bagging and boosting for financial decision applications, it shown better accuracy rate and generalization ability[11]. Ensemble learning is an open issue in recent year's studies[13, 14].

Because of robustness and transparency needs and also the auditing process done by regulators on the credit scoring in some countries, Banks cannot use many of mentioned techniques [15].By using rule bases, banks can easily interpret the results and explore the rejecting reasons to the applicant and regulatory auditors. There is actually a little literature in the field of rule based credit scoring. Ben-Davide provides a new method for rule pruning and examined his method on the credit scoring data set[16]. Hoffmann et.al introduced a new learning method for fuzzy rule induction based on the evolutionary algorithms[17].Martens et al used the support vector machine for rule induction in the credit scoring problems[18]. Malhotra et. al. used the adaptive neuro fuzzy inference systems(ANFIS) for rule induction and showed that this method works betters from discriminant analysis on their own credit scoring dataset which is gathered from credit unions[19], they used the back propagation method to learn their Rules membership function to fit on the data. Baesens et.al. use and evaluate three neural network rule extraction techniques including Neurorule, Trepan, and Nefclass, for rule extraction in three real life data bases including German credit database, Bene1 and Bene2 credit database[20]. They showed Nerorule and Trepan yield better classification accuracy compared to the C4.5 algorithm and the logistic regression. Finally they visualize the extracted rule sets using decision table[21].

In the credit scoring context, imbalanced data sets frequently occur as the number of good loans in a portfolio of loans are usually much higher than the number of loans that default[22]. Its reported that defaults ratio are ten percent of the whole bank's loan portfolio on average[23]. As mentioned practical studies show that the real credit scoring datasets are imbalanced. There are some but few studies which investigate imbalanced credit scoring data sets. Huang, Hung and Jiau proposed a strategy of data cleaning for handling imbalanced distribution of credit data is in order to avoiding problems of over fitting and relevance of classifiers[24]. Brown and Mues run several experiments based on different classifiers on five UCI and non UCI credit datasets, they balanced their samples on 70(good)/30(bad) [22]. Their experiments show that random forest and gradient boosting classifiers perform very well in the credit scoring context.

The aim of this paper is to use previous applied rule bases in credit scoring, including RIPPER, OneR, Decision table, PART and C4.5 to study the reliability and results of sampling on its own dataset. In order to extract invaluable rules bases the results are compared in terms of area under the receiver operating characteristic curve (AUC), Accuracy and number of rules.

This study is divided into four other major parts: section 2 describes the classification techniques used. Section 3 introduces the data, experiments settings, Section 4 discussed their results and finally study concluded in section 5.

2. Overview of Classification Techniques

This paper aims to extract the best rules from imbalanced data in the credit scoring context. For this purpose 5 rule based and tree induction (with the aim of rule induction) classifiers are selected. A brief description of these techniques is presented below.

2-1- C4.5

Decision trees split the data into smaller subsets using their nodes and at the end of each node there is a series of leaf nodes assigning a class to each of the observations. C4.5 build trees based on the concept of information theory[25].the entropy of a sample of K, can be computed by[25]:

$$\text{Entropy (k)} = -p_1 log_2(p_1) - p_2 log_2(p_2) \qquad (1)$$

Where $p_1(p_0)$ are the proportions of the class values 1(0) in the sample K, respectively. The attribute with the highest normalized information is used for division. The algorithm then occurs on the smaller subsets iteratively.

2-2- RIPPER

Repeated Incremental Pruning toProduce Error Reduction (RIPPER), is a rule based learning that builds a setof rules by considering minimizing the amount of error[26]. In the optimization step if the modified rule is better according to an MDL heuristic, rules are replaced with a modified one in order to reach a small rule set.

2-3- OneR

OneR is a one-level decision tree algorithm, which selects attributes one-by-one from a dataset and generates a different set of rules based on error rate. At last the attribute and its appropriate rule set with minimum error is selected[27].

2-4- Decision table

Decision Table algorithm build tables using a simple decision table majority classifier[28].It uses a 'decision table' to summarize the dataset. After finding the line in the decision table that fits the non-class values, a new data item is assigned a category. Then the wrapper method is employed to find a good subset of attributes for inclusion

in the table. The likelihood of over-fitting is reduced by eliminating attributes that contribute little or nothing to a model of the dataset and at last a smaller, well-defined decision table is reached.

2-5- PART

Partial decision tree algorithm (PART) is a developed version of RIPPER and C4.5[29]. Its main improvement is that it does not need to perform global optimization like C4.5and RIPPER to produce rules. It uses the standard covering algorithm to generate a decision list, and avoids over pruning by inducing rules from partial decision trees.

3. Empirical Evaluation

In this section first the data set characteristics is described. Secondly dataset samples are explained and finally the performance analyses are done**.**

3-1- Data sets characteristics

An Iranian commercial bank real export loan dataset is used to evaluate the proposed algorithm. Table (1) shows the characteristics of the dataset. The initial dataset include 1109 corporate applicants' and 46 financial and non financial data in the period from 2007 to 2012. First, the data cleaning is done; it includes removing redundant, outlier's data and missing values. There were a few missing Values for some corporate, some of them lack financial data and others lack the result of their loans, in fact they were in the process of debt repay, some of them haven't applied for loan yet. So 387corporate are excluded. From 722 remainedcorporate,652 are credit worthy (90.3%) and other 70 was unworthy (9.7%). Dummy variables were created for the categorical variables (ex. Type of industry).Using dummy variables number of variables increased to 55.Table (1) summarizes the dataset characteristics before and after cleaning step.

Table 1: dataset description

status	Data size	Inputs variables		
		Total	Continuous	Categorical
Before cleaning	1109	46	38	8
After cleaning	722	55	34	21

Delinquency status was defined by Basel committee definition of "default" and used to generate a 1/0 target variable for modeling purposes (good = 1, bad = 0). Accounts with no more than three months or more in arrears were classified as good. Those that were currently three or more months in arrears, or had been three months in arrears, were classified as bad. The results and descriptions of the variables used are shown in table (5) in appendix (1).

3-2- Re-sampling setup

Table (2) shows the main imbalanced dataset and samples built in order to consider imbalanced issue. The main dataset has a 90/10 class distribution, a 75/25 ratio in percent class distribution is selected for balancing the data and the main database is altered in different scenarios to meet this distribution. The two most common preprocessing techniques are random minority oversampling (ROS) and random majority under sampling (RUS). In ROS, instances of the minority class (bad applicants) are randomly duplicated in the dataset. In RUS, instances of the majority class (good applicants) are randomly discarded from the dataset.

In this study four different balanced datasets are created using two mentioned techniques. First using ROS bad instances are duplicated and the "Oversampled dataset" is created. This duplication is done until the distribution of good/bad meets to 75/25 so the number of bad instances increased from 70 to 217 samples. In another re sampling scenario, using RUS, three different "Under sampled datasets" are created. In order to use all of the datasets, simple random sample without replacement is done. The Under sampled dataset are designed in a manner that each good applicant in the main dataset is included in one and only one of three different under sampled datasets. This reduction is done until the distribution of good/bad meets nearly to 75/25 so the number of good instances decreased for these three under sampled datasets sequentially to 218,226 and 208 samples.

Table 2: Different samples of dataset used

Dataset name	Data size	Good	Bad	Good/All percent
Main imbalanced dataset	722	652	70	90.3
Oversampled dataset	869	652	217	75.02
Under sampled dataset No.1	288	218	70	75.74
Under sampled dataset No.2	297	226	70	76.9
Under sampled dataset No.3	278	208	70	74.82

3-3- Performance analysis

Five different measures are used to analysis the performance of the constructed rule bases. The performance criterion chosen to measure the effect of significant difference in number of observations is the area under the receiver operator characteristic curve(AUC) statistic[22]. Confusion matrix is another favorable instrument used in performance evaluations as shown in table (3).Overall accuracy, Good precision and bad precision are important measures after the ROC measure, as they shown the classifications quality as another dimension.

The overall accuracy of successfully identifying loans is computed using equation (2)

$$\text{Overall accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \quad (2)$$

Table 3: The confusion matrix

		PREDICTED CLASS	
ACTUAL CLASS		Class= Worthy	Class= Unworthy
	Class=Worthy	a(TP)	b(FN)
	Class= Unworthy	c(FP)	d(TN)

The precision of successfully identifying non-default loans is computed using equation (3)

$$\text{Good precision}=\frac{TP}{TP+FP} \quad (3)$$

The precision of successfully identifying default loans is computed using equation (4)

$$\text{Bad precision}=\frac{TN}{TN+FN} \quad (4)$$

Compactness of rules is another issue in rule base systems. At a defined level of ROC and accuracy measures for two rule bases, the rule base which has lower number of rules is preferred.

4. Results and Discussions

All the experiments in this paper are done using 10 fold cross validation. Table (4) shows classification accuracy, number of rules and area under curve for five datasets. The best classification accuracy, the lowest number of rules and area under curve for each data set are bolded. The best results for all of experiments are also underlined. Three groups of experiments are done and their results are presented below:

Table 4: Performance measures on different datasets and classifiers

dataset	Method	AUC	Accuracy(ALL)%	Precision(Bad)%	Precision (good)%	Number of rules
Main imbalanced dataset	RIPPER	0.531	89.47	**31.3**	90.8	2
	Decision table	0.499	**90.3**	0	90.3	**1**
	OneR	0.494	89.20	0	90.2	3
	PART	**0.612**	87.40	27.7	**91.6**	28
	C4.5	0.574	87.11	20.5	90.9	19
Over sampled dataset	RIPPER	0.881	87.45	72.3	93.3	**15**
	Decision table	0.887	80.21	57.5	92.3	575
	OneR	0.643	76.87	55.2	81.5	45
	PART	**0.941**	**90.22**	75.8	**96.2**	22
	C4.5	0.93	90.1	**76.1**	95.8	48
Under sampled dataset No.1	RIPPER	0.594	72.92	37.5	77.3	3
	Decision table	0.492	**73.95**	0	75.3	**1**
	OneR	0.544	73.61	40	**77.5**	7
	PART	**0.667**	72.22	**42.6**	71.4	22
	C4.5	0.595	69.79	36.1	78.9	24
Under sampled dataset No.2	RIPPER	0.517	73.99	34.8	77.3	**1**
	Decision table	0.511	**75.67**	25	76.4	**1**
	OneR	0.518	71.62	29.4	77.1	6
	PART	**0.656**	71.28	**38.8**	**80.8**	17
	C4.5	0.535	69.93	32.7	78.4	25
Under sampled dataset No.3	RIPPER	0.538	71.94	38.9	76.9	2
	Decision table	0.525	**73.02**	22.2	74.7	**1**
	OneR	0.504	71.22	27.3	75	7
	PART	0.581	71.58	**42.4**	**79.5**	20
	C4.5	**0.596**	68.70	38	79.2	20

4-1-First group experiments (Data sets performance comparisons)

First a test set at the 5% level of importance from the best performer using Friedman's testis done against different datasets for all of performance measures. Its findings are as follows:

- It shows that the results of oversampling data set have a significant difference compared to other four datasets; it can be seen that oversampling and increasing the number of observations increase the results performance compared with other reduction techniques at a defined level of good/bad ratio (75/25).
- The three under sampled datasets haven't any significant difference in their results; it can be concluded that different good observations in three different datasets don't have an importance issue in the results.

- The main dataset and three under sampled datasets haven't any significant difference; another separated Friedman test for AUC confirmed this hypothesis.
- Number of rules doesn't have significant change in all of the datasets and techniques, exclude decision table. It shows significant difference and increase in number of rules in oversampled dataset.

The results can be also used to evaluate sampling for different scenarios. Fig. (1) Shows that from one hand oversampling enhance the performance measures totally except number of rules and on the other hand under sampled datasets have no important difference in performance measures. It can be concluded that although the under sampled datasets have lower records but this do not affected their results comparing to the main imbalanced data set.

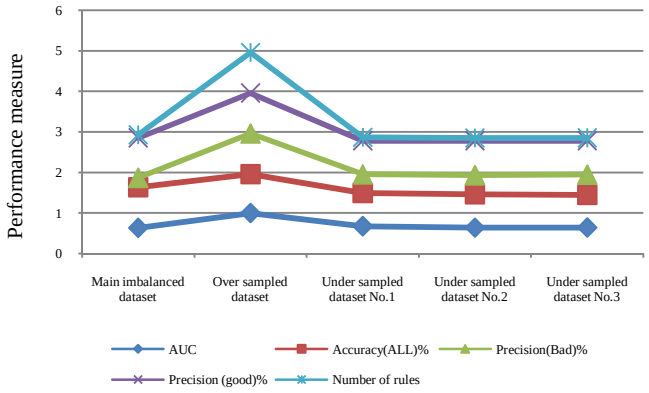


Fig. 1. Status of different performance measures for different samples (the results are standardized).

4-2-Second group experiments (Classifiers performance comparisons)

Second test set at the 5% level of importance using Friedman's test is done for different classifiers against all of performance measures. Its findings are as follows:

- The OneR, decision table and RIPPER haven't any significant difference between each other but have significant difference with other classifiers, they are the worst performers.
- The PART have significant difference with other classifiers, it is the best performer.

4-3-Third group experiments (Classifiers performance comparisons)

Third test set at the 5% level of importance using Friedman's testis done for different classifiers against three main performance measures. Its findings are sorted by their importance and presented below:

- **AUC measure:** The OneR, decision table and RIPPER haven't any significant difference between each other and they are the worst players, but PART and C4.5 have major difference with worst players and with each other. PART is the best performer throughout this measure.
- **Accuracy measure:** All of the classifiers haven't any significant difference between each other under the accuracy measure.
- **Number of rules measure:** The OneR, decision table and RIPPER haven't any significant difference between each other, they are the best players, also PART and C4.5haven't any significant difference between each other and they are the best players.

In brief, when considering different measures based on their importance it can be concluded that PART and after it C4.5 have a very good performance in different levels of class imbalance. However decision table, OneR and RIPPER are the worst performers. The mentioned results were attractive in the oversampled dataset and the results of two best classifiers on this dataset can be used for credit scoring classification. Fig (2) shows the results in brief.

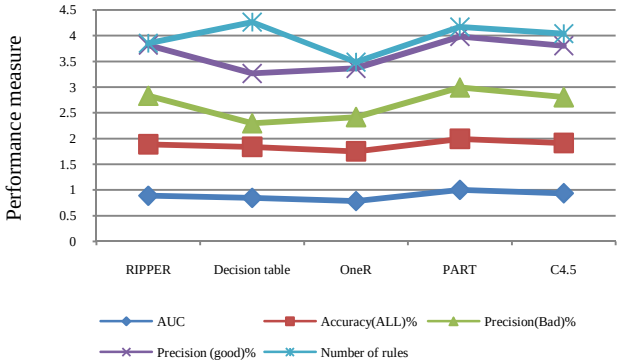


Fig. 2. Status of different performance measures for different classifiers (the results are standardized).

Table (5) shows the ranks against each performance measure for different classifiers. Note that statistical significant test do not checked for this table.

Table 5: Classifiers rank against different measures

Performance measure name	Classifiers rank
AUC	PART> C4.5>RIPPER> Decision table> OneR
Accuracy(ALL)%	RIPPER> Decision table>PART> C4.5> OneR
Precision(Bad)%	PART>RIPPER> C4.5> OneR> Decision table
Precision (good)%	C4.5> PART>RIPPER> Decision table> OneR
Number of rules	RIPPER> OneR>PART> C4.5> Decision table

5. Conclusion

In this paper, a number of different classifiers are used and compared on various balanced and imbalanced datasets. These techniques include RIPPER,C4.5, PART, OneR and Decision table. An imbalanced dataset from a major Iranian bank is applied and balanced using oversampling and several random under sampling techniques. Classifiers and datasets are compared using five different performance measures and Friedman's test. The results of the study shows that random oversampling of bad loans yield to better performance measures for all of the classifiers. It is also found that

PART classifier is perform better on imbalanced data than other classifiers and that it's the best performer at all of the experiments and performance measures except number of rules. On the other hand OneR and decision table techniques are the worst classifiers at all.

Next researches can focus on using other oversampling methods and their effect on the classifiers training. Studying the effect of different sampling methods on feature selection is also another open area of future researches.

Acknowledgement

The authors' kindly acknowledge MR. zekavat for his kind cooperation.

Appendix (1) Variables included in Iran credit dataset and their types are shown in table (6).

Table 6: list of variables in Iran commercial bank credit dataset

Variable	Type	Variable	Type
Net profit	Continuous	Type of industry: industry and mine (=1, other =0)	Categorical
Active in internal market	Categorical	Type of industry: agricultural (=1, other =0)	Categorical
number of countries that the company export to	Categorical	Type of industry: oil and petrochemical (=1, other =0)	Categorical
Sales growth	Categorical	Type of industry: infrastructure and service(=1, other =0)	Categorical
Target market risk (from 1 to 5)	Categorical	Type of industry: chemical (=1, other =0)	Categorical
Seasonal factors	Categorical	Year of financial ratio	Continuous
Company history(number of years)	Categorical	Type of book: Tax declaration(=1,other=0)	Categorical
Top Mangers history	Categorical	Type of book: Audit Organization (=1,other=0)	Categorical
Type of company: Cooperative (=1, other =0)	Categorical	Type of book: Accredited auditor (=1,other=0)	Categorical
Type of company: Stock Exchange(LLP) (=1, other =0)	Categorical	Inventory cash	Continuous
Type of company: Generic join stock(PJS) (=1, other =0)	Categorical	Accounts receivable	Continuous
Type of company: Limited and others (=1, other =0)	Categorical	Other Accounts receivable	Continuous
Type of company: Stock Exchange (=1, other =0)	Categorical	Stock	Continuous
Experience with Bank (number of years in 5 categories)	Categorical	Current assets	Continuous
Audit report Reliability	Categorical (binary)	Non-current assets	Continuous
Current period sales	Continuous	Total assets	Continuous
Prior period sales	Continuous	Short-term financial liabilities	Continuous
Two-Prior period sales	Continuous	Current liabilities	Continuous
Current period assets	Continuous	Long-term financial liabilities	Continuous
Prior period assets	Continuous	Non-current liabilities	Continuous
Two-Prior period assets	Continuous	Total liabilities	Continuous
Current period shareholder Equity	Continuous	Capital	Continuous
Prior period shareholder Equity	Continuous	Accumulated gains or losses	Continuous
Two-Prior period share holder Equity	Continuous	shareholder Equity	Continuous
checking accounts creditor turn over	Continuous	Sale	Continuous
checking Account Weighted Average	Continuous	Gross profit	Continuous
Average exports over the past three years	Continuous	Financial costs	Continuous
Last three years average imports	Continuous	y (nonworthy/worthy)	Categorical (binary)

References

[1] Van Gestel, T. and B. Baesens, Credit risk management: basic concepts: financial risk components, rating analysis, models, economic and regulatory capital. 2009: Oxford University Press, USA.

[2] Wiginton, J.C., A note on the comparison of logit and discriminant models of consumer credit behavior. Journal of Financial and Quantitative Analysis, 1980. 15(03): p. 757-770.

[3] Harrell, F.E. and K.L. Lee, A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. Biostatistics: Statistics in Biomedical; Public Health; and Environmental Sciences. The Bernard G. Greenberg Volume. New York: North-Holland, 1985: p. 333–343.

[4] Crook, J.N., D.B. Edelman, and L.C. Thomas, Recent developments in consumer credit risk assessment. European Journal of Operational Research, 2007. 183(3): p. 1447-1465.

[5] Huang, Z., et al., Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision support systems, 2004. 37(4): p. 543-558.

[6] Ong, C.S., J.J. Huang, and G.H. Tzeng, Building credit scoring models using genetic programming. Expert Systems with Applications, 2005. 29(1): p. 41-47.

[7] Lee, T.S., et al., Credit scoring using the hybrid neural discriminant technique. Expert Systems with Applications, 2002. 23(3): p. 245-254.

[8] Lee, T.S. and I. Chen, A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. Expert Systems with Applications, 2005. 28(4): p. 743-752.

[9] Tsai, C.F. and M.L. Chen, Credit rating by hybrid machine learning techniques. Applied soft computing, 2010. 10(2): p. 374-380.

[10] Huang, C.L., M.C. Chen, and C.J. Wang, Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, 2007. 33(4): p. 847-856.

[11] West, D., S. Dellana, and J. Qian, Neural network ensemble strategies for financial decision applications. Computers & operations research, 2005. 32(10): p. 2543-2559.

[12] Tsai, C.F. and J.W. Wu, Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Systems with Applications, 2008. 34 (4): p. 2639-2649.

[13] Louzada, F., et al., Poly-bagging predictors for classification modelling for credit scoring. Expert Systems with Applications: An International Journal, 2011. 38(10): p. 12717-12720.

[14] Finlay, S., Multiple classifier architectures and their application to credit risk assessment. European Journal of Operational Research, 2010.

[15] Thomas, L.C., Consumer credit models: pricing, profit, and portfolios. 2009: Oxford University Press, USA.

[16] Ben-David, A., Rule effectiveness in rule-based systems: A credit scoring case study. Expert Systems with Applications, 2008. 34(4): p. 2783-2788.

[17] Hoffmann, F., et al., Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms. European Journal of Operational Research, 2007. 177(1): p. 540-555.

[18] Martens, D., et al., Comprehensible credit scoring models using rule extraction from support vector machines. European Journal of Operational Research, 2007. 183(3): p. 1466-1476.

[19] Malhotra, R. and D.K. Malhotra, Differentiating between good credits and bad credits using neuro-fuzzy systems. European Journal of Operational Research, 2002. 136(1): p. 190-211.

[20] Baesens, B., et al., Using neural network rule extraction and decision tables for credit-risk evaluation. Management Science, 2003: p. 312-329.

[21] Baesens, B., et al., Using neural network rule extraction and decision tables for credit-risk evaluation. Management Science, 2003. 49(3): p. 312-329.

[22] Brown, I. and C. Mues, An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 2011.

[23] Dinh, T.H.T. and S. Kleimeier, A credit scoring model for Vietnam's retail banking market. International Review of Financial Analysis, 2007. 16(5): p. 471-495.

[24] Huang, Y.M., C.M. Hung, and H.C. Jiau, Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. Nonlinear Analysis: Real World Applications, 2006. 7(4): p. 720-747.

[25] Quinlan, J.R., C4. 5: programs for machine learning. 1993: Morgan kaufmann.

[26] Cohen, W.W., Learning Trees an ules with Set-val ed Features. 1996.

[27] Holte, R.C., Very simple classification rules perform well on most commonly used datasets. Machine learning, 1993. 11(1): p. 63-90.

[28] Kohavi, R., The power of decision tables. Machine Learning: ECML-95, 1995: p. 174-189.

[29] Frank, E. and I.H. Witten, Generating accurate rule sets without global optimization. 1998.

Seyed mahdi sadatrasoul is a Ph.D student in industrial engineering and systems management at Iran University of Science and Technology (IUST), Tehran. He received his Bs degree in Industrial Engineering from IUST, in 2006 and obtained M.S. degree in information technology management from Tarbiat modares university (TMU), Tehran, in 2009. Presently he is the assistant of faculty member at IT Group in School of Industrial Engineering and is actively engaged in conducting academic, research and development programs in the field of data and process mining. He has contributed more than 20 research papers to many national and international journals and conferences. He has also published two books by reputed publishers. His research interests are including data mining and its synergies with operation research, credit allocation and scoring, e- commerce and financial information systems (FIS).

Mohammad Reza Gholamian is an Assistant Professor in School of Industrial Engineering at the Iran University of Science and Technology (IUST), Tehran. He received his M.S. degree in Industrial Engineering from Isfahan University of Technology (IUT), Isfahan in 1998 and obtained Ph.D. degree in Industrial Engineering from Amirkabir University of Technology (AUT), Tehran in 2005 for the work in the field of Hybrid Intelligent Decision Making Systems. Presently he is a faculty member of IT Group in School of Industrial Engineering and is actively engaged in conducting academic, research and development programs in the field of Industrial Engineering and Information Technology. He has contributed more than 105 research papers to many national and international journals and conferences. Besides this, he has published four books by reputed publishers. His research interests include data mining, soft computing, and decision theory and e-business models.

Kamran Shahanaghi is an Assistant Professor in School of Industrial Engineering at the Iran University of Science and Technology (IUST), Tehran. He received his M.S. degree in Industrial Engineering from IUST in 1986 and obtained Ph.D. degree in 2000. Presently he is a faculty member of optimization Group in School of Industrial Engineering and is actively engaged in conducting academic, research and development programs in the field of Industrial Engineering and optimization. He has contributed more than 140 research papers to many national and international journals and conferences. His research interests include operation research and uncertainty.

Joint Relay Selection and Power Allocation in MIMO Cooperative Cognitive Radio Networks

Mehdi Ghamari Adian*
Department of Electrical Engineering, University of Zanjan, Zanjan, Iran
mehdi.ghamari@aut.ac.ir

Hassan Aghaeinia
Department of Electrical Engineering Amirkabir University of Technology, Tehran, Iran
aghaeini@aut.ac.ir

Received: 01/Sep/2013 Revised: 07/Aug/2014 Accepted: 11/ Sep/2014

Abstract
In this work, the issue of joint relay selection and power allocation in Underlay MIMO Cooperative Cognitive Radio Networks (U-MIMO-CCRN) is addressed. The system consists of a number of secondary users (SUs) in the secondary network and a primary user (PU) in the primary network. We consider the communications in the link between two selected SUs, referred to as the desired link which is enhanced using the cooperation of one of the existing SUs. The core aim of this work is to maximize the achievable data rate in the desired link, using the cooperation of one of the SUs which is chosen opportunistically out of existing SUs. Meanwhile, the interference due to the secondary transmission on the PU should not exceed the tolerable amount. The approach to determine the optimal power allocation, i.e. the optimal transmits covariance and amplification matrices of the SUs, and also the optimal cooperating SU is proposed. Since the proposed optimal approach is a highly complex method, a low complexity approach is further proposed and its performance is evaluated using simulations. The simulation results reveal that the performance loss due to the low complexity approach is only about 14%, while the complexity of the algorithm is greatly reduced.

Keywords: Cognitive Radio Networks; Cooperative Communications; MIMO Systems; Low Complexity Approach.

1. Introduction

Since the issuance of the report of Federal Communications Commission (FCC) in 2002, which revealed the spectrum inefficiency in the incumbent wireless communication systems, cognitive radio (CR) has been regarded as one potential technology to activate the utilization of spectrum resources in the recent evolution of wireless communication systems [1]. As a consequence, the overlay and underlay modes can be developed, based on the definitions of spectrum holes in [1] and the operation modes in [2, 3], to use the white and gray spectrum holes, respectively.

To further enhance the system performance, a cooperative relay network can be incorporated into secondary system (SS). Thus, in the underlay CR system with an IT limit, the cooperative relay networks can also be applied to have a better capacity and error rate performance [5], trade-off between achievable rate and network lifetime [6], maximum signal-to-interference-plus-noise ratio (SINR) at the destination node [7], better channel utilization by multi-hop relay [8], maximum throughput and reduced interference via beam forming [9], and maximum SINR using cooperative beam forming [10].

Multiple-input/multiple-output (MIMO) systems have a great potential to enhance the throughput in the framework of wireless networks [11, 12]. Using M transmits antennas at the transmitter and N receive antennas at the receiver, the capacity of a MIMO single user is equal to $m \quad \{M, N\}$ times the capacity of a single-input/single-output (SISO) system [11, 12]. Multiple antennas can be applied to achieve many desirable goals, such as capacity increase without bandwidth expansion, transmission reliability enhancement via space-time coding, and co-channel interference suppression for multi-user transmission.

The method on relay selection and channel allocation in [13] greedily searches the most profitable pair to maximize system throughput, without considering the interference with primary users, which is the case for CR networks. The problem of joint relay selection and power allocation to maximize system throughput with limited interference to licensed (primary) users in cognitive radio networks was investigated in [14]. In [15], the structure of an optimal relay precoder design for Amplify-and-Forward based Underlay MIMO cognitive relay was studied.

