International Journal of Electrical and Computer Engineering (IJECE)
Vol. 3, No. 6, December 2013, pp. 797~804
ISSN: 2088-8708

The Security of Arithmetic Compression Based Text Steganography Method

Reihane Saniei[1], Karim Faez[2]
[1]Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University
[2]Departement of Electrical Engineering, Amirkabir University of Technology

Article Info

Article history:

Received Jul 30, 2013
Revised Oct 12, 2013
Accepted Nov 1, 2013

Keyword:

Arithmetic coding
Data compression
Information hiding
Security
Text steganography

ABSTRACT

Security of a modern design of steganography on lossless compression is studied in this paper. Investigation of a set of methods presented here indicates that there are various approaches to establish a hidden and safe relationship with the minimum cost for text files. Although, steganography of information in text is one of the most difficult areas of steganography, many efforts were made in this regard. With regard to the spread of this category and existence of wide volume of approaches, this paper deals with comparison and evaluation of steganography security by a statistical compression method called arithmetic coding and other methods of text steganography. Moreover, this method is available for audio-visual and video files. In addition, stego key was placed in a format that it would not arouse any suspicions. It is notable that this new method of steganography or rewriting and syntactic and semantic review does not reveal the secret message and results in 82.88% improvement in security.

Corresponding Author:

Reihane Saniei
Faculty of Computer and Information Technology Engineering,
Qazvin Branch, Islamic Azad University,
Nokhbegan Blvd, Qazvin, Iran.
Email: saniei_re@yahoo.com

1. INTRODUCTION

In today's world, we cannot imagine life without a computer and digital communications have become a vital part of our life. However, the use of computer accompanies by a discussion of security in data transfer, information coding and encoding. In addition, an efficient privacy has been provided through a combination of techniques of hiding information and coding. Since many applied programs are internet-based, having a secret communication is important and ensuring the security of information passing an open canal has been a crucial issue. Therefore, dependability and integrity of data need protection against unauthorized access and utilization.

Today, confidential data can be protected by different approaches of hiding information. Cryptography, steganography, and watermarking are three general techniques to conceal information, one of which hides the existence of a message, and the other means hiding information as a media format such as image, audio, video, and even a text so that other people do not notice the existence of information in above-mentioned format, and finally watermarking means to protect copyright. In recent years, approach to conceal information has paid great attention to steganography and watermarking techniques.

Steganography is science and art of embedding information in a block of host data in conditions where considerable changes in host data are unacceptable [1]. Steganography techniques have a high dependence to the characteristics of host data. For example, steganography of an image may produce delicate changes in color or hiding information in the sound use the limitations of human hearing for data coding [1]. Information steganography in text is one of the most difficult areas of steganography. Since eye can easily

distinguish difference between original and cover texts and in spite of image, audio, and video, text files do not have redundancy, thus many changes cannot produced in them [2]. One of the first methods of text steganography is the use of a system that extracts message automatically or hides it in n^{th} characters or hides coded message by changing distance after or between characters [2]. Of course, all of these methods had a low security and some of them have attracted hackers' attention and or the coded message was removed by rewriting the original text [2].

Another published method of steganography was to put data as noise in a covering media or change the available text format and style. However, such changes attracted hackers' attention and facilitated detection of coded text [3]. With time and advancement of steganography, the use of covering media accompanied by a coded key was prevalent, which had a higher degree of security.

Despite great difficulties of this work, enormous efforts were made to design text steganography methods in English, Chinese, Persian, Arabic, and so forth that can generally be divided into three main categories as shown in figure 1.

The format-based method of steganography includes changing the words, lines, spaces and features of the original text and the linguistic methods is divided into syntactic and semantic classes.

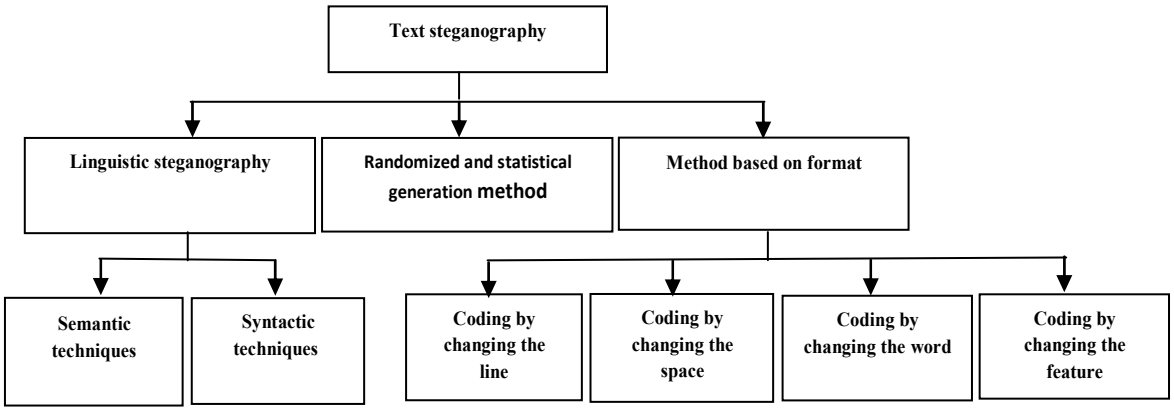

Figure 1. Classification of Steganography in Text

However, what has been common in all these proposed methods was an attempt to meet three main conditions of hiding information system; that is, security, resistance, and capacity [4]. Capacity refers to the amount of data hidden in covering media. Security relates to the ability of eavesdropping to detect hidden information and prevent detection of the message presence. Resistance means probable strength against change, noise, and manipulation of information in a general canal without prevention of a hidden relationship [4, 5, 6].

The method recommended in this paper is to use computational compression method for improvement in capacity and security of steganography.

According to Shannon's theory, applied programs of compression employ various techniques that have a different degree of complexity and compress our data as first order entropy. They can be divided into two categories of lossless and lossy. Since we deal with text thus there is no redundancy and inevitably, only lossless methods will be used [7].

In general, different algorithms were proposed to compress text data, all of which are lossless. We use arithmetic coding method here that has a higher rate of compression than other methods, produces an optimal prefix code, has logarithm time complexity, and can make algorithms more complicated.

Arithmetic coding is one of the popular methods of lossless statistical compression. In fact, it is an effective mechanism to eliminate redundancy in coding information that easily adjusts to statistical inputs. Its main idea was set and formulated by Elias in the early 1960s [8]. The first stage of practical implementation of this idea was carried out by Rissanen and Pasco (1976) [9, 10].

Production of optimal prefix code is one of the advantages of this coding. Prefix code is a one that is unique and the probability of its production error is zero. Its' other advantage is to have the highest compression rate in lossless methods after Huffman's coding. Despite Huffman's method, this one can have a high bit rate in situations where the size of symbols is small or probability of one of them is large. Moreover,

this method has the least time complexity among all other lossless methods and just needs to draw a table in an adaptive mode.

One of the main applications of arithmetic coding is to use it for coding a specific and unique string (tag) without coding for the same strings where this