Joint problems of relay selection and resource allocation in CR networks (CRNs) have attracted extensive

* Corresponding Author

research interests due to its more effective spectrum utilization [13]-[18]. The authors in [13] consider a cooperative cognitive radio network (CCRN) in which the relays are selected among the existing SUs. For CCRNs with decode-and-forward strategy, two relay selection schemes, namely, full-channel state information (CSI)-based best relay selection (BRS) and a partial CSI-based best relay selection (PBRS) were proposed in [14]. In order to obtain an optimal subcarrier pairing, relay assignment and power allocation in MIMO-OFDM based CCRNs, the dual decomposition technique was recruited in [15] to maximize the sum rate subject to the interference temperature limit of the PUs. The issue of joint relay selection and power allocation in two-way CCRN was considered in [16]. A suboptimal approach for reducing the complexity of joint relay selection and power allocation in CCRN was proposed in [17]. The network coding opportunities was exploited in [18].

The issue of resource allocation in MIMO CRNs was explored in [19]-[22]. The authors in [19] presented a low complexity algorithm for resource allocation in MIMO-OFDM based CR networks, using game theory approach and the primal decomposition method. In [20], the authors extended the pricing concept to MIMO-OFDM based CR networks and presented two iterative algorithms for resource allocation in such systems. To obtain an optimal subcarrier pairing, relay assignment and power allocation in MIMO-OFDM based CCRNs; the dual decomposition technique was recruited in [21] to maximize the sum-rate subject to the interference temperature limit of the PUs. Moreover, due to high computational complexity of the optimal approach, a suboptimal algorithm was further proposed in [21] and [22].

In this paper, we consider the opportunistic spectrum access in MIMO cognitive radio networks (MIMO-CRN). More specifically, we propose a Cognitive Cooperative communication protocol based on Beam forming (CCB) in MIMO-CRN which ensures the SU's continuous transmission and reduces its outage probability without interfering the PUs. The desired link is considered as the MIMO link between two SUs, the SU TX and SU RX. Meanwhile, CCB adopts beam forming at the SU RX and the cooperating SU. As a result, the SU RX only receives signals from the SU TX and the best relay, and the interferences from the PUs are suppressed. The same story applies to the cooperating SU as a result of beam forming. To be more accurate, when a PU transmits signal in the system, the joint problems of opportunistic relay selection and power allocation in the context of MIMO CR networks to maximize the end-to-end achievable data rate of Underlay MIMO CR networks need to be considered. Our focus is on the amplify-and-forward (AF) relay strategy. An obvious reason is that AF has low complexity since no decoding/encoding is needed. This benefit is even more attractive in MIMO-CRN, where decoding multiple data streams could be computationally intensive. In addition to simplicity, a more important reason is that AF outperforms decode-and-forward (DF) in terms of network capacity scaling: in general, as the number of relays increases in MIMO-CRN, the effective signal-to-noise ratio (SNR) under AF scales linearly, as opposed to being a constant under DF [30].

The remainder of this paper is organized as follows. Section 2 presents the system model and general formulation of the problem. In Sections 3, the structure of optimal power allocation matrices is studied. Based on these structural results, we simplify and reformulate the optimization problem. The optimization algorithms, including the optimal and suboptimal approach are discussed in Section 4. In Section 5, the outage probability of the desired link is analyzed. Numerical results are provided in Section 6 to show the efficacy of the proposed algorithms and Section 7 concludes this paper.

Notation: The following notation is used throughout the paper. The operators $(.)^H$, $|.|$, $Tr(.)$ and $(.)^+$ are Hermitian (complex conjugate), determinant, trace and pseudo-inverse operators, respectively.

2. System Model

We consider a scenario where a CR network, consisting of $N_{SU}+2$ SUs, coexists with a primary network, consisting of N_{PU} PU pair. In this paper the communication between two SUs is considered, which is also referred to as the desired SU link. The SU transmitter (SU TX) transmits signals to SU receiver (SU RX) either in the direct link or taking advantage of the cooperation of one of the SUs, depending on the presence of the PUs in the system. When the PUs are absent, the SU TX simply communicates the SU RX directly. Therefore, throughout this and next sections, we assume that the PU pairs are present and, as discussed in the previous section, it is inevitable for the SU TX to take advantage of the cooperation of one the SUs to keep the imposed interference on the PUs in the allowed region.

2-1- The transmission process at the presence of PUs

When the PU pairs are present, the direct communications between the SU TX and SU RX may impose intolerable interference on the PUs. The cooperation of one of SUs with the desired SU link can provide the possibility of reducing the transmit power of the SUs and thereby less interference is imposed on the PU pairs. A transmission from SU TX to SU RX at the presence of PUs takes two time-slots. In the first time-slot, the SU TX transmits signals to all the existing SUs in the CR network and the SUs employ beamforming to only receive signal of the SU TX. In the second time-slot, one of the SUs is selected to cooperate with the SU TX by amplifying its received signal and forwarding it to the SU RX, without decoding the message. All the transmissions in the SU system need to be regulated in order to avoid excessive interference on the PU pair. Meanwhile, the interference from the PUs in the SU TX is avoided by employing beam forming. The set of candidate SUs to cooperate with the desired SU link is denoted by S_R . Besides, the set of PU pairs is also denoted by S_{PU} . It is

further assumed that all the users, including the SUs and the PUs are equipped with multiple-antennas. Without loss of generality and for ease of exposition, we assume that the entire candidate SUs to cooperate with desired link are equipped with N_r antennas and the PUs with N_p antennas. The number of antennas at SU TX and SU RX are also N_s and N_d, respectively. $\mathbf{H}_{sr,i} \in \mathbb{C}^{N_r \times N_s}$ represents the channel matrix form SU TX to SU *i* and $\mathbf{H}_{rd,i} \in \mathbb{C}^{N_r \times N_d}$ represents the channel form SU *i* to SU RX. All the channels are modeled as Rayleigh fading channels and invariant during one time slot. It is further assumed that all the instantaneous channel matrices are perfectly known at the SU TX. The assumption of perfect knowledge of all the channel gains is a typical assumption in this area [31, 32]. In the presence of PUs, the amplify-and-forward (AF) relaying protocol is used.

2-2- Problem Formulation

The received signal at *i*-th SU can be written as

$$\mathbf{y}_{r,i} = \mathbf{H}_{sr,i}\mathbf{x}_{s,i} + \mathbf{n}_{r,i}, \quad \forall i \in S_R \tag{1}$$

where the transmit signal of SU TX, intended for SU *i*, is denoted by $\mathbf{x}_{s,i} \in \mathbb{C}^{N_s \times 1}$. $\mathbf{n}_{r,i} \in \mathbb{C}^{N_r \times 1}$ is the additive white Gaussian noise at SU *i*. Note that in (1) the negative effect of the PU signal on received signal of the candidate SUs is canceled, due to employment of beam forming. Suppose that SU *i* is selected to cooperate with the desired SU link. Then, the received signal at SU RX from SU *i* is given by

$$\begin{aligned}\mathbf{y}_d &= \mathbf{H}_{rd,i}\mathbf{A}_i\mathbf{y}_{r,i} + \mathbf{n}_d \\ &= \mathbf{H}_{rd,i}\mathbf{A}_i\mathbf{H}_{sr,i}\mathbf{x}_{s,i} + \mathbf{H}_{rd,i}\mathbf{A}_i\mathbf{n}_{r,i} + \mathbf{n}_d\end{aligned} \tag{2}$$

where $\mathbf{A}_i$ represents the amplification matrix, used at SU *i*; $\mathbf{n}_d \in \mathbb{C}^{N_d \times 1}$ is the additive white Gaussian noise at SU RX. Once again, it is presumed that the interference from the PUs is eliminated at the SU RX, by recruiting the appropriate beam forming. As a result of cooperation of one of the SUs, SU *i*, the achievable data rate in the desired link can be written as

$$\begin{aligned}R_i = \frac{1}{2}\log_2 \Big| &\mathbf{I}_{N_d} + \mathbf{H}_{rd,i}\mathbf{A}_i\mathbf{H}_{sr,i}\mathbf{Q}_i\mathbf{H}_{sr,i}^H\mathbf{A}_i^H\mathbf{H}_{rd,i}^H \\ &\times\left(\sigma_d^2\mathbf{I}_{N_d} + \sigma_r^2\mathbf{H}_{rd,i}\mathbf{A}_i\mathbf{A}_i^H\mathbf{H}_{rd,i}^H\right)^{-1}\Big|\end{aligned} \tag{3}$$

where σ_r^2 and σ_d^2 denote the variances of $\mathbf{n}_{r,i}$ and $\mathbf{n}_d$, and $\mathbf{Q}_i$ denotes the transmit covariance matrix of SU TX, intended for SU *i*. The transmit power of SU TX is restricted to P_T, i.e. $Tr(\mathbf{Q}_i) \le P_T$.

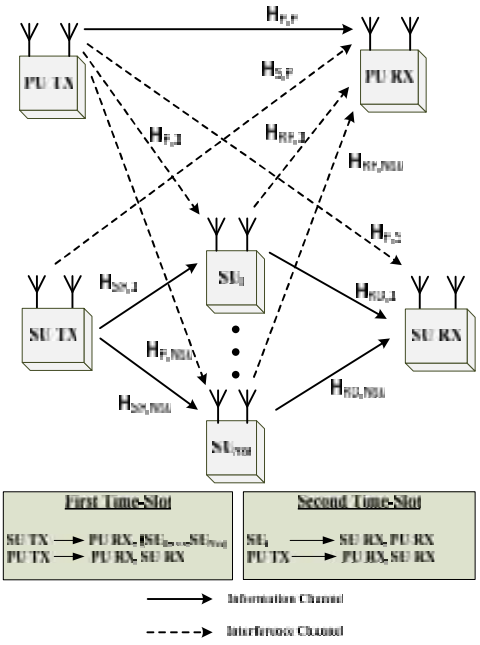


Fig. 1 System Model

Furthermore, the maximum transmit power of the SU *i*, if selected as the cooperative relay, is P_R. The PUs must not be disturbed as a result of transmission by SU TX and further the cooperation of the selected SU with the SU TX. In this way, the interference power constraints on the PUs are provided by $Tr\left(\mathbf{H}_{s,p,n}\mathbf{Q}_i\mathbf{H}_{s,p,n}^H\right) \le P_{I,1}$ and $Tr\left\{\mathbf{H}_{i,p,n}\left[\mathbf{A}_i\left(\sigma_r^2\mathbf{I}_{N_r} + \mathbf{H}_{sr,i}\mathbf{Q}_i\mathbf{H}_{sr,i}^H\right)\mathbf{A}_i^H\right]\mathbf{H}_{i,p,n}^H\right\} \le P_{I,2}$, for all $n \in S_{PU}$, where the SU *i* is selected to cooperate with the SU TX. Moreover, $\mathbf{H}_{i,p,n}$ and $\mathbf{H}_{s,p,n}$ represent the channel from SU *i* and SU TX to *n*-th PU RX, respectively. Evidently, the maximum tolerable interference at the PUs is $P_{I,1} + P_{I,2}$. One of the aims of this work is to optimally select the cooperating SU and also calculate the optimum power allocation in the proposed system, which can be formulated as:

$$\begin{aligned}&\left(\mathbf{Q}_i^*, \mathbf{A}_i^*\right) = \arg\max_{\mathbf{Q},\mathbf{A}_i} \; R_i\left(\mathbf{Q}_i, \mathbf{A}_i\right) \\ &\quad i = \arg\max_i \; R_i\left(\mathbf{Q}_i^*, \mathbf{A}_i^*\right) \\ &\text{s.t. } Tr\left(\mathbf{Q}_i\right) \le P_T \\ &\quad Tr\left[\mathbf{A}_i\left(\sigma_r^2\mathbf{I}_{N_r} + \mathbf{H}_{sr,i}\mathbf{Q}_i\mathbf{H}_{sr,i}^H\right)\mathbf{A}_i^H\right] \le P_R \\ &\quad Tr\left(\mathbf{H}_{s,p,n}\mathbf{Q}_i\mathbf{H}_{s,p,n}^H\right) \le P_{I,1}, \quad \forall n \in S_{PU} \\ &\quad Tr\left\{\mathbf{H}_{i,p,n}\left[\mathbf{A}_i\left(\sigma_r^2\mathbf{I}_{N_r} + \mathbf{H}_{sr,i}\mathbf{Q}_i\mathbf{H}_{sr,i}^H\right)\mathbf{A}_i^H\right]\mathbf{H}_{i,p,n}^H\right\} \le P_{I,2}, \forall n \in S_{PU} \\ &\quad \mathbf{A}_i \ge 0, \mathbf{Q}_i \ge 0\end{aligned} \tag{4}$$

where $\mathbf{Q}_i^*$ and $\mathbf{A}_i^*$ are the optimum transmit covariance and amplification matrices. For convenience, we define two constraint sets according to the following:

$$\Phi_i \sqcup \left\{ \mathbf{Q}_i \,|\, Tr\left(\mathbf{Q}_i\right) \le P_T, \; Tr\left(\mathbf{H}_{s,p,n}\mathbf{Q}_i\mathbf{H}_{s,p,n}^H\right) \le P_I, \mathbf{Q}_i \succeq 0, \forall n \in S_{PU} \right\} \tag{5}$$

$$\Psi_i \sqcup \left\{ \mathbf{A}_i \left| \begin{array}{l} Tr\left[\mathbf{A}_i\left(\pm_r^2\mathbf{I}_{N_r} + \mathbf{H}_{sr,i}\mathbf{Q}_i\mathbf{H}_{sr,i}^H\right)\mathbf{A}_i^H\right] \le P_R, \mathbf{A}_i \succeq 0 \\ Tr\left\{\mathbf{H}_{i,p,n}\left[\mathbf{A}_i\left(\pm_r^2\mathbf{I}_{N_r} + \mathbf{H}_{sr,i}\mathbf{Q}_i\mathbf{H}_{sr,i}^H\right)\mathbf{A}_i^H\right]\mathbf{H}_{i,p,n}^H\right\} \le P_i \end{array} \right. \right\} \tag{6}$$

It is easy to verify that (4) can be *decomposed* into three parts as follows:

$$\max_{i \in S_R}\left(\max_{\mathbf{Q}_i \in \Phi_i}\left(\max_{\mathbf{A}_i \in \Psi_i} R_i\left(\mathbf{Q}_i, \mathbf{A}_i\right)\right)\right) \tag{7}$$

Hence, solving (4) reduces to *iteratively* solving a sub-problem with respect to A_i **,** for all $i \in S_R$ and $n \in S_{PU}$ (with $\mathbf{Q}_i$ fixed), then another sub-problem with respect to $\mathbf{Q}_i$ (with A_i fixed, $\forall i \in S_R$ and $n \in S_{PU}$) and finally a main problem with respect to i . Directly tackling problem (4) is intractable in general. However, we will exploit the inherent special structure to significantly reduce the problem complexity and convert it to an equivalent problem with scalar parameters. In what follows, we will first study the optimal structural properties of A_i and $\mathbf{Q}_i$. Based on these properties, we will reformulate (4).

3. Optimal Power Allocation in the SU TX and Cooperating SU

In the first subsection, the structure of the optimal amplification matrix in *i*-th SU for a given $\mathbf{Q}_i$ is investigated. Then, the optimal structure of $\mathbf{Q}_i$ is studied in second subsection. Finally, based on these optimal structures, the problem in (4) is reformulated in third subsection.

3-1- The Structure of the optimal amplification matrices

For now, we assume that $\mathbf{Q}_i$ is given. Let the eigenvalue-decomposition of $\mathbf{H}_{sr,i}\mathbf{H}_{sr,i}^H$ and $\mathbf{H}_{rd,i}^H\mathbf{H}_{rd,i}$ be

$$\mathbf{H}_{sr,i}\mathbf{H}_{sr,i}^H = \mathbf{U}_{sr,i}\Sigma_{sr,i}\mathbf{U}_{sr,i}^H, \;\; \mathbf{H}_{rd,i}^H\mathbf{H}_{rd,i} = \mathbf{V}_{rd,i}\Sigma_{rd,i}\mathbf{V}_{rd,i}^H \tag{8}$$

where $\mathbf{U}_{sr,i}$ and $\mathbf{V}_{rd,i}$ are unitary matrices, $\Sigma_{sr,i} = diag\left\{\sim_1, \sim_2, \ldots, \sim_{N_r}\right\}$ with $\sim_l \ge 0$, and $\Sigma_{rd,i} = diag\left\{\ _1, \ _2, \ldots, \ _{N_r}\right\}$ with $\ _l \ge 0$.

Proposition 1: *The optimal amplification matrix of SU i,* $\mathbf{A}_i$ *, has the following structure*

$$\mathbf{A}_{i,opt} = \mathbf{V}_{rd,i}\Lambda_{\mathbf{A}_i}\tilde{\mathbf{U}}_{sr,i}^H \tag{9}$$

Where $\tilde{\mathbf{U}}_{sr,i}$ *is obtained by eigenvalue decomposition of* $\tilde{\mathbf{H}}_{sr,i}\tilde{\mathbf{H}}_{sr,i}^H$ *and* $\tilde{\mathbf{H}}_{sr,i} = \mathbf{H}_{sr,i}\mathbf{Q}_i^{\frac{-1}{2}}$ *, i.e.* $\tilde{\mathbf{H}}_{sr,i}\tilde{\mathbf{H}}_{sr,i}^H = \mathbf{H}_{sr,i}\tilde{\mathbf{Q}}_i\mathbf{H}_{sr,i}^H = \tilde{\mathbf{U}}_{sr,i}\tilde{\Sigma}_{sr,i}\tilde{\mathbf{U}}_{sr,i}^H$.

Proof. Please refer to appendix A.

Let the singular value decomposition (SVD) of $\mathbf{H}_{sr,i}$ and $\mathbf{H}_{rd,i}$ be

$$\mathbf{H}_{sr,i} = \mathbf{U}_{sr,i}\Lambda_{sr,i}\mathbf{V}_{sr,i}^H, \;\; \mathbf{H}_{rd,i} = \mathbf{U}_{rd,i}\Lambda_{rd,i}\mathbf{V}_{rd,i}^H \tag{10}$$

which satisfies (8). Then exploiting (9), (10) and (3), the achievable data rates of the desired link can be written as

$$\begin{aligned} & R_i\left(\mathbf{Q}_i, \mathbf{A}_{i,opt}\right) = \\ & \frac{1}{2}\log_2\left|\mathbf{I}_{N_d} + \mathbf{H}_{rd,i}\mathbf{A}_{i,opt}\mathbf{H}_{sr,i}\mathbf{Q}_i\mathbf{H}_{sr,i}^H\mathbf{A}_{i,opt}^H\mathbf{H}_{rd,i}^H\right. \\ & \left.\times\left(\pm_d^2\mathbf{I}_{N_d} + \pm_r^2\mathbf{H}_{rd,i}\mathbf{A}_{i,opt}\mathbf{A}_{i,opt}^H\mathbf{H}_{rd,i}^H\right)^{-1}\right| \\ & = \frac{1}{2}\log_2\left|\mathbf{I}_{N_d} + \Lambda_{rd,i}^2\Lambda_{\mathbf{A}_i}^2\tilde{\Sigma}_{sr,i}\left(\pm_d^2\mathbf{I}_{N_d} + \pm_r^2\Lambda_{rd,i}^2\Lambda_{\mathbf{A}_i}^2\right)^{-1}\right| \end{aligned} \tag{11}$$

According to (11), the achievable data rate in the desired SU link only depends on $\tilde{\Sigma}_{sr,i}$ but not on $\tilde{\mathbf{U}}_{sr,i}$. Then, it can be concluded that for any matrix $\hat{\mathbf{Q}}_i$ which satisfies $\mathbf{H}_{sr,i}\hat{\mathbf{Q}}_i\mathbf{H}_{sr,i}^H = \hat{\mathbf{U}}_{sr,i}\tilde{\Sigma}_{sr,i}\hat{\mathbf{U}}_{sr,i}^H$, the optimal data rate is the same as when the transmit covariance matrix in the desired link is $\tilde{\mathbf{Q}}_i$. Therefore (9) can be written as

$$\mathbf{A}_{i,opt} = \mathbf{V}_{rd,i}\Lambda_{\mathbf{A}_i}\mathbf{U}_{sr,i}^H \tag{12}$$

3-2- The Structure of the optimal transmit covariance Matrix

In this subsection, the optimal structure of the transmit covariance matrix of the desired link is determined.

Proposition 2: *The structure of optimal transmits covariance matrix of SU TX is as follow:*

$$\mathbf{Q}_i = \mathbf{V}_{sr,i}\Lambda_{\mathbf{Q}_i}\mathbf{V}_{sr,i}^H \tag{13}$$

where $\Lambda_{\mathbf{Q}_i}$ *is a diagonal matrix and must be determined such that the achievable data rate in the desired link is maximized.*

Proof. Suppose that $\tilde{\Sigma}_{sr,i,1}$ is $r \times r$, then

$$\begin{aligned} & \mathbf{H}_{sr,i}\hat{\mathbf{Q}}_i\mathbf{H}_{sr,i}^H = \hat{\mathbf{U}}_{sr,i}\tilde{\Sigma}_{sr,i}\hat{\mathbf{U}}_{sr,i}^H \\ & = \left[\hat{\mathbf{U}}_{sr,i,1} \quad \hat{\mathbf{U}}_{sr,i,2}\right]\begin{bmatrix}\tilde{\Sigma}_{sr,i} & \\ & \mathbf{0}\end{bmatrix}\left[\hat{\mathbf{U}}_{sr,i,1} \quad \hat{\mathbf{U}}_{sr,i,2}\right]^H \end{aligned} \tag{14}$$

where $\hat{\mathbf{Q}}_i$ is any PSD [1] matrix which satisfies $\mathbf{H}_{sr,i}\hat{\mathbf{Q}}_i\mathbf{H}_{sr,i}^H = \hat{\mathbf{U}}_{sr,i}\tilde{\Sigma}_{sr,i}\hat{\mathbf{U}}_{sr,i}^H$. Hence the singular value decomposition of matrix $\mathbf{H}_{sr,i}$ with rank r can be expressed as

1. Positive Semi-Definite

$$\begin{aligned}\mathbf{H}_{sr,i} &= \mathbf{U}_{sr,i}\Lambda_{sr,i}\mathbf{V}_{sr,i}^{H} \\ &= \begin{bmatrix}\mathbf{U}_{sr,i,1} & \mathbf{U}_{sr,i,2}\end{bmatrix}\begin{bmatrix}\Lambda_{sr,i,1} & \\ & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_{sr,i,1} & \mathbf{V}_{sr,i,2}\end{bmatrix}^{H}\end{aligned} \quad (15)$$

where $\Lambda_{sr,i,1}$ is $r\times r$. It can be shown that $\mathbf{U}_{sr,i,1}$ is orthogonal to $\hat{\mathbf{U}}_{sr,i,2}$. Moreover, $\mathbf{U}_{sr,i,2}$ is orthogonal to $\hat{\mathbf{U}}_{sr,i,1}$. The pseudo-inverse of $\mathbf{H}_{sr,i}$ is denoted by $\mathbf{H}_{sr,i}^{+}$. Then from

$$\mathbf{H}_{sr,i}\hat{\mathbf{Q}}_i\mathbf{H}_{sr,i}^{H} = \hat{\mathbf{U}}_{sr,i}\tilde{\Sigma}_{sr,i}\hat{\mathbf{U}}_{sr,i}^{H}$$

we have

$$\begin{aligned}&\mathbf{H}_{sr,i}^{+}\mathbf{H}_{sr,i}\hat{\mathbf{Q}}_i\mathbf{H}_{sr,i}^{H}\mathbf{H}_{sr,i}^{+H} = \\ &\begin{bmatrix}\mathbf{V}_{sr,i,1} & \mathbf{V}_{sr,i,2}\end{bmatrix}\begin{bmatrix}\Lambda_{sr,i,1}^{-1} & \\ & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{U}_{sr,i,1} & \mathbf{U}_{sr,i,2}\end{bmatrix}^{H} \\ &\times\begin{bmatrix}\hat{\mathbf{U}}_{sr,i,1} & \hat{\mathbf{U}}_{sr,i,2}\end{bmatrix}\begin{bmatrix}\tilde{\Sigma}_{sr,i,1} & \\ & \mathbf{0}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{U}}_{sr,i,1} & \hat{\mathbf{U}}_{sr,i,2}\end{bmatrix}^{H} \\ &\times\begin{bmatrix}\mathbf{U}_{sr,i,1} & \mathbf{U}_{sr,i,2}\end{bmatrix}\begin{bmatrix}\Lambda_{sr,i,1}^{-1} & \\ & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_{sr,i,1} & \mathbf{V}_{sr,i,2}\end{bmatrix}^{H} \\ &= \begin{bmatrix}\mathbf{V}_{sr,i,1} & \mathbf{V}_{sr,i,2}\end{bmatrix} \\ &\times\begin{bmatrix}\Lambda_{sr,i,1}^{-1}\mathbf{U}_{sr,i,1}^{H}\hat{\mathbf{U}}_{sr,i,1}\hat{\mathbf{U}}_{sr,i,2}\tilde{\Sigma}_{sr,i,1}\hat{\mathbf{U}}_{sr,i,1}\mathbf{U}_{sr,i,1}\Lambda_{sr,i,1}^{-1} & \\ & \mathbf{0}\end{bmatrix} \\ &\times\begin{bmatrix}\mathbf{V}_{sr,i,1} & \mathbf{V}_{sr,i,2}\end{bmatrix}^{H}\end{aligned} \quad (16)$$

It can be verified that $\mathbf{U}_{sr,i,1}^{H}\hat{\mathbf{U}}_{sr,i,1}$ is a unitary matrix, because $\mathbf{U}_{sr,i}^{H}\hat{\mathbf{U}}_{sr,i}$ is unitary. Recall that if $\mathbf{A}$ and $\mathbf{B}$ are two positive semi-definite $M\times M$ matrices with eigenvalues $\mathsf{a}_i(\mathbf{A})$ and $\mathsf{a}_i(\mathbf{B})$, arranged in the descending order respectively, then

$$\sum_{i=1}^{M}\mathsf{a}_i(\mathbf{A})\mathsf{a}_{M+1-i}(\mathbf{B}) \le Tr(\mathbf{AB}) \le \sum_{i=1}^{M}\mathsf{a}_i(\mathbf{A})\mathsf{a}_i(\mathbf{B}) \quad (17)$$

Then using the second inequality in (17) and knowing that $\mathbf{H}_{sr,i}^{H}\mathbf{H}_{sr,i}^{+H}\mathbf{H}_{sr,i}^{+}\mathbf{H}_{sr,i}$ is a project matrix with eigenvalues being only 1 and 0, we have

$$\begin{aligned}Tr(\mathbf{Q}_i) &\ge Tr\left(\mathbf{H}_{sr,i}^{+}\mathbf{H}_{sr,i}\mathbf{Q}_i\mathbf{H}_{sr,i}^{H}\mathbf{H}_{sr,i}^{+H}\right) \\ &= Tr\left(\Lambda_{sr,i,1}^{-1}\mathbf{U}_{sr,i,1}^{H}\hat{\mathbf{U}}_{sr,i,2}\tilde{\Sigma}_{sr,i,1}\hat{\mathbf{U}}_{sr,i,1}\mathbf{U}_{sr,i,1}\Lambda_{sr,i,1}^{-1}\right)\end{aligned} \quad (18)$$

Using the first equality in (17) we can conclude that

$$Tr(\mathbf{Q}_i) \ge Tr\left(\tilde{\Sigma}_{sr,i,1}\Lambda_{sr,i,1}^{-2}\right) \quad (19)$$

Therefore, the structure of the optimal transmit covariance matrix in the desired link is given by

$$\mathbf{Q}_{opt,i} = \begin{bmatrix}\mathbf{V}_{sr,i,1} & \mathbf{V}_{sr,i,2}\end{bmatrix}\begin{bmatrix}\tilde{\Sigma}_{sr,i,1}\Lambda_{sr,i,1}^{-2} & \\ & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_{sr,i,1} & \mathbf{V}_{sr,i,2}\end{bmatrix}^{H} \quad (20)$$

which satisfies

$$\mathbf{H}_{sr,i}\mathbf{Q}_{opt,i}\mathbf{H}_{sr,i}^{H} = \mathbf{U}_{sr,i}\begin{bmatrix}\tilde{\Sigma}_{sr,i,1} & \\ & \mathbf{0}\end{bmatrix}\mathbf{U}_{sr,i}^{H} \quad (21)$$

and the proposition is proved.

3-3- Problem reformulation

In the previous section we proved that the structure of the optimal amplification matrix in SU *i* and transmit covariance matrix in the SU TX can be expressed as

$$\mathbf{A}_{i,opt} = \mathbf{V}_{rd,i}\Lambda_{\mathbf{A}_i}\mathbf{U}_{sr,i}^{H}, \quad \mathbf{Q}_i = \mathbf{V}_{sr,i}\Lambda_{\mathbf{Q}_i}\mathbf{V}_{sr,i}^{H} \quad (22)$$

where $\mathbf{H}_{sr,i} = \mathbf{U}_{sr,i}\Lambda_{sr,i}\mathbf{V}_{sr,i}^{H}$ and $\mathbf{H}_{rd,i} = \mathbf{U}_{rd,i}\Lambda_{rd,i}\mathbf{V}_{rd,i}^{H}$. Recall that the received signal in the SU RX, due to the cooperation of SU *i*, is given by

$$\mathbf{y}_d = \mathbf{H}_{rd,i}\mathbf{A}_i\mathbf{H}_{sr,i}\mathbf{x}_{s,i} + \mathbf{H}_{rd,i}\mathbf{A}_i\mathbf{n}_{r,i} + \mathbf{n}_d \quad (23)$$

Using (22), $\mathbf{y}_d$ in (23) can be rewritten as

$$\mathbf{y}_d = \mathbf{U}_{rd,i}\Lambda_{rd,i}\Lambda_{\mathbf{A}_i}\Lambda_{sr,i}\mathbf{V}_{sr,i}^{H}\mathbf{x}_{s,i} + \mathbf{U}_{rd,i}\Lambda_{rd,i}\Lambda_{\mathbf{A}_i}\mathbf{U}_{sr,i}^{H}\mathbf{n}_{r,i} + \mathbf{n}_d \quad (24)$$

Suppose that $\tilde{\mathbf{y}}_d = \mathbf{U}_{rd,i}^{H}\mathbf{y}_d$, $\tilde{\mathbf{x}}_{s,i} = \mathbf{V}_{sr,i}^{H}\mathbf{x}_{s,i}$, $\tilde{\mathbf{n}}_{r,i} = \mathbf{U}_{sr,i}^{H}\mathbf{n}_{r,i}$ and $\tilde{\mathbf{n}}_d = \mathbf{U}_{rd,i}^{H}\mathbf{n}_d$. Then,

$$\tilde{\mathbf{y}}_d = \Lambda_{rd,i}\Lambda_{\mathbf{A}_i}\Lambda_{sr,i}\tilde{\mathbf{x}}_{s,i} + \Lambda_{rd,i}\Lambda_{\mathbf{A}_i}\tilde{\mathbf{n}}_{r,i} + \tilde{\mathbf{n}}_d \quad (25)$$

Clearly, the relay channel between the SU TX and SU RX has been decomposed into a set of parallel SISO sub channels. Therefore, the achievable data rate in the desired link as result of the cooperation of SU *i* can be expressed as

$$R_i = \log_2\left|\mathbf{I}_{N_d} + \Lambda_{rd,i}^2\Lambda_{\mathbf{A}_i}^2\Lambda_{sr,i}^2\Lambda_{\mathbf{Q}_i}\left(\pm_r^2\Lambda_{rd,i}^2\Lambda_{\mathbf{A}_i}^2 + \pm_d^2\mathbf{I}_{N_d}\right)^{-1}\right| \quad (26)$$

Suppose that the eigenvalue decomposition of $\mathbf{H}_{s,p,n}^{H}\mathbf{H}_{s,p,n}$ and $\mathbf{H}_{i,p,n}^{H}\mathbf{H}_{i,p,n}$ is

$$\begin{aligned}\mathbf{H}_{s,p,n}^{H}\mathbf{H}_{s,p,n} &= \mathbf{U}_{s,p,n}\Lambda_{s,p,n}\mathbf{U}_{s,p,n}^{H} \\ \mathbf{H}_{i,p,n}^{H}\mathbf{H}_{i,p,n} &= \mathbf{U}_{i,p,n}\Lambda_{i,p,n}\mathbf{U}_{i,p,n}^{H}\end{aligned} \quad (27)$$

For all $n\in S_{PU}$. We further assume that

$$\begin{aligned}&\Lambda_{\mathbf{A}_i}^2 = \mathrm{diag}\left\{a_{1,i},\ldots,a_{N_r,i}\right\},\ \Lambda_{sr,i}^2 = \mathrm{diag}\left\{b_{1,i},\ldots,b_{N_r,i}\right\} \\ &\Lambda_{rd,i}^2 = \mathrm{diag}\left\{c_{1,i},\ldots,c_{N_r,i}\right\},\ \Lambda_{s,p,n} = \mathrm{diag}\left\{d_{1,n},\ldots,d_{N_s,n}\right\} \\ &\Lambda_{i,p,n} = \mathrm{diag}\left\{e_{1,i,n},\ldots,e_{N_r,i,n}\right\},\ \Lambda_{\mathbf{Q},i} = \mathrm{diag}\left\{q_{1,i},\ldots,q_{N_s,i}\right\}\end{aligned} \quad (28)$$

Then, using (28), (26) can be rewritten as

$$R_i = \sum_{k=1}^{N_r}\log_2\left(1+\frac{a_{k,i}b_{k,i}c_{k,i}q_{k,i}}{\pm_r^2 a_{k,i}c_{k,i} + \pm_d^2}\right) \quad (29)$$

Moreover, the transmit power constraint of the SU TX and SU *i* will become

$$Tr(\mathbf{Q}_i) \le P_T \Rightarrow \sum_{k=1}^{N_s} q_{k,i} \le P_T \quad (30)$$

$$\begin{aligned}&Tr\left[\mathbf{A}_i\left(\pm_r^2\mathbf{I}_{N_r} + \mathbf{H}_{sr,i}\mathbf{Q}_i\mathbf{H}_{sr,i}^{H}\right)\mathbf{A}_i^{H}\right] \\ &= Tr\left(\pm_r^2\Lambda_{\mathbf{A}_i}^2 + \Lambda_{sr,i}^2\Lambda_{\mathbf{Q}_i}\Lambda_{\mathbf{A}_i}^2\right) \le P_R \Rightarrow \\ &\sum_{k=1}^{N_r} a_{k,i}\left(\pm_r^2 + b_{k,i}q_{k,i}\right) \le P_R\end{aligned} \quad (31)$$

The interference constraint on PUs, due to transmission by the SU TX, can be written as

$$Tr\left(\mathbf{H}_{s,p,n}\mathbf{Q}_i\mathbf{H}_{s,p,n}^{H}\right) = Tr\left(\mathbf{V}_{sr,i}^{H}\mathbf{U}_{s,p,n}\Lambda_{s,p,n}\mathbf{U}_{s,p,n}^{H}\mathbf{V}_{sr,i}\Lambda_{\mathbf{Q}_i}\right) \quad (32)$$

for all $n \in S_{PU}$. Let $\mathbf{M}_{i,n} = \mathbf{V}_{sr,i}^H \mathbf{U}_{s,p,n}$. Then, (32) can be expressed as

$$Tr\left(\mathbf{M}_{i,n}\Lambda_{s,p,n}\mathbf{M}_{i,n}^H\Lambda_{\mathbf{Q}_i}\right) = \sum_{k=1}^{N_s}\left(\sum_{l=1}^{N_s}\left|m_{i,k,l,n}\right|^2 d_{l,n}\right)q_{k,i} \quad (33)$$

where $m_{i,k,l,n}$ denotes the element of *k*-th row and *l*-th column of matrix $\mathbf{M}_{i,n}$. Let $f_{k,i,n} = \sum_{l=1}^{N_s}\left|m_{i,k,l,n}\right|^2 d_{l,n}$. Therefore, the interference constraint on PUs, due to transmitting by SU TX, is expressed by

$$Tr\left(\mathbf{H}_{s,p.n}\mathbf{Q}_i\mathbf{H}_{s,p,n}^H\right) = \sum_{k=1}^{N_s} f_{k,i,n}q_{k,i} \le P_{I,1} \quad (34)$$

The interference constraint on PUs, due to the cooperation of SU *i* with SU TX, is written by

$$\begin{aligned} &Tr\left\{\mathbf{H}_{i,p,n}\mathbf{A}_i\left(\pm_r^2\mathbf{I}_{N_r} + \mathbf{H}_{sr,i}\mathbf{Q}_i\mathbf{H}_{sr,i}^H\right)\mathbf{A}_i^H\mathbf{H}_{i,p,n}^H\right\} \\ &= Tr\left(\mathbf{V}_{rd,i}^H\mathbf{U}_{i,p,n}\Lambda_{i,p,n}\mathbf{U}_{i,p,n}^H\mathbf{V}_{rd,i}\Lambda_{\mathbf{A}_i}\right. \\ &\left.\times\left(\pm_r^2\mathbf{I}_{N_r} + \Lambda_{sr,i}\Lambda_{\mathbf{Q}_i}\Lambda_{sr,i}\right)\mathbf{U}_{sr,i}\Lambda_{\mathbf{A}_i}\right) \le P_{I,2} \end{aligned} \quad (35)$$

for all $i \in S_R$ and $n \in S_{PU}$. Let $\mathbf{S}_{i,n} = \mathbf{V}_{rd,i}^H\mathbf{U}_{i,p,n}$. The element of *k*-th row and *l*-th column of $\mathbf{S}_{i,n}$ is denoted by $s_{i,k,l,n}$. Hence, it can be shown that (35) can be rewritten as

$$\sum_{k=1}^{N_r}\left(\pm_r^2 + b_{k,i}q_{k,i}\right)a_{k,i}\left(\sum_{l=1}^{N_r}\left|s_{i,k,l,n}\right|^2 e_{l,i,n}\right) \le P_{I,2} \quad (36)$$

Let $g_{k,i,n} = \sum_{l=1}^{N_r}\left|s_{i,k,l,n}\right|^2 e_{l,i,n}$. Thus, the interference constraint on PUs, due to the cooperation of the selected SU in relaying the signals of the SU TX is stated as

$$\sum_{k=1}^{N_r}\left(\pm_r^2 + b_{k,i}q_{k,i}\right)a_{k,i}g_{k,i,n} \le P_{I,2} \quad (37)$$

Let $\mathbf{a}_i = \left[a_{1,i},\ldots,a_{N_r,i}\right]$ and $\mathbf{q}_i = \left[q_{1,i},\ldots,q_{N_s,i}\right]$

Finally, the problem (4) can be expressed according to the following

$$\begin{aligned} &\mathbf{q}_i^*, \mathbf{a}_i^* = \arg\max_{\mathbf{q}_i,\mathbf{a}_i} \sum_{k=1}^{N_r}\log_2\left(1+\frac{a_{k,i}b_{k,i}c_{k,i}q_{k,i}}{\pm_r^2 a_{k,i}c_{k,i} + \pm_d^2}\right) \\ &i = \arg\max_i \sum_{k=1}^{N_r}\log_2\left(1+\frac{a_{k,i}^*b_{k,i}c_{k,i}q_{k,i}^*}{\pm_r^2 a_{k,i}^*c_{k,i} + \pm_d^2}\right) \\ &\text{s. t.} \quad \sum_{k=1}^{N_s} q_{k,i} \le P_T \\ &\sum_{k=1}^{N_r} a_{k,i}\left(\pm_r^2 + b_{k,i}q_{k,i}\right) \le P_R \\ &\sum_{k=1}^{N_s} f_{k,i,n}q_{k,i} \le P_{I,1}, \quad \forall n \in S_{PU} \\ &\sum_{k=1}^{N_r} g_{k,i,n}a_{k,i}\left(\pm_r^2 + b_{k,i}q_{k,i}\right) \le P_{I,2}, \quad \forall n \in S_{PU} \end{aligned} \quad (38)$$

Let $h_{k,i} = a_{k,i}\left(\pm_r^2 + b_{k,i}q_{k,i}\right)$. By some simple derivations, the problem in (38) is equivalent to

$$\begin{aligned} &\mathbf{q}_i^*, \mathbf{h}_i^* = \arg\max_{\mathbf{q}_i,\mathbf{h}_i} \sum_{k=1}^{N_r}\log_2\frac{\left(1+\frac{c_{k,i}h_{k,i}}{\pm_d^2}\right)\left(1+\frac{b_{k,i}q_{k,i}}{\pm_r^2}\right)}{1+\frac{c_{k,i}h_{k,i}}{\pm_d^2}+\frac{b_{k,i}q_{k,i}}{\pm_r^2}} \\ &i = \arg\max_i \sum_{k=1}^{N_r}\log_2\frac{\left(1+\frac{c_{k,i}h_{k,i}^*}{\pm_d^2}\right)\left(1+\frac{b_{k,i}q_{k,i}^*}{\pm_r^2}\right)}{1+\frac{c_{k,i}h_{k,i}^*}{\pm_d^2}+\frac{b_{k,i}q_{k,i}^*}{\pm_r^2}} \\ &\text{s. t.} \quad \sum_{k=1}^{N_s} q_{k,i} \le P_T \\ &\sum_{k=1}^{N_r} h_{k,i} \le P_R \\ &\sum_{k=1}^{N_s} f_{k,i,n}q_{k,i} \le P_{I,1}, \quad \forall n \in S_{PU} \\ &\sum_{k=1}^{N_r} g_{k,i,n}h_{k,i} \le P_{I,2}, \quad \forall n \in S_{PU} \end{aligned} \quad (39)$$

4. Optimization Algorithm

In this section, we develop approaches for joint relay selection and power allocation in cooperative cognitive radio networks. At first, we provide an optimal approach and then develop a low-complexity suboptimal approach.

4-1- Optimal approach

Using the Lagrange multipliers method [26] the Lagrange function for (39) is given by

$$\begin{aligned} &L\left(\mathbf{h}_i, \mathbf{q}_i, \mathtt{a}_1, \mathtt{a}_2, \mathtt{a}_{3,n}, \mathtt{a}_{4,n}\right) = \\ &-\sum_{k=1}^{N_r}\log_2\frac{\left(1+\frac{c_{k,i}h_{k,i}}{\pm_d^2}\right)\left(1+\frac{b_{k,i}q_{k,i}}{\pm_r^2}\right)}{1+\frac{c_{k,i}h_{k,i}}{\pm_d^2}+\frac{b_{k,i}q_{k,i}}{\pm_r^2}} \\ &+\mathtt{a}_1\left(\sum_{k=1}^{N_s} q_{k,i} - P_T\right) + \mathtt{a}_2\left(\sum_{k=1}^{N_r} h_{k,i} - P_R\right) + \\ &\sum_{n=1}^{N_{PU}} \mathtt{a}_{3,n}\left(\sum_{k=1}^{N_r} f_{k,i,n}q_{k,i} - P_{I,1}\right) \\ &+\sum_{n=1}^{N_{PU}} \mathtt{a}_{4,n}\left(\sum_{k=1}^{N_r} g_{k,i,n}h_{k,i} - P_{I,2}\right) \end{aligned} \quad (40)$$

where $\mathtt{a}_1$, $\mathtt{a}_2$, $\mathtt{a}_{3,n}$ and $\mathtt{a}_{4,n}$ are the Lagrange multipliers, $\forall n \in S_{PU}$. According to the KKT conditions, we have

$$
\begin{aligned}
&\alpha_1 \ge 0,\ \alpha_2 \ge 0,\ \alpha_{3,n} \ge 0,\ \alpha_{4,n} \ge 0,\\
&q_{l,i} \ge 0,\ h_{k,i} \ge 0,\ l = 1,\ldots,N_s,\ k = 1,\ldots,N_r\\
&\alpha_1\left(\sum_{k=1}^{N_s} q_{k,i} - P_T\right) = 0\\
&\alpha_2\left(\sum_{k=1}^{N_s} h_{k,i} - P_R\right) = 0\\
&\alpha_{3,n}\left(\sum_{k=1}^{N_s} f_{k,i,n} q_{k,i} - P_{I,1}\right) = 0\\
&\alpha_{4,n}\left(\sum_{k=1}^{N_r} g_{k,i,n} h_{k,i} - P_{I,2}\right) = 0\\
&\frac{\partial L}{\partial q_{k,i}} = 0, \quad k = 1,\ldots,N_s\\
&\frac{\partial L}{\partial h_{k,i}} = 0, \quad k = 1,\ldots,N_r
\end{aligned}
\tag{41}
$$

For all $n \in S_{PU}$ and $i \in S_R$. It can be shown that $h_{k,i}$ and $q_{k,i}$ can be obtained using the following equations

$$
q_{k,i} = \frac{\sigma_r^2}{2b_{k,i}}\left[\sqrt{\frac{c_{k,i}^2}{\sigma_d^4} h_{k,i}^2 - \frac{4b_{k,i}c_{k,i}}{\left(\alpha_1 + f_{k,i,n}\alpha_3\right)\sigma_r^2\sigma_d^2 \ln 2}} h_{k,i} - \left(2 + \frac{c_{k,i}}{\sigma_d^2} h_{k,i}\right)\right]^+ \tag{42}
$$

$$
h_{k,i} = \frac{\sigma_d^2}{2c_{k,i}}\left[\sqrt{\frac{b_{k,i}^2}{\sigma_r^4} q_k^2 - \frac{4b_{k,i}c_{k,i}}{\left(\alpha_2 + g_{k,i,n}\alpha_4\right)\sigma_r^2\sigma_d^2 \ln 2}} q_k - \left(2 + \frac{b_{k,i}}{\sigma_r^2} q_{k,i}\right)\right]^+ \tag{43}
$$

where $[.]^+ = \max(.,0)$. Using dual-domain and sub-gradient methods [27], we can further obtain α_1, α_2, $\alpha_{3,n}$ and $\alpha_{4,n}$ through iteration,

$$
\begin{aligned}
\alpha_1^{(m+1)} &= \left[\alpha_1^{(m)} + \kappa^{(m)}\left(\sum_{k=1}^{N_s} q_{k,i}^{(m)} - P_T\right)\right]^+\\
\alpha_2^{(m+1)} &= \left[\alpha_2^{(m)} + \kappa^{(m)}\left(\sum_{k=1}^{N_r} h_{k,i}^{(m)} - P_R\right)\right]^+\\
\alpha_{3,n}^{(m+1)} &= \left[\alpha_{3,n}^{(m)} + \kappa^{(m)}\left(\sum_{k=1}^{N_r} f_{k,i,n} q_{k,i}^{(m)} - P_{I,1}\right)\right]^+, \forall n \in S_{PU}\\
\alpha_{4,n}^{(m+1)} &= \left[\alpha_{4,n}^{(m)} + \kappa^{(m)}\left(\sum_{k=1}^{N_r} g_{k,i,n} h_{k,i}^{(m)} - P_{I,2}\right)\right]^+, \forall n \in S_{PU}
\end{aligned}
\tag{44}
$$

where m is the iteration index and $\kappa^{(m)}$ is a sequence of scalar step sizes. Once α_1, α_2, $\alpha_{3,n}$ and $\alpha_{4,n}$ are obtained, we can get the optimal power allocation matrices $\mathbf{Q}_i$ and $\mathbf{A}_i$ and the corresponding achievable data rate R_i when the i-th SU acts as the relay for the SU TX. Repeating the above procedures at all SUs, we then find the one with the maximum achievable data rate.

4-2- Low-complexity approach

The optimal approach performs joint opportunistic relay selection and power allocation and results in the maximum data rate. However, the optimal approach is with very high complexity. Here, we aim to develop an alternate low-complexity suboptimal approach for problem (39). At first, we assume that the available source power is distributed uniformly over the spatial modes, i.e. $q_i^{uni} = \frac{P_T}{N_s}$. Similar assumption applies for $h_{k,i}$ ($k = 1,\ldots,N_r$), i.e. $h_i^{uni} = \frac{P_R}{N_r}$. Also assume that the interference introduced to the PU by each spatial mode of SU TX is equal and hence the maximum allowable power that can be allocated to the k-th mode is $q_{k,i}^{\max} = \frac{P_{I,1}}{N_s f_{k,i}^{\max}}$, where $f_{k,i}^{\max} = \max_{n \in S_{PU}} f_{k,i,n}$. Therefore, the allocated power to the k-th mode in the SU TX, intended for SU i, is $q_{k,i}^* = \min\left\{q_i^{uni}, q_{k,i}^{\max}\right\}$ for $k = 1,\ldots,N_s$ and $\forall i \in S_R$. Similarly, we assume that the interference introduced to the PU by each spatial mode of SU i is equal. Therefore, it can be concluded that $h_{k,i}^{\max} = \frac{P_{I,2}}{N_r g_{k,i}^{\max}}$, where $g_{k,i}^{\max} = \max_{n \in S_{PU}} g_{k,i,n}$. Therefore, the power allocation in the SU i is given by $h_{k,i}^* = \min\left\{h_i^{uni}, h_{k,i}^{\max}\right\}$ for $k = 1,\ldots,N_r$ and $\forall i \in S_R$. Afterwards, the SU i is selected as the cooperative relay such that the following is maximized

$$
i = \arg\max_i \sum_{k=1}^{N_r} \log_2 \frac{\left(1 + \frac{c_{k,i}h_{k,i}^*}{\sigma_d^2}\right)\left(1 + \frac{b_{k,i}q_{k,i}^*}{\sigma_r^2}\right)}{1 + \frac{c_{k,i}h_{k,i}^*}{\sigma_d^2} + \frac{b_{k,i}q_{k,i}^*}{\sigma_r^2}} \tag{45}
$$

After determining the cooperative SU, we calculate the optimal transmit covariance matrix, $\mathbf{Q}_i$, and amplification matrix, $\mathbf{A}_i$, using the approach provided in the optimal approach subsection. As we can see from the simulation results, this approach is almost as good as the optimal approach. However, it is with much lower complexity.

5. Outage Analysis

In order to analyze the outage behaviour of the proposed system, we consider the scenario where the PU transmitters, $\text{PU TX}_1,\ldots,\text{PU TX}_{N_{PU}}$, randomly communicate with their respective receivers, $\text{PU RX}_1,\ldots,\text{PU RX}_{N_{PU}}$. The interval between two transmissions of PUs and the duration of one PU transmission are assumed being random and obeying Exponential distribution with two parameters λ and μ,

respectively. According to queuing theory, the probability of the absence of the PUs, $P(A)$, and the probability of the presence of the PUs, $P(\bar{A})$, can be expressed respectively as $P(A)=\left(\sum_{n=0}^{N_{PU}}\frac{N_{PU}!}{(N_{PU}-n)!}\left(\frac{-}{2}\right)^{n}\right)^{-1}=\sim$ and $P(\bar{A})=1-\sim$. In order to facilitate the analysis of outage, we modify the system model as explained below. First of all, we assume that the transmit signal at the SU TX is white and thereby $\mathbf{Q}=°\mathbf{I}_{N_s}$, where $\mathbf{I}_{N_s}$ represents the $N_s\times N_s$ identity matrix and $°N_s$ is the transmit power of the SU TX. Moreover, the cooperation strategy of the selected SU is assumed to be Decode-and-Forward (DF) strategy. This strategy switch is intended for some reasons, which among them is to obtain a lower bound for the outage capacity of the desired MIMO link. Meanwhile, this assumption facilitates the outage probability analysis, as will be shown below.

In the first time-slot, the spectrum sensing is used to detect whether the PUs are absent. When the PUs are absent, SU TX transmits data to SU RX directly. When the PUs are present, the transmit power of SU TX, $°N_s$, should be limited. However, if $°N_s$ is too low, the data from SU TX cannot reach SU RX. Thus, we use cooperative relaying to transmit signal from SU TX to SU RX through the best relay which is selected out of available SUs. In the sequel, we derive the approximate outage probabilities of the desired SU link, when the PUs are present and when no PUs transmit signals or in other words, the PUs are absent.

5-1- Absence of PUs

We firstly assume that no PU link is transmitting signal. Hence, the SU TX communicates directly with the SU RX and the received signal in the SU RX can be written as

$$\mathbf{y}_d=\mathbf{H}_{sd}\mathbf{x}_s+\mathbf{n}_d \tag{46}$$

Based on the assumptions expressed at the beginning of this section, the achievable data rates of the desired link using the direct channel is given by

$$R^D=\log_2\left|\mathbf{I}_{N_d}+\frac{°}{±_d^2}\mathbf{H}_{sd}\mathbf{H}_{sd}^H\right| \tag{47}$$

where $\mathbf{H}_{sd}\in$ [illegible] $^{N_d\times N_s}$ represents the direct channel in the desired link. It is obvious that the achievable data rates in the desired link, R^D, is a random variable which depends on the random nature of $\mathbf{H}_{sd}$. In a full-rank system, (47) can be simplified by using singular value decomposition (SVD) as

$$R^D=\sum_{m=1}^{N_d}\log_2\left(1+\frac{°}{±_d^2}a_{sd,m}\right) \tag{48}$$

where $a_{sd,m}$, $i=1,\ldots,N_d$ are the non-negative eigenvalues of the channel covariance matrix $\mathbf{H}_{sd}\mathbf{H}_{sd}^H$. The joint pdf of $a_{sd,m}$, $i=1,\ldots,N_d$ is given by [11]

$$p\left(a_{sd,1},\ldots,a_{sd,N_d}\right)=\left(N_d!K_{N_d,N_s}\right)^{-1}\left(\prod_{m=1}^{N_d}a_{sd,m}^{N_r-N_d}\right)\times\left(\prod_{m<n}\left(a_{sd,m}-a_{sd,n}\right)^2\right)\exp\left(-\sum_{m=1}^{N_d}a_{sd,m}\right) \tag{49}$$

where K_{N_d,N_s} is a normalizing factor. To ensure QoS for the desired link, it needs to support a minimum rate. When than the instantaneous achievable data rate is less than the minimum rate, $R_{\min}$, an outage event occurs. In quasi-static fading, since the fading coefficients are constant over the whole frame, we cannot average them with an ergodic measure. In such an event, Shannon capacity does not exist in the ergodic sense [28-30]. The probability of such an event is normally referred to as outage probability. As described in [31], the distribution of the random achievable data rate can be viewed as Gaussian when the number of transmit and/or receive antennas goes to infinity. It is also a very good approximation for even small N_d and N_s, e.g. $N_s=N_d=2$ [24]. As such, for a sufficiently large N_d and N_s, the achievable data rate of the desired link is approximated as [31]

$$R^D\rightarrow N\left(N_d\log_2(1+°),\frac{N_d°^2}{(\ln 2)^2N_s(1+°)^2}\right) \tag{50}$$

Then, we proceed by considering the distribution of the achievable data rate in the desired link as Guassian with the pdf given in (50). Consequently, it can be shown that the outage probability of the desired link in the absence of the PUs can be written as

$$P_{out}^D=P\left(R^D<R_{\min}\right)=Q\left(\frac{N_d\log_2(1+°)-R_{\min}}{\sqrt{\frac{N_d°^2}{(\ln 2)^2N_r(1+°)^2}}}\right) \tag{51}$$

where $Q(.)$ denotes the Q-function.

5-2- Presence of PUs

As described in the previous section, when PUs transmit signals, the direct communication in the desired link must be avoided and the cooperation of the best SU is employed instead. The received signal in the SU RX using the cooperation of *i*-th SU can be expressed as

$$\mathbf{y}_d=\mathbf{H}_{rd,i}\mathbf{x}_{s,i}+\mathbf{n}_d \tag{52}$$

Thus, the achievable data rates of the desired link is given by

$$R_i^C=\frac{1}{2}\log_2\left|\mathbf{I}_{N_d}+\frac{°}{±_d^2}\mathbf{H}_{rd,i}\mathbf{H}_{rd,i}^H\right| \tag{53}$$

It can be concluded that for the case of present PUs, the achievable data rates in the desired link, R_i^C, can be expressed as

$$R_i^C = \frac{1}{2}\sum_{m=1}^{N_d}\log_2\left(1+\frac{\circ}{\pm_d^2}\,\mathsf{a}_{rd,i,m}\right) \tag{54}$$

where $\mathsf{a}_{rd,i,m}$, $m=1,\ldots,N_d$ are the non-negative eigenvalues of the channel covariance matrix $\mathbf{H}_{rd,i}\mathbf{H}_{rd,i}^H$. The joint pdf of $\mathsf{a}_{rd,i,m}$, $m=1,\ldots,N_d$ is given by [11]

$$p\left(\mathsf{a}_{rd,i,1},\ldots,\mathsf{a}_{rd,i,N_d}\right)=\left(N_d!K_{N_d,N_r}\right)^{-1}\exp\left(-\sum_{m=1}^{N_d}\mathsf{a}_{rd,i,m}\right) \times\left(\prod_{m=1}^{N_d}\mathsf{a}_{rd,i,m}^{N_r-N_d}\right)\left(\prod_{m<n}\left(\mathsf{a}_{rd,i,m}-\mathsf{a}_{rd,i,n}\right)^2\right) \tag{55}$$

where K_{N_d,N_r} is a normalizing factor. Once again and similar to the previous discussions, the achievable data rate of the desired link is approximated as [24]

$$R_i^C \to N\left(N_d\log_2(1+\circ),\frac{N_d\circ^2}{4(\ln 2)^2 N_r(1+\circ)^2}\right) \tag{56}$$

Note that the coefficient $1/4$ in the variance of the pdf in (56) is due to the multiplication of $1/2$ in (54). Therefore, the outage probability of the desired link in the presence of the PUs can be written as

$$P_{out}^{C,i}=P\left(R_i^C<R_{\min}\right) =Q\left(\frac{N_d\log_2(1+\circ)-R_{\min}}{\sqrt{\frac{N_d\circ^2}{4(\ln 2)^2 N_r(1+\circ)^2}}}\right) \tag{57}$$

5-3- The outage probability

In this subsection, the outage probability of the proposed Cognitive Cooperative communication protocol based on Beam forming (CCB) is obtained. However, in the case that the DF cooperation strategy is employed and the PUs are present, another possible case in the proposed protocol is when no SU can decode the signal from SU TX. This may be due to detrimental effects of fading and path loss in the link from the SU TX to SUs. In this case, the SU TX indispensably transmits data to SU TX directly with limited power $\circ' N_s$ in order not to disturb the PUs. Assume that Δ_u is a non-empty sub-set of the N_{SU} secondary users who can decode the data of SU TX, i.e. $\Delta_u \subseteq S_R$, and $\bar{\Delta}_u$ is the complementary set of Δ_u. Suppose that α is a null set. Then, the probability of existing no SU to decode the data of SU TX, P_{out}^{α}, can be written as

$$P_{out}^{\alpha}=P\left(\Delta=\alpha\right)=\prod_{m=1}^{N_{SU}}P_{out}^{R,m} \tag{58}$$

and $P_{out}^{R,m}$ (where $m\in S_R$) denotes the outage probability in the link from SU TX to the SUs in the first time-slot. Similar to previous subsections, a good approximate for $P_{out}^{R,m}$ can be obtained as

$$P_{out}^{R,m}=Q\left(\frac{N_r\log_2(1+\circ')-R_{\min}}{\sqrt{\frac{N_r\circ'^2}{(\ln 2)^2 N_s(1+\circ')^2}}}\right) \tag{59}$$

In the following theorem, we derive the outage probability of the desired SU link using the CCB.

Theorem 1: *The outage probability of the desired SU link using the proposed cognitive cooperative communication protocol based on beamforming is*

$$P_{out}=(1-\sim)\left(P_{out}^{\alpha}+\sum_{u=1}^{2^{N_{SU}}-1}P\left(\Delta=\Delta_u\right)P_{out}^{\Delta_u}\right)+\sim P_{out}^{D} \tag{60}$$

where $P_{out}^{\Delta_u}$ is the outage probability of the desired link in the presence of PUs and when the one SUs in the sub-set Δ_u is cooperating with desired link.

Proof. Consider the case that the PUs are present. Then, the probability of event $\{\Delta=\Delta_u\}$, i.e. there exist some SUs which can decode the signal from SU TX, can be written as

$$P\left(\Delta=\Delta_u\right)=\left(\prod_{m\in\Delta_u}\left(1-P_{out}^{R,m}\right)\right)\left(\prod_{m\in\bar{\Delta}_u}P_{out}^{R,m}\right)= \quad u \tag{61}$$

The outage probability of the desired link in the presence of PUs and when the one SUs in the sub-set Δ_u is cooperating with desired link is given by

$$P_{out}^{\Delta_u}=\prod_{i\in\Delta_u}P_{out}^{C,i} \tag{62}$$

Then, the outage probability of the desired SU link in the presence of the PU signals can be written as

$$P_{out}^{\bar{A}}=P_{out}^{\alpha}+\sum_{u=1}^{2^{N_{SU}}-1}P\left(\Delta=\Delta_u\right)P_{out}^{\Delta_u} \tag{63}$$

Finally, it can be concluded that the outage probability of the desired link using the proposed cognitive cooperative communication protocol based on beam forming is given by

$$P_{out}=P\left(\bar{A}\right)P_{out}^{\bar{A}}+P\left(A\right)P_{out}^{A} =(1-\sim)\left(P_{out}^{\alpha}+\sum_{u=1}^{2^{N_{SU}}-1}P\left(\Delta=\Delta_u\right)P_{out}^{\Delta_u}\right)+\sim P_{out}^{D} \tag{64}$$

where P_{out}^{D} is the outage probability of the desired link, when the PUs are absent and is given in (51) and the proof is complete in this way.

6. Simulation Results

In this section, the performance of the proposed CCB protocol is evaluated using simulations. For better comprehending the merit of the proposed low complexity approach (LCA), we will also compare the proposed approach with the approaches using random cooperative SU selection with optimal power allocation matrices (transmit covariance matrix and amplification matrix), referred to as RS-OPA (Random SU-Optimal Power Allocation) and non-optimal power allocation, i.e., the amplification matrix of the randomly selected SU and the

transmit covariance matrix are obtained as described in 4.2, respectively, which is referred to as RS-EPA (Random SU-Equal Power Allocation). All users are assumed to be equipped with the same number of antennas, denoted by *N*.

We set interference limits, $P_{I,1} = P_{I,2} = 0.1$ mW , otherwise stated. There exist 5 PU pairs in the system, otherwise stated. The elements of the channel matrices follow a Rayleigh distribution and are independent of each other. The path-loss exponent is 4, and the standard deviation of shadowing is 6 dB. The number of existing SUs in the system is 20, otherwise stated. The level of noise is assumed identical in the system and equal to 10^{-6} W/Hz.

The data rate in the desired SU link versus the maximum transmit power of SU TX for different number of antennas and various scenarios is shown in Fig. 2. The maximum transmit power of each SU *i*, for all $i \in S_R$, is $P_R = 0.7$ W . Using the Low Complexity Approach (LCA), 50% achievable data rate gain over the RS-OPA is obtained, when $N = 2$. Moreover, LCA leads to only 14% data rate degradation compared with OA, with much lower complexity. When P_T is small, the achievable data rate in the desired SU link increases rapidly with P_T . However, for large amounts of P_T , due to restrictions by the interference limits, the data rate is not sensitive to the P_T . As another observation, it can also be seen that the Random SU and Optimal Power Allocation scheme (RS-OPA) achieves a significant gain in the data rate over the Random SU and Non-Optimal (Equal) Power Allocation scheme (RS-EPA), especially when P_T is small.

The data rate of the desired SU link versus the maximum transmit power of the cooperating SU (P_R) is depicted in Fig. 3. The maximum transmit power of the SU TX is fixed at $P_T = 0.7$ W and the number of existing SUs in the secondary network, N_{SU} , is 20.

As shown in Fig. 4, the achievable data rate in the desired link grows with the number of existing SUs in the CR network. However, this growth saturates from a particular number of SUs which shows that the increasing the number of existing SUs will not necessarily result in the similar increase in the data rate of the desired link. Moreover, deploying larger number of antennas in users, i.e. larger *N*, compensates for the less maximum transmit power of SU TX and the cooperating relay. It must also be noted that the achievable data rate in the system is increased with the number of existing SUs due to multiuser diversity.

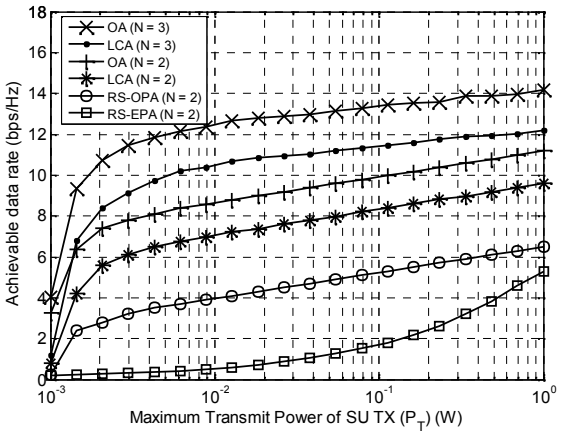


Fig. 2 Achievable data rate in the desired link versus the maximum transmit power of SU TX (P_T)

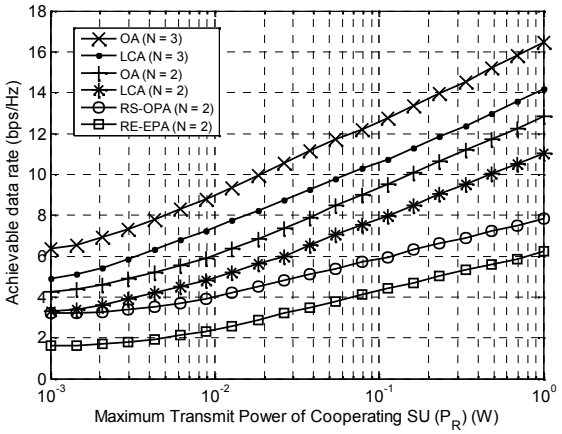


Fig. 3 Data rate in the desired link versus maximum transmit power of cooperating SU (P_R)

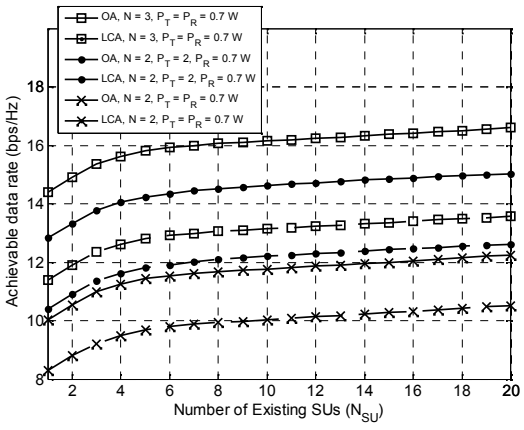


Fig. 4 Data rate in the desired link versus the number of existing SUs

7. Conclusions

In this work, an adaptive transmission protocol based on beam forming for underlay MIMO cognitive radio networks was proposed. It is assumed that when PUs are present, the direct transmission by PUs introduces intolerable interference on PUs. As a remedy, the cooperation of one of SUs was proposed to not only reduce the imposed interference on PUs, but also to maximize the data rates in the SU link. Based on the proposed Cognitive Cooperative communication protocol based on Beamforming (CCB), the joint problems of optimal power allocation and relay selection were solved in the optimal manner. However, due to high complexity of the optimal approach, a suboptimal approach with less complexity was further suggested. Finally, an outage probability analysis was provided to examine the performance of the proposed CCB protocol.

Appendix

Proof of Proposition 1.

It was shown in [25] that if the SU TX works in spatial multiplexing mode, i.e., the SU TX transmits independent data streams from different antennas, the amplification matrix of SU *i* can be written as

$$\mathbf{A}_i = \mathbf{V}_{rd,i} \Lambda_{\mathbf{A}_i} \mathbf{U}_{sr,i}^H \tag{65}$$

where $\Lambda_{\mathbf{A}_i}$ is a diagonal matrix. Therefore, $\mathbf{A}_i$ can be considered as a matched filter along the singular vectors of the channel matrices. In order to use the results of [25] for the case of non-white transmit data of the SU TX and equivalently the transmit covariance matrix is any arbitrary matrix $\mathbf{Q}_i$, we define the equivalent channel matrix $\tilde{\mathbf{H}}_{sr,i} = \mathbf{H}_{sr,i} \mathbf{Q}_i^{\frac{-1}{2}}$. Hence, by adopting the same method as in [25], for any given pair of $\mathbf{A}_i$ and $\tilde{\mathbf{Q}}_i$, there always exists another pair $\mathbf{A}_{i,opt}$ and $\tilde{\mathbf{Q}}_i$ that achieves better or equal data rate in the desired link. In this case, for the case of known $\mathbf{Q}_i$, (65) must be modified as

$$\mathbf{A}_{i,opt} = \mathbf{V}_{rd,i} \Lambda_{\mathbf{A}_i} \tilde{\mathbf{U}}_{sr,i}^H \tag{66}$$

where $\tilde{\mathbf{U}}_{sr,i}$ is obtained by eigenvalue decomposition of $\tilde{\mathbf{H}}_{sr,i} \tilde{\mathbf{H}}_{sr,i}^H$, i.e. $\tilde{\mathbf{H}}_{sr,i} \tilde{\mathbf{H}}_{sr,i}^H = \mathbf{H}_{sr,i} \tilde{\mathbf{Q}}_i \mathbf{H}_{sr,i}^H = \tilde{\mathbf{U}}_{sr,i} \tilde{\Sigma}_{sr,i} \tilde{\mathbf{U}}_{sr,i}^H$.

References

[1] S. Haykin, “Cognitive radio: Brain-empowered wireless communications,” IEEE Journal on Selected Areas in Communications, vol. 23, pp. 201–220, Feburary 2005.

[2] Q. Zhao and B. M. Sadler, “A survey of dynamic spectrum access,” IEEE Signal Process. Mag., vol. 24, no. 3, pp. 79–89, May 2007.

[3] V. Chakravarthy, X. Li, Z. Wu, M. A. Temple, F. Garber, R. Kannan, and A. Vasilakos, “Novel overlay/underlay cognitive radio waveforms using sd-smse framework to enhance spectrum efficiency-part I: Theoretical framework and analysis in AWGN channel,” IEEE Trans. on Communications, vol. 57, no. 12, pp. 3794–3804, December 2009.

[4] R. Zhang, X. Kang, and Y. C. Liang, “Protecting primary users in cognitive radio networks: Peak or average interference power constraint?” in IEEE International Conference on Communications, Dresden, Germany, June 2009, pp. 1–5.

[5] S. Yan and X. Wang, “Power allocation for cognitive radio systems based on nonregenerative OFDM relay transmission,” in International Conference on wireless communications, networking and mobile computing, Beijing, China, Sept. 2009, pp. 1–4.

[6] C. Luo, F. R. Yu, H. Ji, and V. C. Leung, “Distributed relay selection and power control in cognitive radio networks with cooperative transmission,” in IEEE International Conference on Communications, Cape Town, South Africa, May 2010, pp. 1–5.

[7] J. Mietzner, L. Lampe, and R. Schober, “Distributed transmit power allocation for multichip cognitive radio systems,” IEEE Trans. on Wireless Communications, vol. 8, no. 10, pp. 5187–5201, Oct. 2009.

[8] M. Xie, W. Zhang, and K.-K. Wong, “A geometric approach to improve spectrum efficiency for cognitive relay networks,” IEEE Trans. On Wireless Communications, vol. 9, no. 1, pp. 268–281, Jan. 2010.

[9] C. Sun and K. B. Letaief, “User cooperation in heterogeneous cognitive radio network with interference reduction,” in IEEE International Conference on Communications, Beijing, China, May 2008, pp. 3193–3197.

[10] M. A. Beigi and S. M. Razavizadeh, “Cooperative beam forming in cognitive radio networks,” in IFIP Wireless Days, Paris, France, Dec. 2009, pp. 1–5.

[11] E. Telatar, “Capacity of multi-antenna Gaussian channels," Eur. Trans. Telecomm., vol. 10, pp. 585-598, Nov. 1999.

[12] G. J. Foschini and M. J. Gans, “On limits of wireless communications in a fading environment when using multiple antennas," Wireless Personal Commun., vol. 6, pp. 311-335, Mar. 1998.

[13] Y. Yu, W. Wang, C. Wang, F. Yan, and Y. Zhang, “Joint relay selection and power allocation with QoS support for cognitive radio networks,” IEEE WCNC 2013, pp. 4516-4521.

[14] J. Kim, T. Q. Duong and X. Tran, “Performance analysis of cognitive spectrum-sharing single carrier systems with relay selection,” IEEE Transactions on Signal Processing, vol. 60, no. 12, pp 6435-6449, 2012.

[15] M. G. Adian, and H. Aghaeinia, “Optimal resource allocation for opportunistic spectrum access in multiple-input multiple-output orthogonal frequency division multiplexing based cooperative cognitive radio networks,” IET Signal Processing, vol. 7, no. 7, pp. 549-557, 2013.

[16] P. Ubaidulla and S. Aissa, “Optimal Relay Selection and Power Allocation for Cognitive Two-Way Relaying Networks,” IEEE Wireless Communications Letters, vol. 1, no. 3, pp. 225-228, 2012.

[17] L. Li, X. Zhou, H. Xu, G. Y. Li, D. Wang and A. Soong, "Simplified relay selection and power allocation in cooperative cognitive radio systems," IEEE Transactions on Wireless Communications, vol. 10, no. 1, pp. 33-36, 2011.
[18] P. Li, S. Guo, W. Zhuang, and B. Ye, "On efficient resource allocation for cognitive and cooperative communications," IEEE Journal on Selected Areas in Communications, vol. 32, no. 2, pp. 264-273, 2013.
[19] M. G. Adian, H. Aghaeinia, and Y. Norouzi "Spectrum sharing and power allocation in multi input-multi-output multi-band underlay cognitive radio networks," IET Communications, vol. 7, no. 11, pp. 1140-1150, 2013.
[20] M. G. Adian, and H. Aghaeinia, "Spectrum sharing and power allocation in multiple-in multiple-out cognitive radio networks via pricing," IET Communications, vol. 6, no. 16, pp. 2621-2629, 2012.
[21] M. G. Adian, and H. Aghaeinia, "Resource allocation in MIMO-OFDM based cooperative cognitive radio networks," IEEE Transactions on Communications, doi: 10.1109/TCOMM.2014.2327063, 2014.
[22] M. G. Adian, and H. Aghaeinia, "Low complexity resource allocation in MIMO-OFDM-based cooperative cognitive radio networks," Transactions on Emerging Telecommunications Technology, doi: 10.1002/ett.2799, 2014.
[23] J. Liu, N. B. Shroff and H. D. Sherali, "Optimal Power Allocation in Multi-Relay MIMO Cooperative Networks: Theory and Algorithms," IEEE Journal on Selected Areas in Communications, vol. 30 , no. 2, 2012 , pp. 331 – 340.
[24] X. C.X. Wang, H. H. Chen and M. Guizani, "Cognitive radio network management," IEEE Vehicular Technology Magazine, vol. 3, no. 1, 2008, pp. 28-35.
[25] P. Gong, P. Xue, D. Park, and D. K. Kim "Optimum power allocation in a nonorthogonal amplify-and-forward relay-assisted network" IEEE Trans. Veh. Technol., vol. 60, no. 3, Mar. 2011, pp. 890–900.
[26] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press, 2004.
[27] Y. Wei and R. Lui "Dual methods for nonconvex spectrum optimization of multicarrier systems" IEEE Trans. Commun., vol. 54, no. 7, July 2006, pp. 1310–1322.
[28] L. H. Ozarow, S. Shamai, and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," IEEE Trans on Vehicular Technology, May 1994, vol. 43, pp. 359-378.
[29] R. Knopp and P. A. Humblet "On coding for block fading channels" IEEE Trans. on Information Theory, vol. 46, Jan. 2000, pp. 189-205.
[30] E. Biglieri, G. Caire, and G. Taricco "Limiting performance of block-fading channels with multiple antennas" IEEE Trans. on Information Theory, vol. 47, May 2001, pp. 1273-1289.
[31] B. M. Hochwald, T. L. Marzetta, and V. Tarokh "Multiple antenna channel-hardening and its implications for rate feedback and scheduling" IEEE Trans. on Information Theory, vol. 50, Sep. 2004, pp. 1893-1909.
[32] X. Tang and Y. Hua, "Optimal design of non-regenerative MIMO wireless relays,"IEEE Trans. on Wireless Communications, vol. 6, no. 4, 2007, pp. 1398-1407.

Mehdi Ghamari Adian received his B.Sc. degree from Amirkabir University of Technology (Tehran Polytechnic) Tehran, Iran in 2004, his M.Sc. degree from Sharif University of Technology, Tehran, Iran in 2006 and his Ph.D. degree from Amirkabir University of Technology (Tehran Polytechnic) Tehran, Iran in 2014, both in Electrical Engineering (Communication Systems). He is currently an assistant professor in the Electrical Engineering department in University of Zanjan, Zanjan, Iran.
His current research focus is in the areas of cognitive radio networks, cooperative communications and the applications of game theory and benefits of incorporating the MIMO systems in the cooperative cognitive radio networks.

Hassan Aghaeinia received his B.Sc. degree from Amirkabir University of Technology (Tehran Polytechnic) in electronic engineering, in 1987. In 1989 he finished his M.Sc. in Amirkabir University of Technology. Then, in 1992 he received the M.Sc. from Valenciennes University (UVHC), Valenciennes, France. Then he continued his studies towards Ph.D. in UVHC in electronic engineering. He finished his Ph.D. in 1996. From 1996 till present, he is a faculty member in Amirkabir University of Technology, where he is an associate professor in the Communication Engineering Group. His research includes work in digital communications, spread spectrum and advanced communication systems and digital image processing.

Detection and Removal of Rain from Video Using Predominant Direction of Gabor Filters

Gelareh Malekshahi
Department of Electrical Engineering, Sahand University of Technology, Tabriz, Iran
g_malekshahi@sut.ac.ir

Hossein Ebrahimnezhad*
Department of Electrical Engineering, Sahand University of Technology, Tabriz, Iran
ebrahimnezhad@sut.ac.ir

Received: 18/Apr/2014 Revised: 16/Aug/2014 Accepted: 18/Nov/2014

Abstract
In this paper, we examine the visual effects of rain on the imaging system and present a new method for detection and removal of rain in a video sequences. In the proposed algorithm, to separate the moving foreground from the background in image sequences that are the frames of video with scenes recorded from the raindrops moving, a background subtraction technique is used. Then, rain streaks are detected using predominant direction of Gabor filters which contains maximum energy. To achieve this goal, the rainy image is partitioned to multiple sub images. Then, all directions of Gabor filter banks are applied to each sub image and the direction which maximizes the energy of the filtered sub image is selected as the predominant direction of that region. At the end, the rainy pixels diagnosed in per frame are replaced with non-rainy pixels background of other frames. As a result, we reconstruct a new video in which the rain streaks have been removed. According to the certain limitations and existence of textures variation during time, the proposed method is not sensitive to these changes and operates properly. Simulation results show that the proposed method can detect and locate the rain place as well.

Keywords: Inpainting; Background Subtraction; Removal of Rain; Gabor Filters; Rain Detection.

1. Introduction

In indoor environment, video is recorded ideally because of artificial lighting. But in outdoor environment, it is important to remove weather effect because natural environment have diverse noise such as steady disturbances (fog and mist) and dynamic (rain and snow). To improve image quality, we need some methods to remove them. The rain in the foreground is an unwanted component of the background. The rain has distribution of droplets that falling at high speeds. Each drop with refraction and reflection of the environment causes extreme changes in the intensity of an image. Moreover, the intensity of rain appears as blurred movement and therefore it dependent on the background intensity.

Rain detection in images is one of the most challenging problems in computer vision. Recent years due to advances in computer graphics tools, much research has been done in this field and several methods have been designed that aim to determine the location and range of rain streaks in video and image and of course removal of them to improve the quality.

At first, removing the effects of bad weather was developed by Nayar and Garg in which the constant weather, water or smoke, were considered very small floating particles in the air. Using the dispersion model, the color of pictures achieved good results. In the dynamic weather, they developed rain physical properties where the rain area is detected by using properties of pixels in the certain time intervals. To reduce rain effect they used the false positives and correlated in space and time [1].

Zhang proposed a method which includes both temporal and chromatic properties of the rain on the video. In temporal properties, a pixel is not always impressed by rain in video and the color properties that contain change of value by R, G and B, in the rainy pixels are almost identical. K-Means clustering was used to divide the background and rain; however this method is not suitable for real-time processing [2].

Garg and Nayar proposed another method by adjusting camera parameters such as camera exposure time, diaphragm, etc. which remove the rain for the period of video recording, but this method cannot be applied to heavy rainfall and vague moves [3].

Shen and Xue detected and removed the rain by using the definition of the character of light field and motion field. In this case, the intensity of each pixel of the current frame is compared with the intensity of pixel of same frame in space neighborhoods and intensity of pixel of another frame in temporal neighborhoods. Also, using the motion field, distinction between the rain streak and other moving objects are created. So, using data of only three frames identified the rain, and for removal of the pixels in the detected rain area, an anisotropic diffusion smoothing method was proposed as a temporal-spatial filter [4].

* Corresponding Author

In the research of Barnum and Narasimhan that was based on the combination of the modeling of realistic streaks and knowledge of dynamic weather, unlike previous works which were seek to detect rain pixels or individual pieces, this method was dealing to rain and snow as a general phenomenon. In order to determine the influence of the rain and snow in the video, it was developed a general model in frequency space [5].

Other work was performed by Hi Lee and Joo Park in which by using the recursive data processing and using the Kalman filter, they estimate the intensity of each pixel and by comparing the intensities, detect rain streaks [6]. Further work in this field was done by Bossu, Hautière and Tarel, at that, a system based on computer vision is presented which detects the presence of rain or snow. To separate the foreground from the background in image sequences, a classical Gaussian Mixture Model is used. The foreground model serves to detect rain and snow, since these are dynamic weather phenomena. Selection rules based on photometry and size are proposed in order to select the potential rain streaks. Then a Histogram of Orientations of rain or snow Streaks (HOS), estimated by the method of geometric moments, is computed and it is assumed to follow a model of Gaussian uniform mixture [7].

Due to the random spatial distribution and fast motion of rain, removal of rain in video is a more difficult problem, Geng and Qi Submitted a background subtraction based on sample model to remove the rain. They analyze the properties of rain and establish the sample model with values randomly taken in the spatial neighborhood of each pixel on the first frame so better to classify detected rain by background subtraction. In addition, the movement of objects will cause the corresponding color pixel brightness values to change significantly. The H component of the HSI color space was applied to reduce the impact of moving objects on rain removal. Experimental results show that this method not only can eliminate a good rain compared with existing methods, but also have faster processing speed [8].

The other method is using the properties of the image. Detection and removal of rain requires discrimination of the rain and non-rain pixels. Accuracy of the algorithm depends upon this discrimination. This is done by Tripathi and Mukhopadhyay, where the merits and demerits of the algorithms are discussed that motivate the further research. A rain removal algorithm has a wide application in tracking and navigation, consumer electronics and entertainment industries [9].

A method based on the framework of a single image based on the proper technique for formulating rain removal remove rain as a problem of image analysis based on morphological component analysis of expression has been proposed by Li-Wei Kang ،Chia-Wen Lin and Yu-Hsiang Fu. Rather than the direct application of conventional image analysis techniques, in the proposed method the first two parts of the image using a bilateral filter decays to low-frequency and high-frequency (HF). At that time, performing dictionary learning and sparse coding portion of the HF portion of the "rain component" and "component without Rain" is resolved. As a result, the rainy part of the image can be removed successfully, while preserving the most image detail [10].

The next algorithm, which is highly efficient and simple and has been proposed by A.K. Tripathi and Mukhopadhyay for detection and removal of rain from video sequence, is used for the spatial and temporal properties. Fortunately, the spatial and temporal properties of the pixel are involved to separate the non-rainy and rainy pixels. Therefore, the proposed algorithm, may be involved fewer consecutive frames to reduce the buffer size and Latency which significantly reduces the complexity and run-time. This new algorithm is not supposed to operate as well for different shape, size and speed of raindrops. The proposed method reduces the buffer size and therefore the cost of the system, the delay and energy consumption slows down, significantly. For performance evaluation, as well as to avoid errors and misdiagnosis, a new metric is introduced in the temporal and spatial variation [11].

In the work of Xudong Zhao and Peng Liu which is used for both static and dynamic scenes, K-means classification algorithm is used to identify and detect rain. Then, the rain histograms removal is performed [12]. There are other methods in which the median filter is applied to each pixel. Of course, precision of this method is very low.

In the rest of paper, in section (2), we pay attention to express a variety of weather conditions and physical characteristics of raindrops. In section (3), we provide the proposed method in which the original video frames or subtracted frames are extracted at first and then to detect the rainy pixels, dominant direction of Gabor filters are used. Finally, image completion method is employed to remove the effect of rainy pixels. In Section (4), implementation results are obtained and discussion is performed. Conclusion is performed in Section (5).

2. Types of Weather Conditions and Physical Characteristics of Raindrops

Outdoor vision systems are used for various purposes such as surveillance and navigation.

2-1- Types of weathering

In order to develop a vision system that can be used under all weather conditions, it is essential to get the visual impacts of different climate models and develop algorithms to remove them. Weather conditions can be broadly classified into Fixed (fog, mist, smoke, cloud) and Moving or dynamic (rain, snow, hail). In the case of steady weather, droplets are too small (1–10 μm) to be individually detected by a camera. The intensity produced at a pixel is due to the collective effect of a large number of droplets within the pixel's solid angle (see Fig. 1(a)). Therefore, volumetric scattering models such as attenuation and air light can be used to effectively describe the appearance of steady weather.

In this paper, we will focus on the problem of rain. Rain includes the distribution of a large number of drops with different sizes that are falling at high speed. Each droplet behaves like a transparent sphere that performs reflection and refraction of light from the environment into the camera. Thus, Rain drops are the sample of the rain and

they have all of rain's features, but we worked on the drops to have simple examine and consider the more details.

The result of drops falling at high speed is the light intensity fluctuations at different times in images and movies. In addition to the limited time exposure of camera, the changes in light intensity caused by the movement of rain causes the context to be ambiguous. Thus, the visual effects of rain are the combined effects of rain and metering dynamic environment. Rain has been widely studied in the fields of atmospheric science, remote sensing and communication signals. However, the effect of rain on camera to view a scene in the natural environment is very different and sometimes remains unknown.

2-2- Physical Characteristics

Shape of a Raindrop. A raindrop experiences rapid shape distortions as it goes down, a phenomenon often referred to as oscillations. For most vision tasks, the effects of oscillations are insignificant and for this reason, a raindrop can be assumed to have a fixed shape often referred to as an *equilibrium* shape. The equilibrium shape of a drop depends on its size. Smaller drops are usually spherical in shape. However, as the size of the drop increases, it tends towards an oblate spheroid shape.

Velocity of a Raindrop. As a raindrop falls, it attains a constant speed, called the terminal speed (Manning, 1993). Gunn and Kinzer (1949) present an empirical study of the terminal velocities of falling raindrops for different drop sizes. Their observations show that the terminal speed (*v)* of a raindrop can be expressed as a function of its radius.

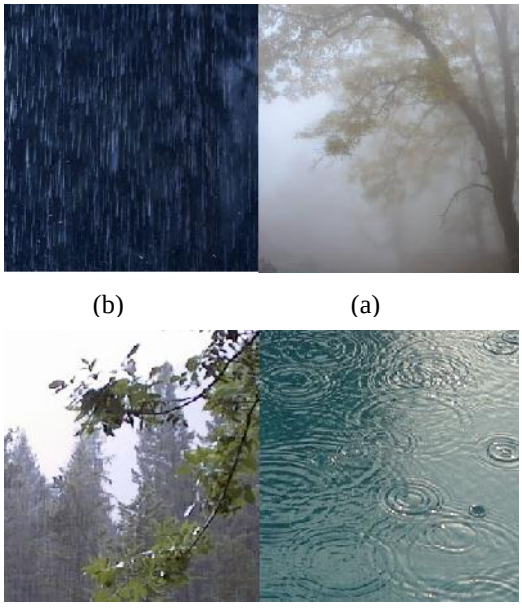


Fig 1: The different climatic conditions. Image (a) is foggy scenes pixel intensity caused by tiny flakes fog and haze are too many of them. Figure (b) shows an image of the scene taken on a rainy day and the strands are observed due to the motion of individual droplets. Blue waves in the image (c) is seen to represent the rain with variable tissue with time. Figure (d) the picture is Rain with scenes with complex motion.

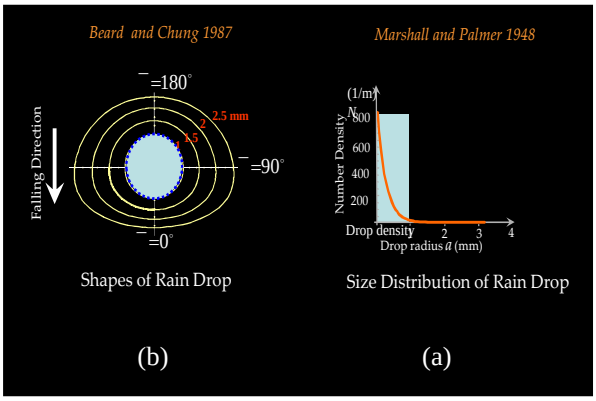


Fig 2: Distribution Marshall - Palmer count raindrops, as a function of their size. Figure (a) Note that the density drops exponentially with the size of the rain falls. Image (b) is the shape of raindrops of different sizes (0.5 to 2.5 mm). Due to the drop in flat form (an oblate spheroid shape) are falling [15].

Raindrops line. As we explained in the previous paragraphs, assuming the small amount of distortion and considering the constant speed of falling raindrops, the raindrops can fall in constant line with fix direction. With these assumptions, we use some approaches to detect streaks of rain drops in a certain direction.

Raindrop size. Raindrops turn to show a wide distribution of sizes. A typical ***distribution*** process used for the rain drop size distribution of the Marshall - Palmer [15].

2-3- Physical properties of rain

Rain is a collection of water droplets of different shapes and sizes with randomly distributed that move at high speeds. The physical properties of rain have been widely studied in atmospheric sciences. Here, we summarize these properties in brief and make observations that are relevant to our goal of modeling the appearance of rain. The size of a raindrop typically varies from 0.1mm to 3.5 mm. The distribution of drop sizes in rain is given by the Marshall-Palmer distribution. Figure 2(a) shows the distribution for a typical rainfall. Note that the density of drops decreases exponentially with the drop size. The shape of a drop can be expressed as a function of its size. Figure 2(b) shows the shapes of raindrops of various sizes.

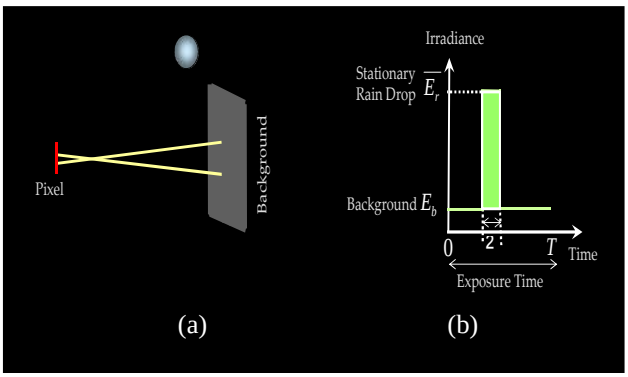


Fig 3: Changes in pixel intensity due to rain drops falling. (a) it is seen that the average pixel intensity because of the rain and because of the background scene. It should be noted that the energy drop is greater than the background energy. (b) is also shown that in $_2 \leq 1.18^{ms}$ less than the camera exposure time (T) is a drop of rain on the pixels of the image [1].

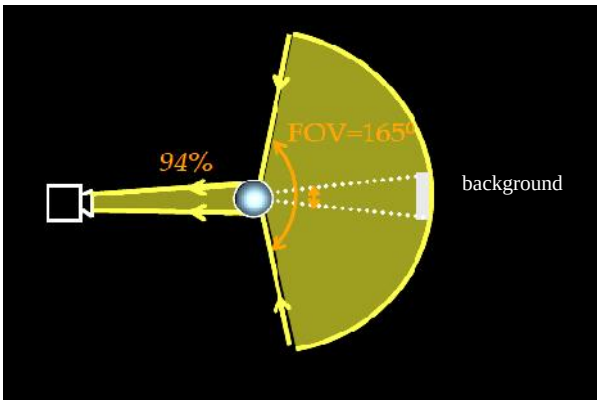


Fig 4: Lighting and visibility drops. Represents the field of view is approximately 165 degrees of raindrops.

Smaller raindrops are generally spherical in shape while larger drops are similar to oblate spheroids. In a typical rainfall, most of the drops are less than 1mm in size, as seen in Figure 2(a). Therefore, most raindrops are spherical and we will use this approximation in our work.

As a drop falls through the atmosphere, it reaches a constant terminal speed. The terminal speed v of a drop is also related to its size and is given by:

$$V = 200\sqrt{a} \quad (1)$$

Raindrops are distributed randomly in 3D space. This distribution is generally assumed to be uniform. Moreover, it can be assumed that the statistical properties of the distribution remain constant over time. These assumptions are applicable in most computer vision scenarios. Complexity of spatial and temporal fluctuations in rainfall recorded image depends on several factors:

• drop distribution and its speed
• Ambient lighting and background scene
• The intrinsic parameters of the camera

The changes in intensity of all pixels along a streak line, has a linear relationship with the intensity of the background light occluded by the streak. So, there is a limit for parameter τ. The maximum value of τ is about 1.18*ms* that is much lower than the camera exposure time (T=30*ms*). Time τ is the time that a raindrop projected on the intended pixel. As illustrated in Fig. (3-b), in the time interval zero to τ and τ to T, there is no change in light intensity since in these time intervals the drop has not reached to the pixel or has passed from the pixel. Therefore, the light intensity raindrops are much more than the intensity of background and during the drop passing from background we face with drastic changes in intensity.

2-4- Brightness of a stationary raindrop

Raindrops act similar to lenses refracting and reflecting environment radiances towards the camera. Detailed geometric and photometric models for the refraction through and reflection from a spherical raindrop have been developed. These models show that raindrops have a large field of view of approximately 165◦ (see Figure 4) and the incident light that is refracted towards the camera is weaken by only 6%.

3. Proposed Method

The proposed algorithm has several sections of which the most important ones are rain detection and removal. In the following, different stages of the proposed method are presented. To increase the quality of rainy clip, texture synthesis theory can be used because rain has regular and integrated tissue despite the integrated stripes which it has in a special direction. One of the tools which can be used for extracting features is use of Gabor filter which detects the oriented lines well and our goal is to identify the rain stripes which have individual directions. However, we have used a special kind of detection to get better result and achieve more successful detection and better removal and restoration. In our method, direction has not been directly introduced to Gabor filters. However, in each window with small dimensions (each sub image) the dominant direction is found. Using Gabor filter with dominant direction, accuracy of detection increases and rain stripes are better detected. Process of filling the lost zones of image by preserving statistical characteristics and integration in the entire image is called image completion. In other words, this process completes an imperfect image which has lost zones such that final results can be visually acceptable. In addition, image completion process should fulfill the following specifications:

- It should be able to complete complex natural images.
- It should be implemented on the large lost zones.
- All of the steps should be done automatically without human involvement.
- Image completion process should be able to solve texture synthesis problem.

Considering the mentioned reasons, image completion is a very challenging and considerable problem. Of course, image completion is applied in many fields such as graphic application, edition, restoration of film and photo. In recent years, many studies have been conducted on image completion.

3-1- Conversion of video to frames

In this stage, we select a video with complex textures which has been recorded from rain scene as input and divide them into their related frames. The selected video has many frames some of which are shown in Figure 5.

3-2- Rain detection

Considering structure of rain stripes which are usually placed in one direction and have a texture like structure, we use set of Gabor filters for detecting them. Gabor filter is used as a linear filter for extracting features from images. One of these features is extraction of directional lines which we consider here. In addition, optimal localization properties are observed in both amplitude and frequency spaces. As a result, it is a suitable method for dividing textures, detecting goal, analyzing document, detecting edge, identifying retina and representing image. Gabor filter $h(x, y)$ can be considered as a sinuous signal in special frequency and direction which has been modulated with a Gaussian push based on Eq (2).

$$h(x,y) = s(x,y)g(x,y) \quad (2)$$

In this equation, s is a complex sinuous function and g is a two-dimensional Gaussian function:

$$s(x,y) = e^{-j2\pi(u_0x+v_0y)} \quad (3)$$

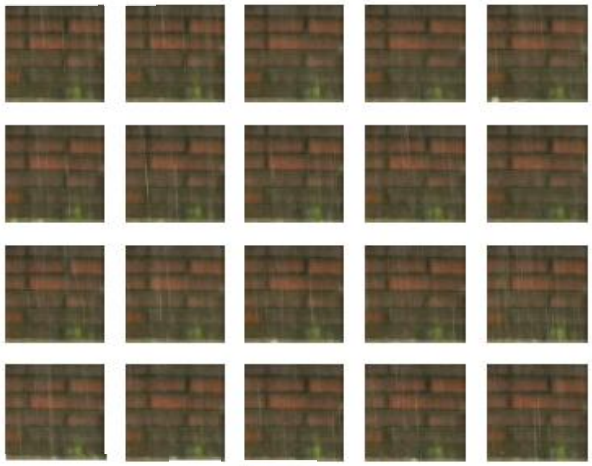


Fig 5: Some video frames for rainy scene

$$g(x,y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)} \quad (4)$$

As a result, Gabor filter impulse response is defined with the following equation:

$$h(x,y) = e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)} . e^{-j2\pi(u_0x+v_0y)} \quad (5)$$

Frequency response of Gabor filter is obtained from:

$$H(u,v) = G(u-u_0, v-v_0) \quad (6)$$

Where,

$$G(u,v) = 2\pi\sigma_x\sigma_y e^{-2\pi^2(u^2\sigma_x^2+v^2\sigma_y^2)} \quad (7)$$

In Figure (6) shows frequency response of Gabor Filter for given angles. Gabor filter has strong response in a direction where variety is in the direction of the same angle. In this case, simple smoothing action which follows a special threshold is sufficient for segmenting image into four zones relating to lines in four directions (see Figure 7). We obtain the all components of image which their energy is centralized in (u_0, v_0) by passing image through a Gabor filter with the defined parameters $(u_0, v_0, \sigma_x, \sigma_y)$. Spatial response of Gabor filter is shown in Figure (8) for some different angles and frequencies. [13]. In Gabor filters bank, a similar set of continuous and interrelated family is produced by changing delay and rotation angle. In rain detection application, rain lines are identified by introducing special direction to filter and mentioning the related mathematical relations with acceptable accuracy. Of course, a threshold limit is defined for increasing accuracy to express Gabor filter and its different values change accuracy of result. In this case, using the higher threshold limit, the more rain stripes are detected.

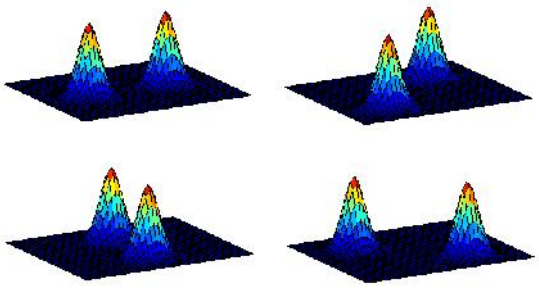


Fig 6: Diagram of frequency response of Gabor filter for different values of u0 ،v0 relating to four directions of 0, 45, 90, 135.

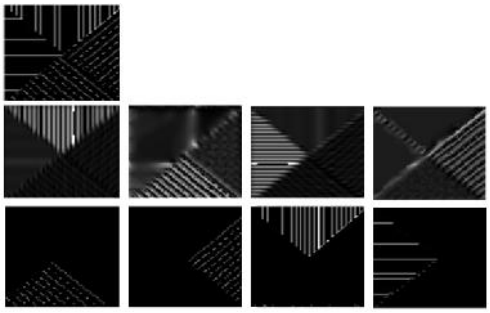


Fig 7: Output of Gabor filter in four directions when input of an image contains lines in similar frequencies but in different angles [16].

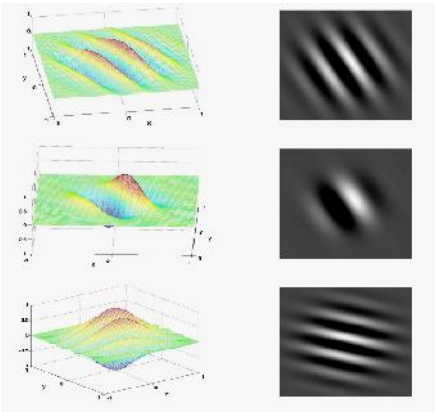


Fig 8: An example of the performance of Gabor filters with different frequencies and directions. Left: Three-dimensional graph of the Gabor filter Right: Gabor filter with intensity image [13].

This result is desirable in terms of high detection accuracy but it decreases accuracy of removal which is explained later. Therefore, we cannot consider its value very high. Figure 8 shows example of Gabor filter performance with different frequencies and directions. The first column is three-dimensional representation and the second column is diagram of its amplitude in gray level of image [13].

Since, rain stripes suddenly collapse and may be exposed to turbulences such as wind and changes from a frame to another frame for collapse, we should add a stage before applying the Gabor filter which intelligently detects direction of rain stripes in each desirable filter considering

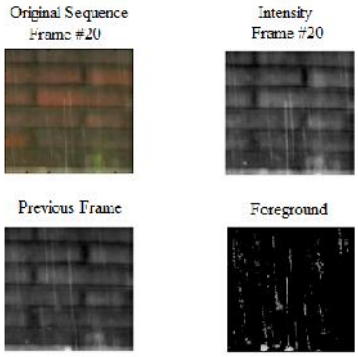


Fig 9: Separating foreground from background considering the previous and next frames. In this image which relates to frame 20 of the main video, mobile foreground is separated from fixed background considering place of mobile components in the previous and next frames.

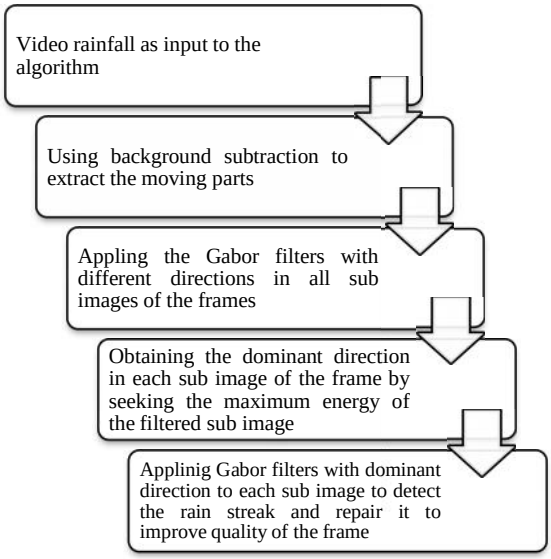


Fig 10: Flowchart of the proposed algorithm for rain detection and removal with dominant direction of Gabor filters.

The previous frame (Fig 9). For this task, we should go forward based on flowchart of Fig (10). It means that we first receive main video which has rain drops as input and then detect background of the moving components of the image which are the rain stripes using subtraction and the result of this section is binary image which indicates moving components. Now, we should identify the pixels which are rainy pixel candidates by inspecting features of rain drops and their sizes to reduce range of selection and finally increase speed of calculations.

Now, we obtain maximum energy by calculating dominant direction of Gabor filter on these rainy pixels through which direction of fully rainy pixels is detected. Using these directions detected in the stage of Gabor filter with dominant direction which is different for each frame, Gabor filter is applied at this time not with unique fix direction but with adaptively detected dominant directions and detection stage is performed. Therefore, we increase accuracy of detection as most as possible and receive more acceptable result by adaptive filtering. Estimation of the dominant direction of Gabor filter is performed by applying Gabor filter defined in Equation (8) to the sub-image I_{in} using Equation (9).

$$g(x,y,f_r,\theta) = e^{-\frac{(x^2+y^2)}{2\sigma^2}} \cos\left(2\pi f_r(x\cos\theta + y\sin\theta)\right) \quad (8)$$

$$I_{out}(x,y,f_r,\theta) = \iint I_{in}(p,q)\, g(x-p, y-q, f_r,\theta)\, dpdq \quad (9)$$

Therefore, dominant direction of Gabor filter is defined as the maximum argument θ in the energy of the filtered sub-image I_{out}:

$$\theta_{dominant} = argmax_\theta \left(\sum_{x=1}^{M}\sum_{y=1}^{N}\left(I_{out}(x,y,f_r,\theta)\right)^2\right) \quad (10)$$

3-3- Rain removal

The method which is applied for removing the effect of rain is the video repair method. This method is applied for repairing the destroyed parts of photo and video. Basis of this algorithm is on displacing of the destroyed pixels with perfect ones.

Inpainting: During the past years, interest in the field of video repairing has increased considerably among the research population. Some examples of the applications of this method are as follows:

Deleting unwanted object: There may be static or dynamic objects in a film which are not desirable.

Revising fiction: Video repairing also can change special scenes of film. For example, it censors a motion which is not suitable for goal.

Repairing video: Some *films* enclose scratch or dust spots or damaged frames which can be deleted with this method.

The applied algorithm is based on this class and repairs the scenes damaged by rain stripes [18].

In this work, a framework is presented for the lost pieces which are moving from background of a recorded video sequence with fixed camera. Generally, the zone which is inpainted may be fixed or moving, located in background or foreground and also may be obstructed with a fence. In painting-based algorithms comprise of two parts. The first part includes simple preprocessing and the second part includes video painting. In the preprocessing stage, almost all sections of each frame are divided into background and foreground and two image strips are created using this division which helps present results based on time and performance of algorithm is enhanced by reducing search space. In inpainting section, we first repair moving objects in the foreground which has been obstructed with the zone which will be inpainted. For this purpose, the holes are filled using priority-based design as far as possible by copying information from the moving foreground in other frames and then the remaining holes bare inpainted with background. Then, frames should be ranked as far as possible to be directly copied. The remaining pixels are filled by expanding texture synthesis spatial techniques in space and time domains. The presented framework has different advantages over artistic algorithms which act with similar types of data and limitation and this permits some camera motions which have simple and rapid implementations not to need statistical models of background and foreground [15].

Generally, we receive binary image using results of Gabor filter recognition and then analyze its pixels. In each frame, the value of each rainy pixel is filled with

corresponding pixel value in the next frame, of course, in case that pixel is not rainy in the next frame, otherwise, it is filled with next frames to fill all rainy pixels in a frame with value of the same pixel which is not rainy in the neighbor frames. Of course, we have performed this action until special stage which is given as threshold limit again and this affects accuracy of removal. In this way, using the higher number of the stage, the better removal action will be performed and the higher quality images will be resulted.

4. Experimental Results

The proposed method in this paper is based on this idea that which direction of Gabor filter must be applied to rain streaks in each zone to detect them as accurate as possible. Basis of the work is such that all zones of the video frames are recognized from 0 to 180 with angular changes of 10 degrees through Gabor filter to find the dominant direction in each zone on each frame.

Therefore, dominant direction which has the maximum energy is found in each frame using 18 Gabor filters. Then, we increase the number of Gabor filters (to make possible applying different Gabor filters for each zone) on each frame for more accuracy. The results are available for (1*1), (3*3), (4*4) and (5*5) zones in Figures 11 to 14. To increase accuracy of detection and improve image quality, differential frame has been used instead of main frame meaning that input image of Gabor filter is subtraction of two sequential frames.

Now, these results of detection section are given as the input image to the stage of rain removal which is video repairing. In the set of images of Figure (15), result of detection using dominant direction of Gabor filter (the above figures) is seen and this has been given from the left to the right for detection using Gabor filter (1*1), (2*2), (3*3), (4*4) and (5*5). As appeared in the images, quality of image has improved more acceptably and more strongly than

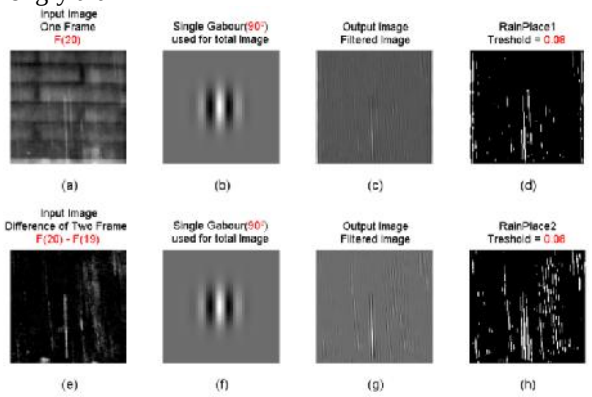


Fig 11: Rain detection of each frame using background subtraction and filtering with Gabor. Figures (a) to (d) relate to frame 20 of main video, Gabor filter in direction of 90 degree, filtered image and rain place after thresholding, respectively. In the next row, image (e) relates to differential image of two sequential frames. Images (f) to (h) are like definitions (b) to (d) which have been executed for differential frame.

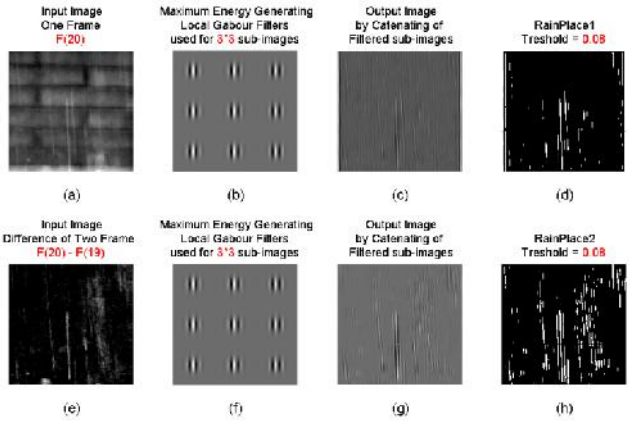


Fig 12: Rain detection of each frame using background subtraction and dominant direction of Gabor filter (3*3). Figures (a) to (d) relate to frame 20 of main video, Gabor filters with dominant directions, filtered image and rain place after thresholding, respectively. Other explanations are similar to Figure (11).

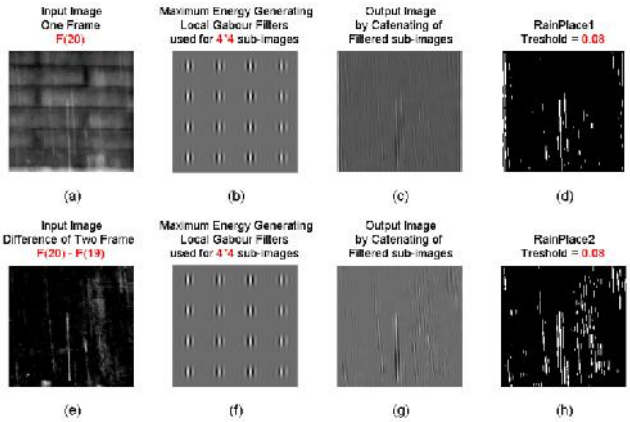


Fig 13: Rain detection of each frame using background subtraction and dominant direction of Gabor filter (4*4). Figures (a) to (d) relate to frame 20 of main video, Gabor filters with dominant directions, filtered image and rain place after thresholding, respectively. Other explanations are similar to Figure (11).

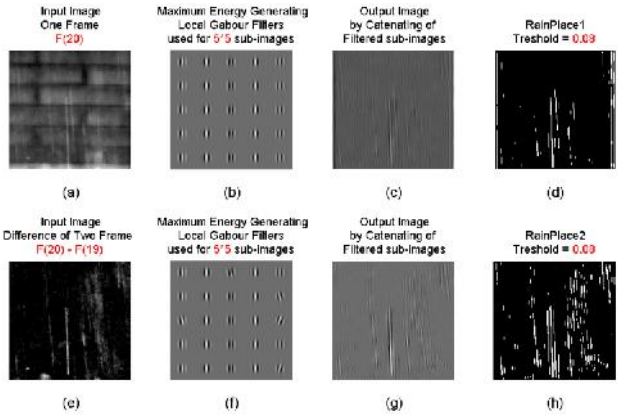


Fig 14: Rain detection of each frame using background subtraction and dominant direction of Gabor filter (4*4). Figures (a) to (d) relate to frame 20 of main video, Gabor filters with dominant directions, filtered image and rain place after thresholding, respectively. Other explanations are similar to Figure (11).

Fig 15: Results of rain removal stage using dominant direction of Gabor filter on comp pool database. In this set of images, result of detection and recognition using dominant direction of Gabor filter (the above Figures) is seen which is given for detection using Gabor filter (1*1) to (5*5) from the left to the right respectively.

Fig 16: Results of rain removal stage using dominant direction of Gabor filter on comp magnolia database. In this set of images, result of detection and recognition using dominant direction of Gabor filter (the above Figures) is seen which is given for detection using Gabor filter (1*1), (3*3) and (5*5) from the left to the right, respectively.

The former algorithms [20] from the left to the right with increase of the number of Gabor filter directions. Of course, it should be noted that the threshold value is not very high because it also leads to image obscurity. For more study and ensuring results of algorithm, operations on other data are tested and its results are given in Figure (16). As it is evident, this image relates to a video with foreground motion of person. Therefore, rain removal is not performed as well with Gabor filter because the foreground is moving and causes error in background directional lines. However, for Gabor filter, there are (3*3), (4*4), and (5*5) from the right to the left, respectively, and obscurity is seen to some extent and progress of this case should be discussed. The proposed method is compared with the results of [1] in Figure 17. As it is seen, the reconstructed frames by the proposed method is less opaque than the method [1] and this is due to the direct placement of information from the rainless frame to the rainy frames and the lack of averaging among the frames. Figure 18 shows another comparison of our method with the methods of Zhang [2] and Nayar [3]. As it is evident, our method has removed the rain effect better than the others.

All of the methods presented in the papers for rain removal are evaluated based on the resulting image and video quality. As far as we know, there is no quantitative evaluation in the papers in this field. With this acclaim, we propose a way to assess the rain removal quality by quantitative values. It is intensity that we report it among the passages. For example, figure (19) is the distribution for a pixel with rain at this image, our result is shown with black and the other method (Kalman Filter [6]) is shown by cyan. We know that the pixel with rain have more light intensity rather than other pixel, our result show it better than other one.

To know how the method works well, we have reported rain rate detection of different methods. In table (1), there are three columns, the first one is measured rain rate by National Climatic Data Center and the second one is the reported rain rate by our method. The result shows that our method could detect rain as well as we expect. At the end, we have reported rain rate by Nayar in third column. By comparison between our method and Nayar's method, we can see the better result for the proposed method. In table (2), we have reported the result of rain removal that shows the success of method in rain removal.

The examined image has 14038 pixels from which 6685 pixels are rainy pixels and the remaining pixels (7353) are not rainy pixels. As the table illustrates, 5421 rainy pixels (TP = 81%) have been removed successfully.

Of course, 327 number of not rainy pixel have been detected as rainy pixels (FP =4%).

5. Conclusion

In this paper, attempt was made to detect and remove rain on video by studying features of rain and extracting them and also using new methods. For this work, the best detection method based on feature has been used. At the beginning, Gabor filter was directly applied with dominant direction introduced to the input data as rainy video frames. Then, to increase accuracy of detection in this algorithm which acted based on dominant direction of Gabor filter, direction of rain stripes were also detected but because this presented method was not able to extract rainy parts, another stage was added to it as background subtraction. In this case, foreground can be separated from background with a background image without any foreground and by reducing it from other frames and applying color zoning and lighting to the images obtained from subtraction. Background subtraction based method was combined with

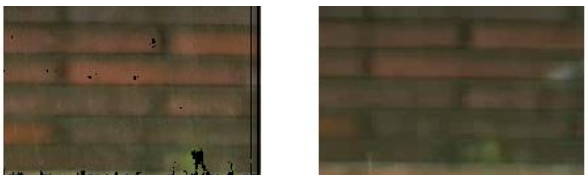


Fig 17: Comparing the proposed method with method [1]. The left image is result of the proposed method and the right image is result of paper [1]. Another advantage is that rain removal in the proposed algorithm causes obscurity of background scene.

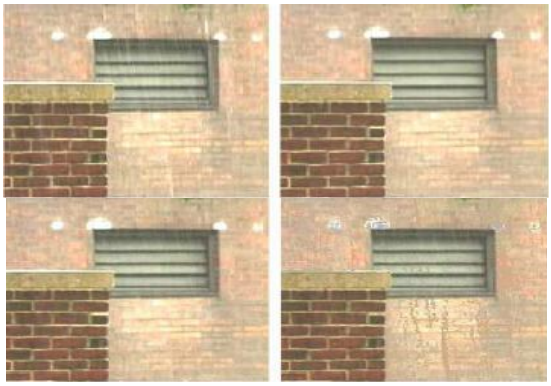


Fig 18: Comparing the proposed method with Zhang's and Nayar's methods. At first row, there are original scene and the result of our method. In the second row, Zhang and Nayar's result are shown. As it is observable, our method has better result for rain removal.

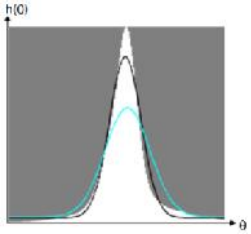


Fig 19: Comparing the proposed method with method [6]. The black one is the intensity of rain pixel for our method and the blue one is for rain removal with kalman filter.

Table 1: A comparison of rain rate measured by rain rate reported.

Type of rainfall	Measured rain rate (mm/hr)	Reported rain rate (mm/hr) by proposed method	Reported rain rate (mm/hr) by Nayar
Light	1.49	0.96	0.93
Moderate	3.631	3.012	2.963
Heavy	14.12	11.47	10.947

Table 2: Removal result based on the use of the Predominant Direction of Gabor Filters.

	Operation of Method	
	Removed	Not Removed
Rainy Pixels	5421	1264
Not Rainy Pixels	327	7026

The method presented based on dominant direction of Gabor filter. Then, detection stage ends by separating foreground including rain from background and video repairing and completion stage can be used for filling zones of pore which are the extracted mobile foregrounds for rain removal stage. In other words, rain foreground is separated from fixed background and we receive the image without recovered rain. This algorithm gives better result than other algorithms considering lack of averaging in neighborhood or sequential frames and also direct placement of frames information. As the future works, we try to remove effects of fixed climate such as fog from video. In rain problem, although we reduced detection error compared with other presented methods, it should be studied yet and better results could be achieved. The proposed algorithm has no high efficiency in the case where mobile object is in background or the background has been obstructed by a person or an object. Therefore, a method should be presented to solve this problem when mobile object is obstructed for some sequential frames.

References:

[1] K. Garg, S.K. Nayar, "Detection and removal of rain from videos", in Computer Vision and Pattern Recognition, Proceedings of the 2004 IEEE Computer Society Conference on, IEEE, 2004, Vol. 521, pp. I-528-I-535.

[2] X. Zhang, H. Li, Y. Qi, W.K. Leow, T.K. Ng, "Rain removal in video by combining temporal and chromatic properties", Multimedia and Expo, IEEE International Conference on, IEEE, 2006, pp. 461-464.

[3] K. Garg, S.K. Nayar, "When does a camera see rain?", Computer Vision, 10th IEEE International Conference on, IEEE, 2005, pp. 1067-1074.

[4] M. Shen, P. Xue, "A fast algorithm for rain detection and removal from videos, in Multimedia and Expo (ICME), IEEE International Conference on, IEEE, 2011, pp. 1-6.

[5] P.C. Barnum, S. Narasimhan, T. Kanade, "Analysis of rain and snow in frequency space", International Journal of Computer Vision, Vol. 86, No. 2-3, Jan. 2010, pp. 256-274.

[6] W.-J. Park, K.-H. Lee, "Rain removal using Kalman filter in video, Smart Manufacturing Application", in ICSMA 2008, International Conference on, IEEE, 2008, pp. 494-497.

[7] J.r.m. Bossu, N. HautiÃ¨re, J.-P. Tarel, "Rain or snow detection in image sequences through use of a histogram of orientation of streaks", International Journal of Computer Vision, Vol. 93, No. 3, Jul. 2011, pp. 348-367.

[8] M. Qi, B. Geng, J. Jiang, T. Wang, "A rain detection and removal method in video image", in Intelligent Visual Surveillance (IVS), 3rd Chinese Conference on, IEEE, 2011, pp. 1-4.

[9] A.K. Tripathi, S. Mukhopadhyay, "Removal of rain from videos: a review", Signal, Image and Video Processing, Vol. 8, No. 8, Sep. 2012, pp. 1421-1430.

[10] L.-W. Kang, C.-W. Lin, Y.-H. Fu, "Automatic single-image-based rain streaks removal via image decomposition", Image Processing, IEEE Transactions on, Vol. 21, No. 4, Apr. 2012, pp. 1742-1755.

[11] A.K. Tripathi, S. Mukhopadhyay, "Video post processing: low-latency spatiotemporal approach for detection and removal of rain", IET image processing, Vol. 6, No. 2, Mar. 2012, pp. 181-196.

[12] X. Zhao, P. Liu, J. Liu, T. Xianglong, "The application of histogram on rain detection in video", Proceedings of the 11th Joint Conference on Information Science, 2008.

[13] K.V. Beard, C. Chuang, "A new model for the equilibrium shape of raindrops, Journal of the Atmospheric sciences", Vol. 44, No. 11, Jun. 1987, pp. 1509-1524.

[14] G.B. Foote, P.S. Du Toit, "Terminal velocity of raindrops aloft", Journal of Applied Meteorology, Vol. 8, No. 2, May 1969, pp. 249-253.

[15] J.S. Marshall, W.M.K. Palmer, "The distribution of raindrops with size", Journal of Meteorology, Vol. 5, Aug. 1948, pp. 165-166.

[16] V.S.N. Prasad, J. Domke, "Gabor filter visualization", Technical Report, University of Maryland, 2005.

[17] J.R. Movellan, "Tutorial on Gabor filters", Open Source Document 2002.

[18] M. Bertalmio, A.L. Bertozzi, G. Sapiro, Navier-stokes, "fluid dynamics, and image and video inpainting", Computer Vision and Pattern Recognition, Proceedings of the 2001 IEEE Computer Society Conference on, IEEE, 2001, Vol. 351, pp. I-355-I-362.

[19] L. Liang, C. Liu, Y.-Q. Xu, B. Guo, H.-Y. Shum, "Real-time texture synthesis by patch-based sampling", ACM Transactions on Graphics, Vol. 20, No. 3, Jul. 2001, pp. 127-150.

[20] G. Malekshahi, H. Ebrahimnezhad, "Detection and removal of rain in video sequence using Gabor filter", in 21th Iranian Conference on Electrical Engineering (ICEE2013), May. 2013 (printed in Persian).

Gelareh Malekshahi received the B.Sc. degree in Electrical Engineering from Karaj Azad University Iran, in 2009 and the M.Sc. degrees in Electrical Engineering from Sahand University of Technology, Iran, in 2013, respectively. Her research interests are, machine vision, image and video processing, and pattern recognition.

Hossein Ebrahimnezhad received the B.Sc. and M.Sc. degrees in Electronic and Communication Engineering from Tabriz University and K.N.Toosi University of Technology in 1994 and 1996, respectively. In 2007, he received the Ph.D. degree from Tarbiat Modares University. His research interests include image and multimedia processing, computer vision, 3D model processing and soft computing. Currently, he is associate professor at Sahand University of Technology.

SRR Shape Dual Band CPW-fed Monopole Antenna for WiMAX / WLAN Applications

Zahra Mansouri*
Department of Electrical engineering, Khodabandeh Branch, Islamic Azad University, Khodabandeh,Iran
zm.mansouri@gmail.com

Ramezan Ali Sadeghzadeh
Department of Electrical and Computer Engineering, K.N Toosi University of Technology, Tehran, Iran
sadeghz@eetd.kntu.ac.ir

Maryam Rahimi
Department of Electrical engineering, Imam Khomeini International University, Qazvin, Iran
maryam.rahimi05@gmail.com

Ferdows B. Zarrabi
Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran
ferdows.zarrabi@modares.ac.ir

Received: 30/Jun/2014 Revised: 04/Sep/2014 Accepted: 07/Oct/2014

Abstract
CPW structure is became common structure for UWB and multi band antenna design and SRR structure is well-known kind of metamaterial that has been used in antenna and filter design for multi band application. In this paper, a SRR dual band monopole antenna with CPW-fed for WLAN and WiMAX is presented. The prototype antenna is designed for wireless communication such as WLAN and WIMAX respectively at 2.4 GHz and 5 GHz. The HFSS and CST microwave studio are used to simulate the prototype antenna for two different FEM and time domain method and they have also been compared with the experimental results. The total size of the antenna is 60mm×55mm×1.6mm and it is fabricated on FR-4 low cost substrate. The antenna is connected to a 50 Ω CPW feed line. Its bandwidth is around 3% for 2.45 GHz (2.4-2.5 GHz) and 33% for 5.15GHz (4.3-6 GHz).Its limited bandwidth in 2.4 GHz frequency is benefit for power saving at indoor application. The antenna has 2-7 dBi gain in the mentioned bands with an Omni-directional pattern. The antenna experimental result shows good similarity to simulation kind for return loss and pattern. Here, the effect of parasitic SRR on current distribution has been studied in presence and absence of parasitic element. The simulation of polarization is confirmed that the antenna has linear polarization. Here comparison between antenna return losses in absence of each parasitic element is presented.

Keywords: CPW-fed; SRR; Dual Band Antenna; WLAN.

1. Introduction

Wireless communication systems have progressed too fast in the last decade. It is widely used in notebooks and cellular phone because of mobility and low cost and These days WLAN systems due to reasonable prices and high-speed data transfers are widely used [1-2]. A WLAN links two or more devices, by providing a high speed connection through an access point to the wider internet. This gives users the ability to move around in a local coverage area and still be connected to the network. Nowadays, broadband systems have been designed for faster communication and high data transfer rate [3]. IEEE 802.11 is a standard for WLAN system. IEEE 802.11a standard considers 5.15-5.35 GHz and 5.725-5.825 GHz as sending and receiving band respectively. The IEEE 802.11bg is applied for 2.4GHz (2.4– 2.484 GHz) applications. The frequency range 3.5 GHz (3-5 GHz) is used for WiMAX applications [4]. It is need to design small size, easy fed and low cost antenna for multi band applications. CPW fed antenna is the common type for UWB applications with all properties that are needed for WLAN communication systems and circular arc is used to increase the bandwidth in antenna at WLAN frequency [5]. Microstrip compact antennas can be used in mobile communication and WLAN systems because of their benefits. Different methods have been used to design multi band antenna, such as notch technique, metamaterial, CRLH or ZOR (zeroth order resonator),slot and fractal methods like Minkowski, Hilbert, Koch, Sierpinski, tree [6-8]. The bandwidth of microstrip antennas is low and it is one of the main drawbacks of them. Different ways are used for improving its bandwidth such as increasing the substrate thickness, using substrate with low permittivity, proper feeding techniques for better impedance matching, multi resonance technique, slots on the antenna and inserting parasitic element in antenna geometry [9-10].

*Corresponding Author

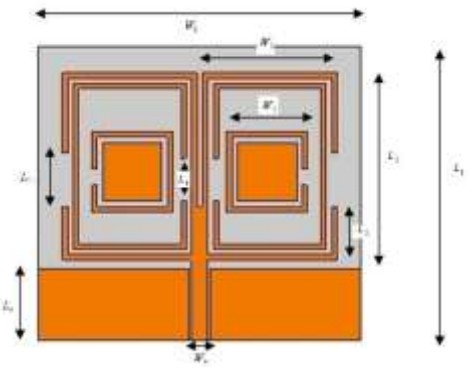
Nowadays, CPW-feed technique has become one of the most popular methods in feeding micros trip antennas for wide band application. Low radiation loss, low leakage, wide band width, improved impedance matching and easy integration with RF circuit are known as benefits for CPW-feed antenna [11-12].The high electromagnetic coupling between the patch and the parasitic element improves impedance bandwidth and miniaturizes the antenna. Parasitic elements resonate near the patch resonance frequency and leads to higher bandwidth. Bandwidth can be controlled by adjusting the distance between parasitic element and the patch .Different model of parasitic element are used in antenna such as U-Shaped parasitic elements or Split ring resonator (SRR) element [13-15]. Parasitic ring element has been employed in slot antenna with microstrip feed for dual band compact antenna [16].Negative index metamaterial are known as composite materials that show negative effective permittivity and negative effective permeability and in those material Negative permeability was obtaining by structure such as SRRs, spiral resonators (SR), V-shaped resonators [17] Metamaterials are artificial structures that have been known with unusual properties such as anti parallel phase and group velocities, and negative reflection index. In Metamaterials relationship between electric field, magnetic field and wave vector follows left-hand rule, so they are called left-handed materials (LHM). They show negative permittivity and permeability. When both ε and μ are negative in a structure, we call it double negative (DNG). Metamaterials make it possible to design a miniaturized, multi band antenna. Split ring resonator (SRR) is a type of metamaterial which is used in microwave devices or absorber to improve the bandwidth and the gain of the antenna. It can produce the negative permeability and positive permittivity around the resonance frequency. The main feature of SRR is the quasi-static resonance at a larger wavelength in spite of its own size. This leads to use of SRR in designing small antenna [18-19]. In this paper, a SRR structure which is most famous in metamaterial is presented. Dual band Microstrip antenna is designed for wireless applications with Omni directional pattern. The effect of parasitic SRR on current distribution has been studied. The simulation results are compared with experimental results for VSWR and radiation pattern. The SRR parasitic structure is used for antenna miniaturization. The antenna experimental result shows good similarity to simulation kind for return loss and pattern. The simulation of polarization is confirmed that the antenna has linear polarization. Here comparison between antenna return loss in absence of each parasitic elements is presented.

2. Antenna Design

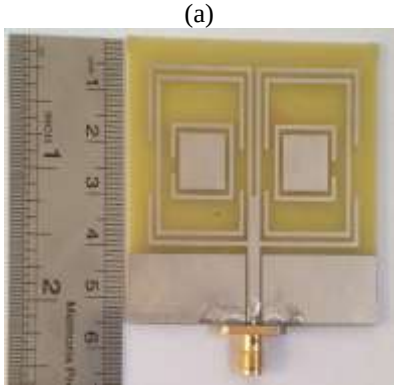
CPW-fed antenna is the common model antenna for UWB and multi band applications [20]. Also SRR structures are used for design multiband and small antenna. Our proposed antenna is a dual band CPW-feed monopole antenna with SRR structure for using in WLAN and WIMAX applications. The effect of parasitic element is studied here. Fig. 1(a) and (b) shows the geometry and fabricated antenna on FR-4 layer respectively. The antenna is fabricated on FR-4 low cost substrate with dielectric constant ε= 4.4 and loss tangent of tan δ=0.02. The substrate height and the substrate dimension are h=1.6 mm and 60mm ×55mm, respectively. The antenna is connected to SMA with CPW 50 Ω feed line. All gaps width of the antenna are assumed 1mm and the gap between CPW grounds and radiation elements are 1.6 mm. The inner square length is 5mm and other dimensions of the antenna are as shows in Table 1:

Table 1. The geometrical parameter of antenna

parameter	mm
L_1	55
L_2	35
L_3	10
L_4	8
L_5	10
L_6	13.4
W_1	60
W_2	24
W_3	15
W_4	3



(a)



(b)

Fig.1 a) Antenna geometry and b) fabricated antenna

3. Simulation Result

The final model of antenna is simulated with HFSS and CST for two different full wave simulation methods. The fabricated final antenna is measured by Agilent 8720ES. The simulated and measured return losses are shown and compared in Fig.2. The presented antenna operates at two bands at 2.45 and 5.15 GHz with return

loss less than -10dB as shows in Fig.2. The first band is occurred from 2.40 to 2.5GHz which allotted for Wi-Fi and WLAN-2.4 application. The second band happened from 4.3 to 6 GHz which are used in WLAN and WiMAX applications such as 5.15-5.35 GHz and 5.725-5.825 GHz as sending and receiving band in WLAN. Its bandwidth is around 3% for 2.45 GHz (2.4-2.5 GHz) and 33% for 5.15GHz (4.3-6 GHz)

Fig. 3 shows the comparison between the effects of parasitic elements at antenna return loss. Fig 3.a, b, c, d shows 4 step of antenna designing. The final antenna is presented in Fig.3.d and this antenna result shows in Fig.2. As shown in Fig. 3 the resonance frequency decreased as the number of parasitic elements increases. Small parasitic element decreases the first resonance to 2.45 GHz. In the first antenna, First resonance is occurred at 2.55-2.675 GHz and second band is at 4.35-6 GHz but at this frequency range from 4.75-5.55 GHz the return loss is increased to -8 dB .In second antenna the first resonance is shifted to higher frequency and it is happened at 3.075-3.2 GHz but second resonance at 4.3-6 GHz with sufficient return loss. At third model antenna the first resonance is matched with our request

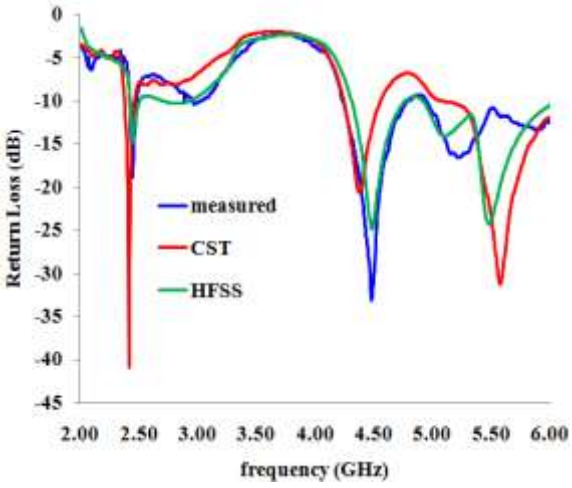
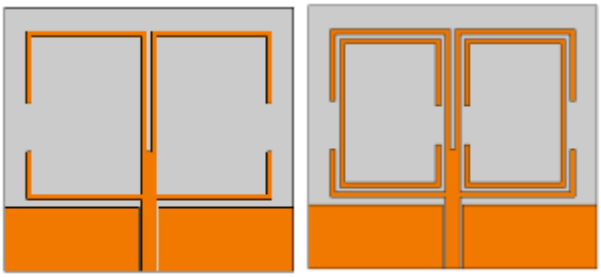
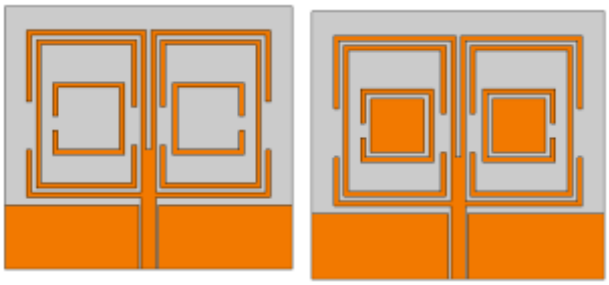


Fig.2 Comparisons of return losses among CST, HFSS and experimental results.

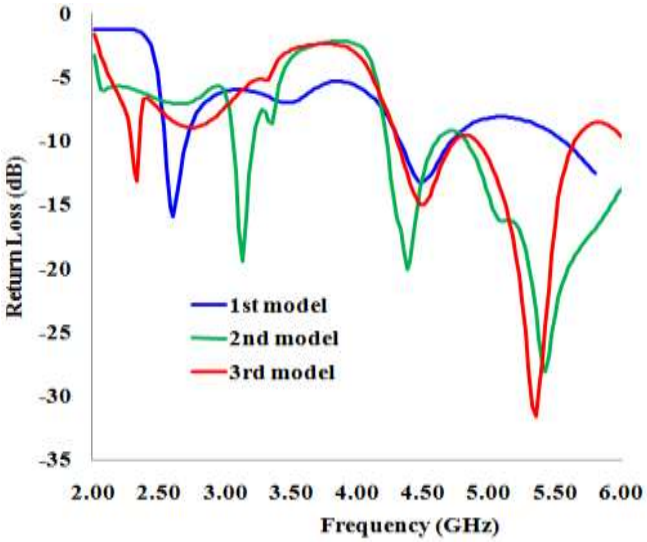
band at 2.45 GHz and second resonance is occurred in range of 4.3-5.6.25 GHz so it cannot cover the receiving band for WLAN at 5.725-5.825 GHz so with adding last parasitic element as shows in Fig.3.d the final antenna is achieved with good performance at both band.



(a) (b)



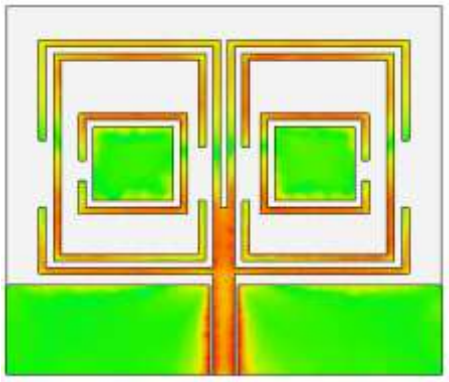
(c) (d)



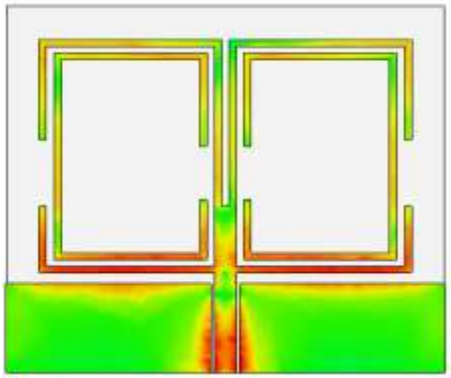
(e)

Fig.3 The prototype antenna return loss in comparison with the antenna without parasitic element and all simulated antenna a)1st model b)2nd model c)3rd model d) final model e) first to third antenna return loss simulation in HFSS

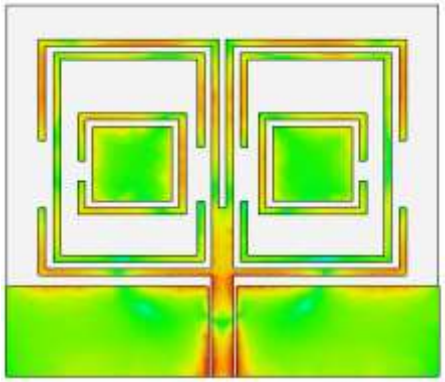
Fig.4 shows the current distribution comparison for final antenna and second model at 2.45 and 5.15 GHz. As shown in Fig.4 (a) the inner small ring helps to increase the antenna effective length based on coupling effect between bottom and top part of the antenna. Based on Fig.4 (a), the current distribution enhance in the outer ring in comparison with second model that given in Fig.4.b. As shows in Fig.4 (b) the antenna current is concentrated at the bottom of the radiator so the antenna effective length is reduced and the antenna first resonance is shifted to higher frequency. Fig.4 (c) and Fig.4 (d) shows the inner parasitic element has less effect at 5.15 GHz and in both structure the antenna has similar current distribution form.



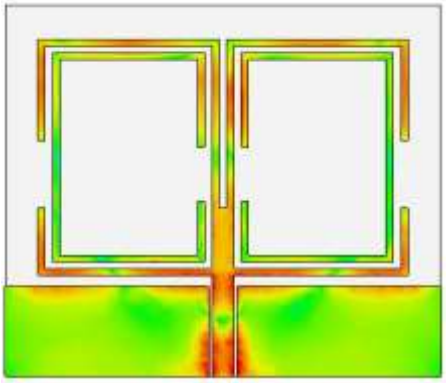
(a)



(b)

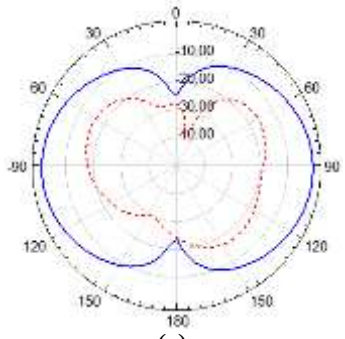


(c)

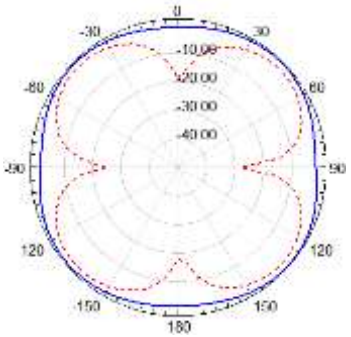


(d)

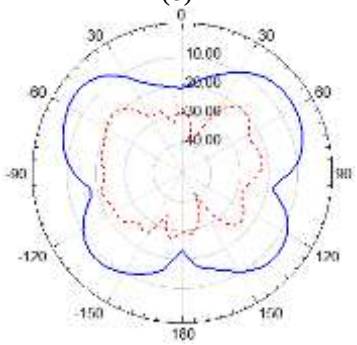
Fig.4 The current distribution comparison between prototype antenna and second antenna at a) final antenna at 2.45 GHz, b) second antenna at 2.45 GHz, c) final antenna at 5.15 GHz d) second antenna at 5.15 GHz



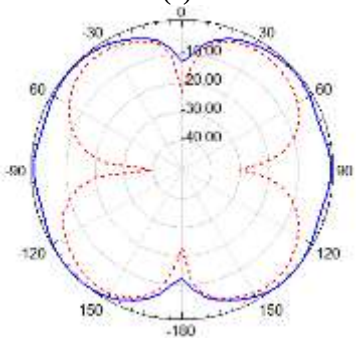
(a)



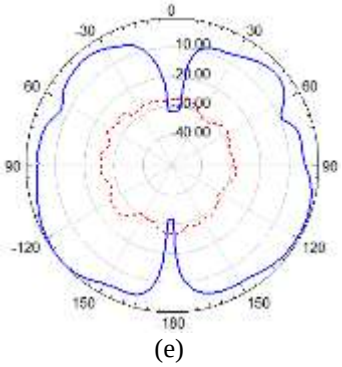
(b)



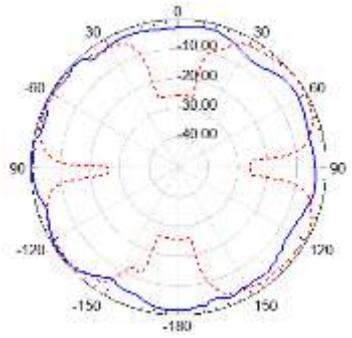
(c)



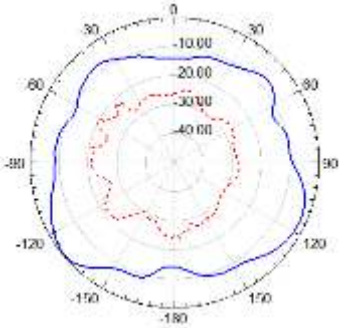
(d)



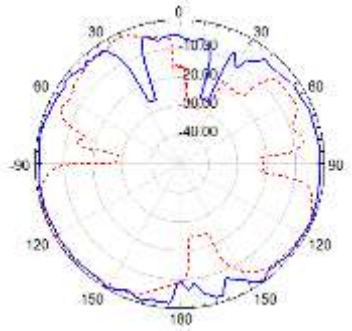
(e)



(f)



(g)



(h)

Fig.5 The prototype antenna pattern in a) E-Plane at 2.45 GHz simulation b) H-Plane at 2.45 GHz simulation c) E-Plane at 5.3 GHz simulation d) H-Plane at 5.3 GHz simulation e) E-Plane at 2.45 GHz measured f) H-Plane at 2.45 GHz measured g) E-Plane at 5.3 GHz measured h) H-Plane at 5.3 GHz measured

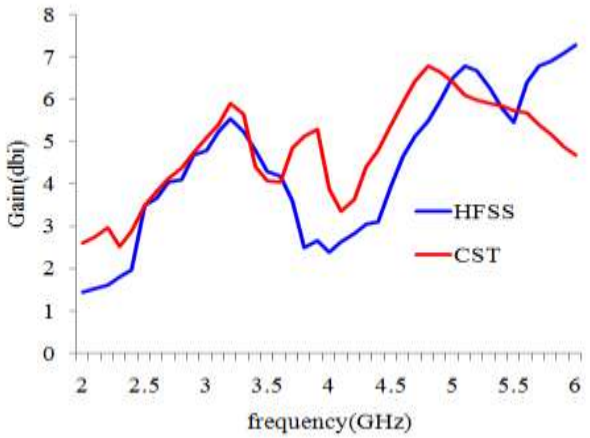


Fig .6 the antenna gain simulation

The measured and simulation radiation patterns (co-polarization and cross-polarization (dash-line)) in E-plane (*y-z* plane) and H-plane (*x-z* plane) at 2.45 and 5.3 GHz of the prototype antenna are shown in Fig. 5. The antenna radiates Omni-directionally in the E-plane and approximately bi-directionally in the H-plane. Obviously, the antenna shows good performances over the frequency range.

Fig. 6 presents gain of the antenna which is between 2 and 7dBi and HFSS result is compared with CST microwave studio.

The axial ratio is used to measure the quality of the field polarization, as shows in (1). AR is the ratio of major and minor axes of the polarization along the bore-sight for the proposed antenna, derived from the measured results on the fabricated antenna.

$$(1)\ AR[dB] = 20\log\frac{E_{\max}}{E_{\min}}$$

So when axial ratio is more than 3 the antenna polarization is linear. Axial ratio that is shown in fig 7, confirmed the vertical linear polarization of antenna.

The simulated efficiency is more than 80% for first band and more than 65% for second band. Fig 8 shows the efficiency of the antenna. Table 2 shows comparison between prototype antenna and some previous research [21-22] for bandwidth, size and gain. The prototype antenna shows good quality and has mediocre gain or bandwidth.

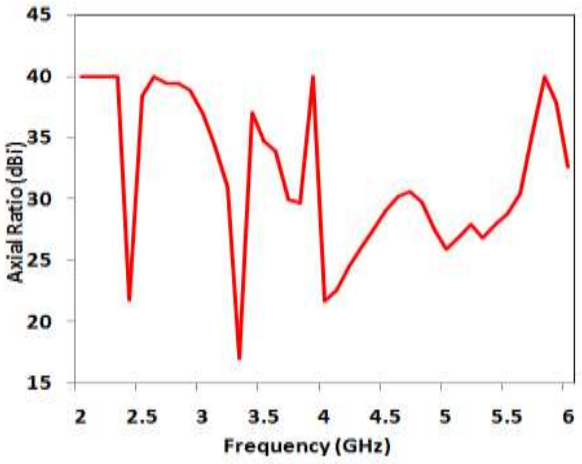


Fig7. Axial ratio of the proposed antenna

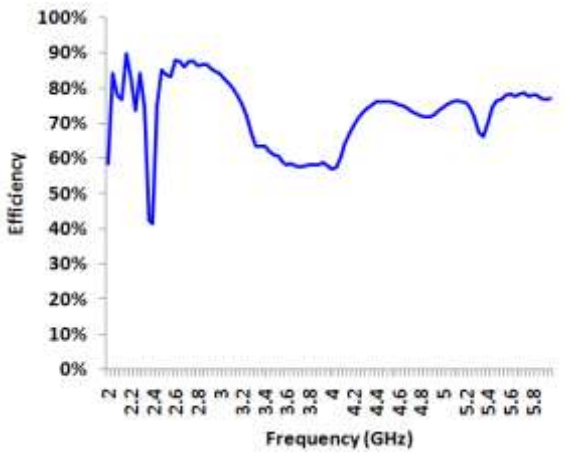


Fig 8. Efficiency of the proposed antenna

Table .2 comparisons between prototype antenna and some previous research

	Our purpose	Ref 20	Ref 21	Ref 22
First Band (GHz)	2.4-2.5	2.32-3.3	2.45	1.8-2.53
Second Band(GHz)	4.3-6	4.96-6.34	5.85	2.9-6.04
Gain (dBi)	2-7	2.4-4.5	-1.7-5.2	3.4-9.2
Size (mm)	60×55	69.3×68	30×30	52×54
First Bandwidth	3%	35.8 %	1.2%	33.7%
Second Bandwidth	33%	24.4 %	12.24%	20%

4. Conclusions

Here presented an antenna with limited bandwidth in comparison to conventional UWB CPW antenna so it is useful for power saving at indoor application and also antenna has Omni directional pattern such as other CPW antenna. In other hand SRR parasitic elements are used to reduce the antenna resonance to 2.4 GHz**.** The bandwidths are around 3% for 2.45 GHz (2400-2500 MHz) and 33% for 5.15GHz (4300-6000MHz). The antenna has 2-7 dBi gain in the operation band. The final model of antenna is simulated by HFSS and CST microwave studio with FEM and time domain full wave methods. The results are compared with experimental**.** The simulation of polarization confirmed that the antenna has linear polarization. The comparison between antennas in absence of each parasitic element is presented. The final antenna shows Omni directional pattern with sufficient gain for wireless applications

References

[1] M. Rahimi, F. B. Zarrabi, R. Ahmadian, Z. Mansouri, and A.Keshtkar."Miniaturization of Antenna for Wireless Application with Difference Metamaterial Structures" Progress In Electromagnetic Research, vol.145, 2014,pp 19-29.

[2] X. Song, " Small CPW- fed triple band microstrip monopole antenna for WLAN applications," Microwave and Optical Technology Letters, vol. 51, no. 3,2009,pp 747-749.

[3] M. Rahimi, R. A. Sadeghzadeh, F. B. Zarrabi and Z.Mansouri "Band-Notched UWB Monopole Antenna Design with Novel Feed for Taper Rectangular Radiating Patch." Progress In Electromagnetics Research C, vol. 47, 2014,pp 147-155.

[4] B. Boroomandisorkhabi, R. A. Sadeghzadeh, F. B. Zarrabi and E. Ghahremani, "A novel UWB circular CPW antenna with triple notch band characteristics," IEEE Conference (LAPC) In Antennas and Propagation, 2013Loughborough, pp. 637-640.

[5] M. Chongcheawchamnan, K. Meelarpkit, S. Julrat, C. Phongchareonpanich and M. Krairiksh, " Extending bandwidth of a CPW- fed monopole antenna using circular arc structure," Microwave and Optical Technology Letters, vol. 54, no. 6, 2012, pp 1412-1415.

[6] D. Li and J.Mao , "A Koch-like sided fractal bow-tie dipole antenna," IEEE Transactions on Antennas and Propagation, vol. 60, no. 5, 2012, pp 2242-2251.

[7] Y.H. Ryu, J.H. Park, J.H. Lee, and H.S. Tae. "Multiband antenna using+ 1,− 1, and 0 resonant mode of DGS dual composite right/left handed transmission line." Microwave and Optical Technology Letters, vol. 51, no. 10, 2009, pp2485-2488.

[8] A. Jafargholi, M. Kamyab, M. Veysi and M. NikfalAzar "Microstrip gap proximity fed-patch antennas, analysis and design" AEU-International Journal of Electronics and Communications, vol. 66, no. 2, 2012, pp 115-121.

[9] C.H. Chen and E. K. N. Yung, "Dual-Band Circularly-Polarized CPW-Fed Slot Antenna with a Small Frequency Ratio and Wide Bandwidths," IEEE Transaction on Antennas and Propagation, vol. 59, no. 4, 2011, pp 1379-1384.

[10] D. M. Pozar, "Microstrip antennas" Proceedings of the IEEE, vol. 80, no. 1, 79-91, 1992.

[11] J.S. Chen "Dual-frequency annular-ring slot antennas fed by CPW feed and microstrip line feed," IEEE Transactions on Antennas and Propagation, vol. 53, no. 1, 2005, pp 569-573.

[12] D.B. Lin, I. Tang and Y.J. Wei, " Compact dual- band- notched CPW- fed wide- slot antenna for WLAN and WiMAX applications," Microwave and Optical Technology Letters, vol. 53, no. 7, 2011,pp 1496-1501.

[13] C. Wood, "Improved bandwidth of microstrip antennas using parasitic elements," In IEE Proceedings H (Microwaves, Optics and Antennas), vol. 127, no. 4, 1980, pp 231-234.

[14] J. Dong, A. Wang, P. Wang and Y. Hou "A Novel Stacked Wideband Microstrip Patch Antenna with U-Shaped Parasitic Elements" Antennas, Propagation and EM Theory, November 2008, pp185–188.

[15] J.- G. Lee and J.H. Lee, "Suppression of spurious radiations of patch antennas using split- ring resonators (SRRs)," Microwave and optical technology letters, vol. 48, no. 2, 2006, pp 283-287.

[16] S. Gai, Y.-C. Jiao, Y.-B. Yang, C.-Y. Li, and J.-G. Gong , "Design of a novel microstrip-fed dual-band slot antenna for WLAN applications," Progress In Electromagnetics Research Letters ,Vol. 13 ,2010,pp75-81.

[17] E. Ekmekci, K. Topalli, T. Akin, and G. Turhan-Sayan, " A feasibility study for tunable metamaterial design using multi-Split SRR structures and RF MEMS switching," IEEE Society International Symposium In Antennas and Propagation , 2009, pp. 1-4.

[18] F. B. Zarrabi, S. Sharma , Z. Mansouri, and F. Geran, " Triple Band Microstrip Slot Antenna for WIMAX/WLAN Applications with SRR Shape Ring," Fourth International Conference on In Advanced Computing & Communication Technologies (ACCT), 2014, pp. 368-371

[19] A. Sharma, S. K. Gupta, B. K. Kanaujia, G. P. Pandey and M. Sharma, "Design and analysis of a meandered multiband antenna based on split ring resonator," Microwave and Optical Technology Letters, vol. 55, no. 11, 2013,pp. 2787-2795.

[20] W-C. Liu, and C-M. Wu "Broadband dual-frequency CPW-fed planar monopole antenna with rectangular notch" Electronics Letters, Vol.40, no. 11, 2004, pp. 642-643.

[21] H.H. Li, , X.Q. Mou , Z. Ji, H. Yu, Y. Li, and L. Jiang, "Miniature RFID tri-band CPW-fed antenna optimized using ISPO algorithm," Electronics Letters ,Vol. 47, no. 3 ,2011, pp.161-162.

[22] C. Wang, Z-H. Yan, P. Xu, J-B. Jiang, and B. Li "Trident-shaped dual-band CPW-fed monopole antenna for PCS/WLAN applications" Electronics letters, Vol.47, no. 4, 2011, pp.231-232.

Zahra Mansouri was born in Zanjan, Iran in 1987. She received her B.Sc. degree in Electrical Engineering (Telecommunication) from Zanjan University, Zanjan, Iran, in 2008 and M.Sc. degree in Electrical Engineering (Telecommunication) from Islamic Azad University, Science and Research Branch, Tehran, Iran, in 2012. She is now PHD student in Science and Research Branch, Islamic Azad University, Tehran, Iran. Her primary research interests are in microwave components such as couplers, power dividers and metamaterials and also UWB antenna.

Ramezan Ali Sadeghzadeh received the B.Sc. degree in telecommunication engineering from K. N. Toosi University of Technology, Tehran, Iran, in 1984, the M.Sc. in digital communication engineering from the University of Bradford, Bradford, U.K., and University of Manchester Institute of Science and Technology (UMIST), Manchester, U.K., as a joint program in 1987, and the Ph.D. degree in electromagnetic and antenna from the University of Bradford in 1991. During 1992 to 1997, he worked as a Postdoctoral Research Assistant in the field of propagation, electromagnetic, antenna, biomedical, and wireless communication with the University of Bradford. From 1984 to 1985, he was with Iran Telecommunication Company, Tehran, Iran, working on networking. Since 1997, he has been with the Faculty of Electrical and Computer Engineering, K. N. Toosi University of Technology. He has published more than 120 referable papers in international journals and conferences. His research interests include numerical techniques in Electromagnetics, antenna, propagation, radio networks, wireless communications, nano antennas, and radar systems.

Maryam Rahimi was born in Tehran, Iran, in 1987. She received the B.Sc. degree in electrical engineering from Islamic Azad University, Qazvin, Iran, with honor the M.Sc. degree with honor in electrical engineering from the University of Imam Khomeini international University, Qazvin ,Iran, in 2013, Her major interest is designing of antenna for wireless and

Ferdows B. Zarrabi was born in Iran, Babol in 1983. He studied the electrical engineering at University of Tabriz in major of communication engineering in 2008. His major interest is designing of antenna for wireless and UWB application for breast cancer detection radar; microwave Devices, Absorber, Metamaterial, Plasmonic, Nano antenna and THz antenna. He author and co author of more than 13 referee paper.

A New Robust Digital Image Watermarking Algorithm Based on LWT-SVD and Fractal Images

Kayvan Ghaderi
Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran
keyvan.ghaderi@uok.ac.ir

Fardin Akhlaghian Tab*
Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran
f.akhlaghian@uok.ac.ir

Parham Moradi
Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran
p.moradi@uok.ac.ir

Received: 30/Jun/2014 Revised: 04/Sep/2014 Accepted: 07/Oct/2014

Abstract

This paper presents a robust copyright protection scheme based on Lifting Wavelet Transform (LWT) and Singular Value Decomposition (SVD). We have used fractal decoding to make a very compact representation of watermark image. The fractal code is presented by a binary image. In the embedding phase of watermarking scheme, at first, we perform decomposing of the host image with 2D-LWT transform, then SVD is applied to sub-bands of the transformed image, and then the watermark, "binary image," is embedded by modifying the singular values. In the watermark extraction phase, after the reverse steps are applied, the embedded binary image and consequently the fractal code are extracted from the watermarked image. The original watermark image is rendered by running the code. To verify the validity of the proposed watermarking scheme, several experiments are carried out and the results are compared with the results of the other algorithms. In order to evaluate the quality of image, we use parameter peak value signal-to-noise ratio (PSNR). To measure the robustness of the proposed algorithm, the NC coefficient is evaluated. The experimental results indicate that, in addition to high transparency, the proposed scheme is strong enough to resist various signal processing operations, such as average filter, median filter, Jpeg compression, contrast adjustment, cropping, histogram equalization, rotation, etc.

Keywords: Image Watermarking; Lifting Wavelet Transforms; Singular Value Decomposition; Fractal Image.

1. Introduction

During the last decade, the availability of information in digital form has increased rapidly. The success of the Internet and cost-effective recording and storage devices have made it possible to easily create, replicate, transmit, and distribute digital content. However, the information security, authentication of data and protection of intellectual property rights have also become an important issue. In such a scenario, a mechanism for copyright protection of multimedia data is essential. Digital watermarking is a process that embeds ownership protection data, named watermark, into the host data. The embedded data could be a signature image, an audio or textual data. This process is highly necessary to protect digital data against unauthorized use [1, 2].

Basically, a set of requirements is evaluated for a watermarking scheme to be effective. These requirements can be categorized as follows: (1) imperceptibility, (2) robustness, (3) capacity [3].

According to operation domain, digital watermarking can be divided into two categories: spatial domain and transform domain. Early image watermarking schemes operated directly in spatial domain, which was mostly associated with poor robustness properties [4]. In contrast, watermarking in the transform domain such as discrete cosine transform (DCT), wavelet transforms (WT) and singular value decomposition (SVD) provide more advantages, and better performances will be obtained compared to those of spatial ones in most of recent research [5-8].

The watermark-extraction techniques can also be classified into non-blind, semi-blind, and blind categories. Non-blind methods need the original signal, which limits their usage since the original media are difficult to obtain sometimes. In semi-blind methods some features of the original signal are needed to be known a priori. Finally, in blind methods, there is no need for the original signal or the watermark in the watermark extraction phase [9-11].

The main motivation of this work is to provide a robust digital image watermarking.

To achieve higher imperceptibility and robustness in the watermarking algorithm, the main idea of this work is based on compact representation of watermark image using fractal encoding. In this way, watermark image is selected from the set of fractal images. In the watermark embedding stage, instead of storing the raw data of the

* Corresponding Author

original watermark image, we embed a binary image which shows the constructor code of the fractal image. In the extraction phase, the binary image is extracted from the host image. Consequently, from this retrieved binary image, the constructor fractal code is also simply extracted. Perhaps the impression is created that when we choose the watermark from the set of fractal images, a limitation is imposed on this algorithm. However, considering the philosophy and application of watermarking algorithms, it is clear that regardless of the watermark image type, the algorithm can achieve its goals such as copyright protection or proving of ownership. Therefore it can be said that selection of fractal images as watermark is not a fundamental limitation and, in favor of its significant improvement, can be entirely disregarded.

In the embedding and extraction phases of the proposed algorithms, LWT and SVD transforms are used. The host image is decomposed by K-level LWT transform, and then the binary image is inserted in the singular values of the K_{th} level of the decomposed host image. In the extraction phase, after reverse steps of the embedding stage, the inserted values (binary image) can be caught up from the watermarked image. The fractal code is normally a text, typed with a keyboard with regular fonts. This text file is then turned to a binary image for watermarking application. Therefore the embedded code can be obtained from the extracted watermark image easily by a user, or automatically by a simple OCR system. Although both methods are acceptable, however considering the focus of this paper on watermarking algorithm, we extract the fractal code by a user. Rendering the fractal code will produce the final watermark image.

The proposed algorithm achieves higher robustness and improved fidelity, which is one of the important challenges of the watermarking schemes. Since the singular values of the original image are required for extracting the watermark, the introduced algorithm is semi-blind.

The proposed watermarking algorithm is tested against different attacks such as average filter, rotation, cropping, Jpeg compression, etc. The experimental results indicate more robustness against different attacks compared to other algorithms, while significant improvement in PSNR value is also achieved.

The rest of the paper is structured as follows. In Section 2, LWT and SVD transforms are briefly described. The proposed algorithm is discussed in Section 3. The experimental results to demonstrate the performance of this scheme are described in Section 4 and finally the conclusion is drawn in Section 5.

2. Preliminaries of LWT and SVD Transforms

In this section, LWT and SVD transforms are briefly described.

2-1- Lifting wavelet transform

Let $I(m,n)$ be a 2D signal. Without loss of generality, we assume that this signal is first performed with 1D wavelet transform on the vertical direction and then on the horizontal direction. With the basic principle of lifting structure given in [12], each 1D wavelet transform can be factored into one or multiple lifting stages. A typical lifting stage consists of three steps: split, predict and update.

In the first step, all samples are split into two parts: the even poly-phase samples and the odd poly-phase samples,

$$\begin{cases} \boldsymbol{Ie(m,n) = I(2m,2n)} \\ \boldsymbol{Io(m,n) = I(2m,2n+1)} \end{cases} \quad (1)$$

In the predict step, the odd poly-phase samples are predicted from the neighboring even poly-phase samples. In the conventional lifting, the predictions always come from the vertical neighboring even poly-phase samples.

LWT which is the second generation fast wavelet transform is a substitute method for DWT to transform images into the transform domain for real time applications. In lifting wavelet transformation, up sampling and down sampling are replaced simply by split and merge in each of the levels. Split and merge process in LWT reduces computational complexity to 50%. Information loss is less as compared to DWT algorithm, because in LWT based algorithm up sampling and down sampling have not been used. The odd poly-phase and even poly-phase components of the signal are filtered in a specific parallel process by using the corresponding wavelet filter coefficients, producing the better result compared to up sampling and down sampling which is required in the traditional DWT approach. In comparison with general wavelets, reconstruction of images by lifting wavelet is a good idea because it increases smoothness and reduces aliasing effects [13]. Employing LWT reduces information loss, increases intactness of embedded watermark in the image and helps to increase the robustness of watermark. Lifting wavelet transform also provides several advantages [14, 15] such as less memory requirements, reduced distortion and aliasing effects, good reconstruction, less computation and computational complexities. In this decomposition, filter coefficients are converted into lifting coefficients (predict s(z), update t(z) and scaling (k)) using Euclidian algorithm, and the original image is split into (odd and even) sets of samples. Further lifting coefficients are applied to the sampled original image to get approximate and detailed sub bands (Fig. 1).

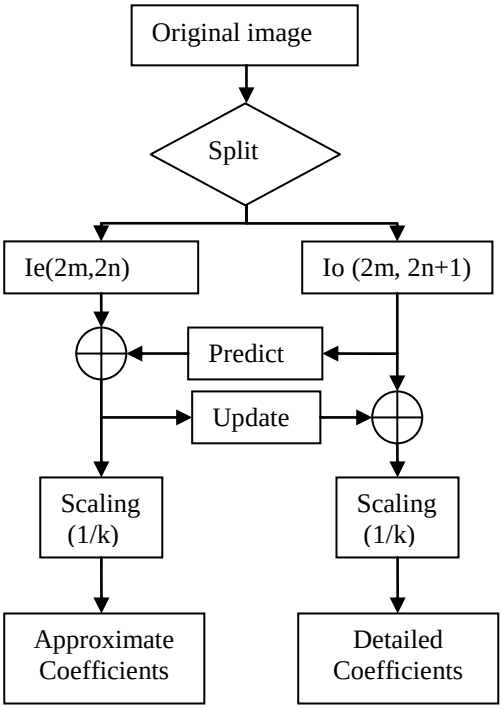


Fig.1. Applying LWT transform to the image to obtain coefficients.

2-2- Singular value decomposition

Let A be a general real (complex) matrix of order m×n. The singular value decomposition is the following factorization

$$(2)\ A = U \times S \times V^T$$

where U and V are orthogonal (unitary) and S = diag $(\sigma_1, \sigma_2, \ldots, \sigma_r)$, where $\sigma_i, i = 1, \ldots, r$ are the singular values of the matrix A with r = min (m, n) and satisfying:

$$(3)\ \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$$

Use of SVD in digital image processing has some advantages which are listed as follows:

1. The size of the matrices from SVD transformation should not necessarily square and can be a rectangle.
2. Singular values in a digital image are less affected if general image processing is performed. It means that for a small perturbation added to an image, its SVs do not change fast.
3. Singular values contain intrinsic algebraic image properties, where singular values correspond to the brightness of the image and singular vectors reflect geometry characteristics of the image. [16-18].

SVD can effectively reveal essential properties of image matrices, so it has been used in a variety of image processing applications such as noise estimation and digital watermarking [19-21].

3. Proposed Watermarking Scheme

This section is divided into three parts as follows: (1) watermark generating, (2) embedding stage and (3) extraction stage.

3-1- Watermark generation using fractal images

At first, considering properties of fractal images, the mathematical equation and then the related fractal based code for producing the selected watermark image is determined. Thus, instead of using the original watermark image, a small binary image which shows the fractal code of the watermark image is used in the watermarking scheme. This stage is shown in TABLE 1. Next, the binary image is called as the substitute watermark image. As can be seen in the aforementioned table, the code section is divided to two parts: Constant and Main. The Constant part is used in the rendered original watermark stage as a key extraction. Also, the Main part is used after converting to a binary image, as the substitute for the original watermark.

In this way, since the volume of inserted information in the host image is much less than the original watermark, and also due to function of the fractal code in producing the original watermark image, watermarking criteria such as transparency and robustness are very well satisfied.

3-2- Embedding stage

The watermark embedding procedure has been represented in Fig. 2, followed by a detailed explanation:

1. Perform K [1] -level 2D-LWT on the host image to provide multi-resolution sub bands: LL_k, LH_k, HL_k and HH_k.
2. Apply SVD transform to sub bands: $LH_1, HL_1, \ldots, LH_k$, and HL_k to get U, V and S matrices at each sub band.

$$I = U_i \times S_i \times V_i^T \; i = 1, 2, \ldots, k \qquad (4)$$

3. Modify the singular values of the host image in LH_i and HL_i sub-bands according to those of the watermark.

$$S_{modified} = S + (\alpha \times watermark) \qquad (5)$$

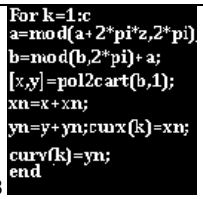
Where α represents the scaling factor.

4. Perform inverse SVD with updated S matrix.
5. Apply K-level 2D-ILWT to obtain watermarked image.

[1] In the proposed algorithm, selection of the number of 2D-DWT level depends on parameters such as size of the host and watermark images, number of sub bands, etc.

Table 1 . Watermark construction process.

<table>
<tr><th>Code</th><th colspan="2">Subject</th></tr>
<tr><td rowspan="2">z=log(2)*0.0185;
a=0;
b=0;
c=15000;
curx=zeros(1,c);
cury=zeros(1,c);
xn=0;
yn=0;

For k=1:c
a=mod(a+2*pi*z,2*pi);
b=mod(b,2*pi)+a;
[x,y]=pol2cart(b,1);
xn=x+xn;
yn=y+yn;
curx(k)=xn;
cury(k)=yn;
end
line (curx,cury ,1 ,'k')</td><td>Constant</td><td>As a reconstruction key</td></tr>
<tr><td>Main</td><td>Original watermark
Size=1024×1024
  </td></tr>
<tr><td colspan="3">Converted code to image as a substitute watermark</td></tr>
<tr><td colspan="3">Size=128×128
  </td></tr>
</table>

3-3- Extraction stage

Our aim in watermark extraction is to obtain embedded substitute watermark. To reconstruct the original watermark image, in the first step, this fractal code and reconstruction key are combined together, so this new code is run. The extraction procedure is explained as follows.

1. Perform K-level 2D-ILWT on the watermarked image to provide multi-resolution sub bands: LL_k, LH_k, HL_k and HH_k.
2. Apply SVD transform to sub bands LH_1, HL_1, …, LH_k and HL_k to get U, V and S matrices.

$$I^* = U_i^* \times S_i^* \times (V_i^*)^T, i = 1,2,...,k \quad (6)$$

3. Extract watermark from S matrices by the following equation:

(7)

$$watermark = (D^* - S) / \alpha \quad (8)$$

4. Experimental Results

In simulation, we use the images "Lena" and “Pepper” whose size is 512×512 pixels as the original image and the embedded binary image, which shows the fractal code, with size

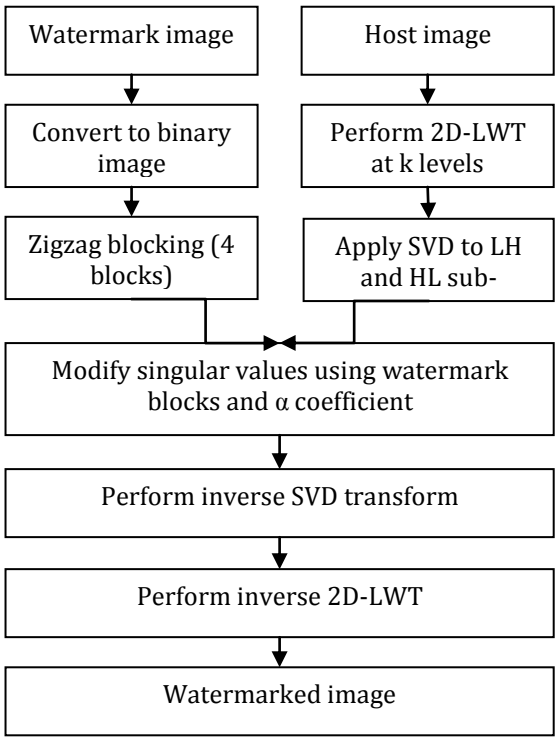


Fig. 2. Embedding stage of watermarking scheme.

128×128 Pixels as the substitute watermark image. In Fig. 3 these images are shown. Considering the size of images, the number of decomposition levels (named*k*) in these experiments is set to 4.

In order to evaluate the transparency of watermarked image, we use the criterion peak value signal-to-noise ratio (PSNR). PSNR is used as an efficient measure of visual fidelity between the host image and the watermarked image. The PSNR in decibels is given by the following equation:

$$PSNR = 10\log_{10}(225^2 / MSE) \quad (9)$$

where

$$MSE = \frac{\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\left[I(i,j) - I'(i,j)\right]^2}{N_1 \times N_2} \quad (10)$$

Where $N_1 \times N_2$ is the size of image, I and I' are the pixel gray values of the host image and the watermarked image respectively. Since the higher value of PSNR presents better transparency, it is desired.

The similarity (evaluate the robustness) between W (the original watermark) and W* (the extracted watermark) can be measured by means of normalized correlation. The normalized correlation is defined as:

$$(11)\ Corr(w, w^*) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(W_{ij} - \bar{W})(W_{ij}^* - \bar{W})}{\sqrt{\sum_{j=1}^{N}(W_{ij} - \bar{W})^2}\sqrt{\sum_{j=1}^{N}(W_{ij}^* - \bar{W}^*)^2}}$$

where $\bar{W} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}$ *and*

$$\bar{W}^* = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}^*$$

```
For k=1:c
a=mod(a+2*pi*z,2*pi)
b=mod(b,2*pi)+a;
[x,y]=pol2cart(b,1);
xn=x+xn;
yn=y+yn;cuwx(k)=xn;
curv(k)=yn;
end
```

(a) (b)

Fig. 3. (a) Original image and (b) Substitute watermark image.

Fig. 4 shows the watermarked image and the extracted binary watermark image without any attacks. The PSNR between embedded watermark image and original image is 71.382 db. The value of α in this paper for trade-off between transparency and robustness is set to 0.07. Fig. 5 shows a relation between α and transparency in terms of the PSNR value. In this table the abbreviation W.F means Fractal watermark.

In order to evaluate the robustness of watermarking algorithm, the watermarked image is attacked by several types of attacks and then correlation coefficients between original watermark w and detected watermark w' is calculated (TABLE 2). In Figures 6 to 14, the sub images (a, b) show the attacked watermarked image with several attacks in common image processing. Similarly, in sub images (c, d) the extracted substitute watermark images, when the watermarked images are attacked with the image manipulation, are shown.

```
For k=1:c
a=mod(a+2*pi*z,2*pi)
b=mod(b,2*pi)+a;
[x,y]=pol2cart(b,1);
xn=x+xn;
yn=y+yn;cuwx(k)=xn;
curv(k)=yn;
end
```

(a) (b)

(c)

Fig. 4. (a) Watermarked image, (b) The extracted substitute watermark image and (c) Rendered original watermark.

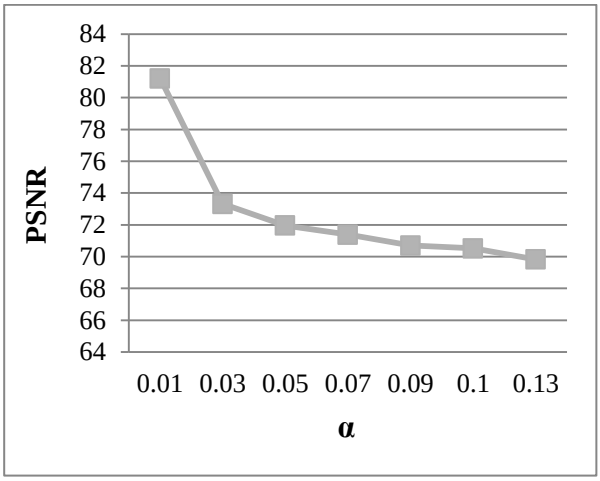


Fig. 5. Relationship between α and PSNR.

Table 2. Correlation coefficients

Attack	**Correlation Coefficients**	
	Host image: **Pepper**	Host image: **Lena**
No	0.9996	0.9998
Average Filter 9×9	0.9802	0.9820
Median Filter 9×9	0.9770	0.9467
Histogram Equalization	0.9614	0.9686
Gaussian Noise 50%	0.9632	0.9399
Contrast Adjustment	0.9831	0.9754
Cropping 50%	0.9810	0.9826
Jpeg Compression with QF=20	0.8726	0.8720
Resizing (100% → 25%)	0.9616	0.9839
Rotation 90˚	0.9840	0.9768

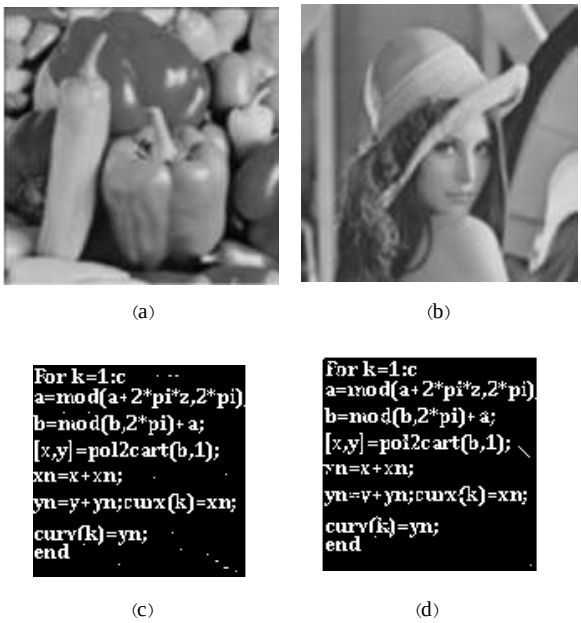


Fig. 6. (a), (b) Attacked watermarked images by Average filter 9×9; (c), (d) Extracted watermark images.

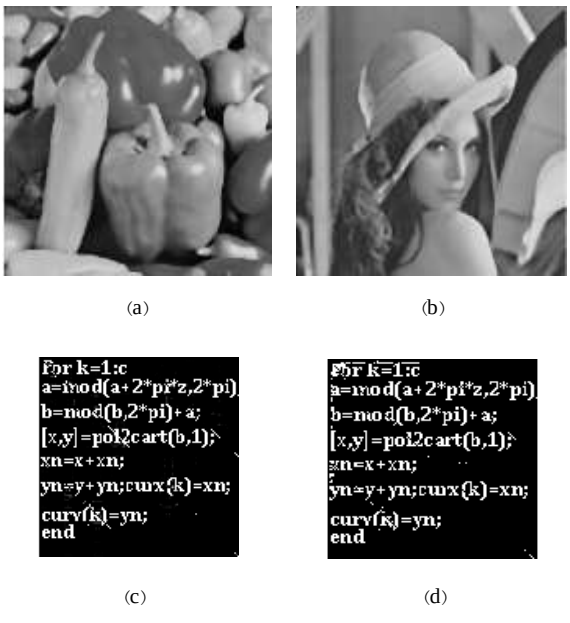


Fig. 7. (a), (b) Attacked watermarked images by Median filter 9×9; (c), (d) Extracted watermark images.

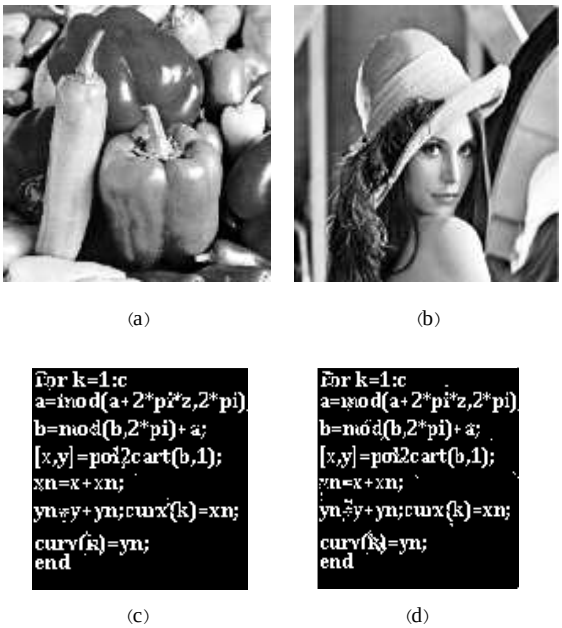


Fig. 8. Attacked watermarked images by Histogram equalization; (c), (d) Extracted watermark images.

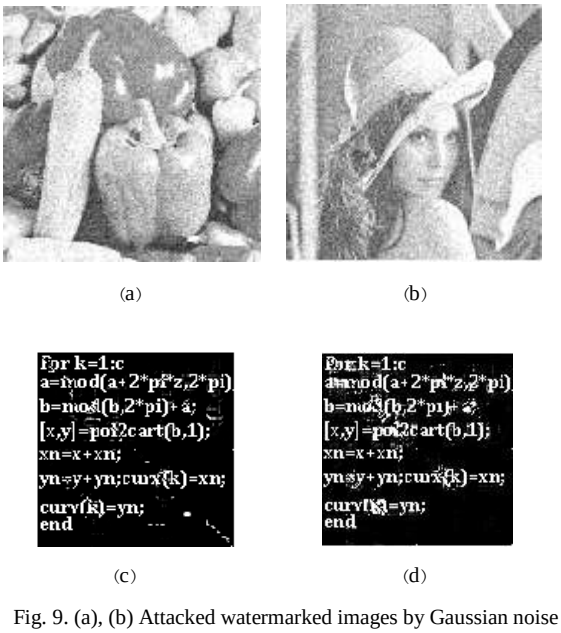


Fig. 9. (a), (b) Attacked watermarked images by Gaussian noise 50%; (c), (d) Extracted watermark images

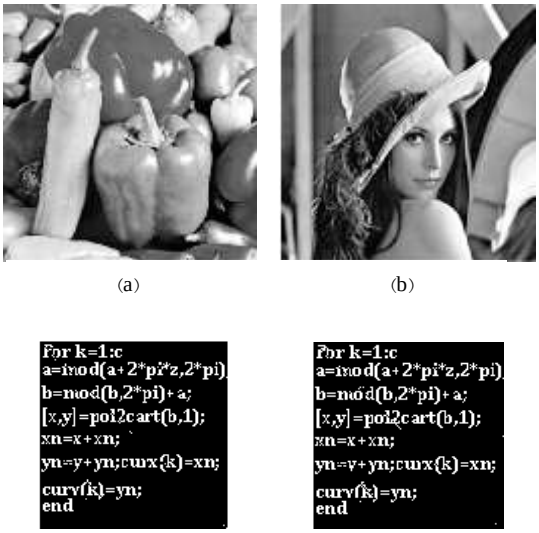


Fig. 10. (a), (b) Attacked watermarked images by Contrast Adjustment; (c), (d) Extracted watermark images.

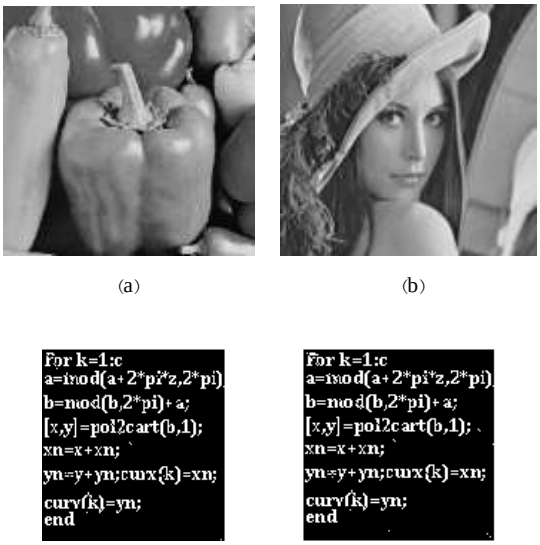


Fig. 11. (a), (b) Attacked watermarked images by Cropping; (c), (d) Extracted watermark images.

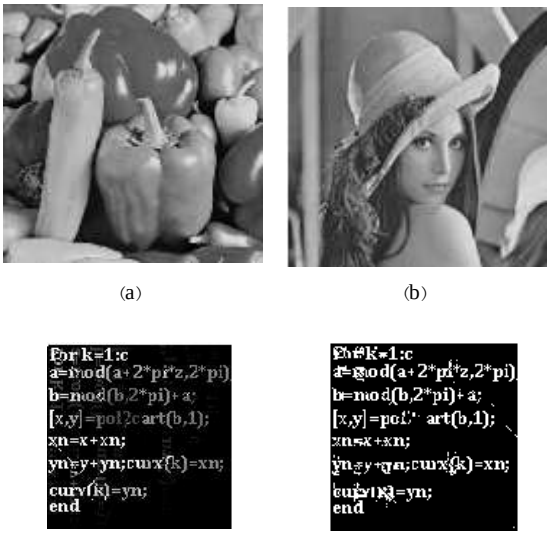


Fig. 12. (a), (b) Attacked watermarked images by Jpeg Compression with QF=20; (c), (d) Extracted watermark images.

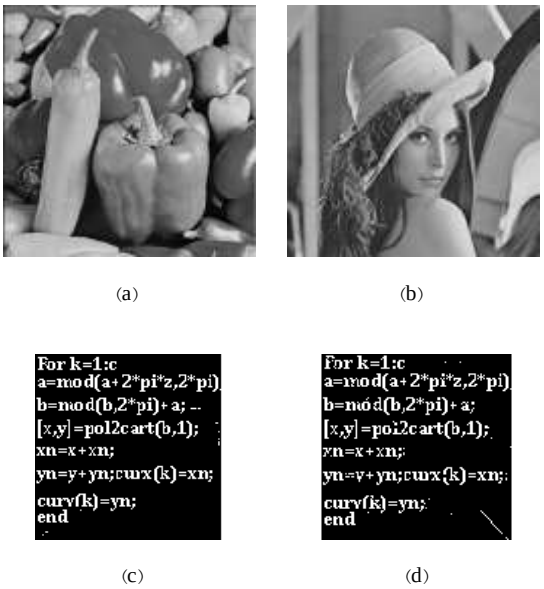
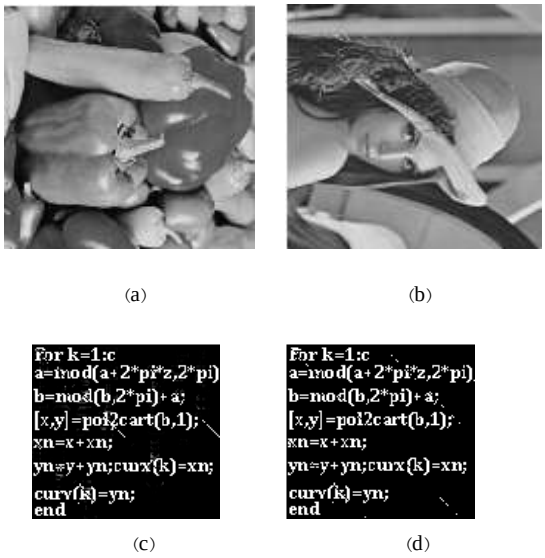


Fig. 13. (a), (b) Attacked watermarked images by Resizing (100% to 25%); (c), (d) Extracted watermark images.



(a) (b) (c) (d)

Fig. 14. (a), (b) Attacked watermarked images by Rotation 90˚; (c), (d) Extracted watermark images.

4-1- Comparison

In this subsection, the results of performing the proposed and the other four watermarking algorithms are compared. The host image is Lena and the Robustness and transparency are the criteria for comparison.

TABLE 3 shows the result of the comparison between the correlation coefficient of the proposed method and four other recently published papers[1] ([22]-[25]).

For the sake of comparison these algorithms are implemented in the same conditions. In this table the last row shows the average amount of correlation coefficients for the applied attacks.

Table 4 shows the result of the PSNR comparison between the mentioned methods. There is no difference between these methods from the image quality aspect. Scaling factor for this comparison is 0.1. The results indicate both higher PSNR values and more robustness for the proposed algorithm than the other compared algorithms. This is in spite of the limitation of the images which can be produced by the fractal images.

In table 5, the blindness of the proposed algorithm with references no. [22-25], has been compared. In this table the non-watermarked parameters/data needed to reconstruct the watermark image moreover the extracted information from the host image has been briefly explained. As the table shows, except the work of Yang et al. [25], the blindness property of proposed algorithm is better or analogous to the other compared works in the table.

[1] Although there are many published watermarking algorithms in the literature, regarding the robustness and transparency criteria, these two algorithms are better than or at the same level with other algorithms. Therefore comparison with these two algorithms are reasonable and enough to show the priority of the proposed algorithm.

Table 3. Correlation coefficients comparison.

Attack	Correlation Coefficients					
	Proposed	[22]	[23]	[25]	[24]	
					First	Second
No	0.9998	0.9980	0.9995	0.9990	0.9724	0.9854
Average Filter 9×9	0.9820	0.9010	0.8990	0.9060	0.8100	0.9500
Un-sharpening	0.9600	0.9610	0.9100	0.8560	0.9100	0.9600
Histogram Equalization	0.9684	0.9862	0.9000	0.9635	0.9800	0.9900
Gaussian Noise σ=0.01	0.9399	0.9575	0.8560	0.9460	0.6800	0.8300
Contrast Adjustment	0.9754	0.9900	0.9450	0.9000	0.7800	0.7850
Gama Correction γ=0.6	0.9693	0.9296	0.9301	0.9554	0.9100	0.9300
Jpeg Compression with QF=75	0.9657	0.9658	0.9010	0.9541	0.9700	0.9800
Resizing	0.9839	0.9700	0.9600	0.8628	0.8500	0.9800
Average	**0.9716**	**0.9621**	**0.9223**	**0.9269**	**0.8736**	**0.9322**

Table 4. PSNR for several watermarking schemes on Lena host image.

Method		PSNR with α=0.1
Run et al.[24]	First method	32.17 db
	Second method	33.93 db
Yang et al.[25]		40.1891 db
L&T [22]		46.02 db
T&J&L [23]		24 db
Proposed Method		70.515 db

Table 5. Comparing the blindness of different schemes

Watermarking Algorithm	Data used in Extraction Watermark Step	Non-Blind, Semi-Blind or Blind
[22]	S matrix after applying SVD transforms to host image.	Semi- Blind
[23]	Using SVR Training	Not Fully Blind(Similar to Semi- Blind)
[24]	Used host image (B_k Sub-bands in embedding stage)and watermark image	Non-Blind
[25]	Neither needs the original host image nor any other side information.	Blind
Proposed	S matrix after applying SVD transforms to host image.	Semi-Blind

5. Conclusion

In this paper, a hybrid image watermarking technique based on LWT and SVD has been presented, where the watermark is embedded on the singular values of the cover image's LWT sub-bands (LH and HL). In this work, original watermark is fractal image that converts to its constructor codes. In this way, instead of the original watermark for embedding, we used fractal coding that is much smaller than original watermark. Our algorithm is robust against various attacks including average filter, median filter, contrast adjustment, Jpeg compression, rotation, scaling, resizing and cropping. Experimental results of the proposed technique have shown both the significant improvement in imperceptibility and the robustness under attacks.

References

[1] H.-T. Wu and Y.-M.Cheung "Reversible watermarking by modulation and security enhancement" IEEE Transactions on Instrumentation and Measurement, vol. 59, no. 1, pp. 221–228, Jan. 2010.

[2] C. Chang, P. Tsai, and C. Lin, "SVD-based digital image watermarking scheme," Pattern Recognition Letters, vol. 26, pp. 1577–1586, 2005.

[3] M. Fan, H. Wang, and S. Li, "Restudy on SVD-based watermarking scheme," vol. 203, pp. 926–930, 2008.

[4] R. Run, S. Horng, J. Lai, T. Kao, and R. Chen, "Expert Systems with Applications An improved SVD-based watermarking technique for copyright protection q," Expert Systems With Applications, vol. 39, no. 1, pp. 673–689, 2012.

[5] G. Bhatnagar "A new facet in robust digital watermarking framework" AEUE-International Journal of Electronics and Communications, vol. 66, no. 4, pp. 275–285, 2012.

[6] W. Lu, W. Sun, and H. Lu, "Robust watermarking based on DWT and nonnegative matrix factorization," Computers and Electrical Engineering, vol. 35, no. 1, pp. 183–188, 2009.

[7] G. Bhatnagar, Q. M. J. Wu, and B. Raman "Robust gray-scale logo watermarking in wavelet domain q" Computers and Electrical Engineering, vol. 38, no. 5, pp. 1164–1176, 2012.

[8] M. Keyvanpour and F. Merrikh-bayat "Procedia Computer Robust Dynamic Block-Based Image Watermarking in DWT Domain" Procedia Computer Science, vol. 3, pp. 238–242, 2011.

[9] S. Rawat and B. Raman "A blind watermarking algorithm based on fractional Fourier transform and visual cryptography" Signal Processing, vol. 92, no. 6, pp. 1480–1491, 2012.

[10] S. Rawat and B. Raman, "Best tree wavelet packet transform based copyright protection scheme for digital images," OPTICS, vol. 285, no. 10–11, pp. 2563–2574, 2012.

[11] G. Bhatnagar, B. Raman, A new robust reference watermarking scheme based on DWT-SVD, Computer Standards & Interfaces (2009), doi: 10. 1016/j. csi. 2008.09.031.

[12] A. Manjunath, "Comparison of Discrete Wavelet Transform (DWT), Lifting Wavelet Transform (LWT) Stationary Wavelet Transform (SWT) and S-Transform in Power Quality Analysis," European Journal of Scientific Research, vol. 39, no. 4, pp. 569–576, 2010.

[13] H. Kiya, Iwahashi and Watanabe "A new class of lifting wavelet transform for guaranteeing losslessness of specific signals," IEEE ICASSP2008, pp. 3273–3276, 2008.

[14] S. K. Jinna and L. Ganesan, "Lossless Image Watermarking using Lifting Wavelet Transform," International Journal of Recent Trends in Engineering, vol. 2, no. 1, pp. 191–195, 2009.

[15] K. Loukhaoukha, "Optimal Image Watermarking Algorithm Based on LWT-SVD via Multi-objective Ant Colony Optimization," Journal of Information Hiding and Multimedia Signal Processing, vol. 2, no. 4, pp. 303–319, 2011

[16] L. Lamarche, Y. Liu, J. Zhao, K. E. Ave, and C. Kn, "Flaw in SVD-based Watermarking," IEEE CCECE/CCGEI, pp. 2082–2085, 2006.

[17] H.-hsu Tsai, Y.-jieJhuang, and Y.-shou Lai, "An SVD-based image watermarking in wavelet domain using SVR and PSO," Applied Soft Computing Journal, vol. 12, no. 8, pp. 2442-2453, 2012.

[18] X. Wu and W. Sun, "Robust copyright protection scheme for digital images using overlapping DCT and SVD," Applied Soft Computing Journal, pp. 1–13, 2012.

[19] C. Lai, "A digital watermarking scheme based on singular value decomposition and tiny genetic algorithm," Digital Signal Processing, vol. 21, no. 4, pp. 522–527, 2011.

[20] N. M. Makbol and B. E. Khoo, "Robust blind image watermarking scheme based on Redundant Discrete Wavelet Transform and Singular Value Decomposition," AEUE - International Journal of Electronics and Communications, pp. 1–11, 2012.

[21] J. Song, J. Song, and Y. Bao "A Blind Digital Watermark Method Based on SVD and Chaos" Procedia Engineering, pp.285-289, 2012.

[22] C. Lai and C. Tsai "Digital Image Watermarking Using Discrete Wavelet Transform and Singular Value Decomposition" IEEE Transactions on Instrumentation and Measurement, vol. 59, no. 11, November 2010.

[23] H. Tsai, Y. Jhuang, and Y. Lai, "An SVD-based image watermarking in wavelet domain using SVR and PSO," Applied Soft Computing Journal, vol. 12, no. 8, pp. 2442–2453, 2012.

[24] R. Run, S. Horng, J. Lai, T. Kao, and R. Chen, "An improved SVD-based watermarking technique for copyright protection q," Expert Systems With Applications, vol. 39, no. 1, pp. 673–689, 2012.

[25] H. Yang, X. Wang, and C. Wang "A robust digital watermarking algorithm in un-decimated discrete wavelet transform domain" Computer and Electrical Engineering, 2012.

Kayvan Ghaderi received his M.Sc. in artificial intelligence from the Faculty of Computer Engineering, University of Kurdistan, Sanandaj, Iran in 2013. His current research areas of interest include digital watermarking, data hiding, image enhancement, multimedia communications and fuzzy systems.

Fardin Akhlaghian Tab was born in Sanandaj, Iran in 1965. He received his bachelor's degree in Electronic Engineering from Isfahan University of Technology in 1989 and his M.Sc. Degree in Telecommunications and Signal Processing from Tehran University of Tarbiat Modarres in 1992. From 1993 until 2000 he was a scientific member of Computer Engineering Department at University of Kurdistan, Sanandaj, Iran. From 2001 until 2005 he studied for his Ph.D. degree at University of Wollongong, Australia and then after receiving the Ph.D. degree in 2005, he again joined the Department of Computer Engineering and Information Technology at University of Kurdistan. His major research interests are pattern recognition, image processing and computer vision.

Parham Moradi received his Ph.D. degree in Computer Science from Amirkabir University of Technology in March 2011. Moreover, he received M.Sc. and B.Sc. degrees in Software Engineering and Computer Science from Amirkabir University of Technology, Tehran, Iran, in 1998 and 2005 respectively. He conducted a part of his Ph.D. research work in the Laboratory of Nonlinear Systems, EPFL (Ecole Polytechnique Federale de Lausanne), Lausanne, Switzerland, form September 2009 to March 2010. Currently he is working as an Assistant Professor in the Department of Computer Engineering and Information Technology, University of Kurdistan, Sanandaj, Iran. His current research areas include Reinforcement Learning, Graph Clustering, Social Network Analysis, Data Mining and Recommender Systems.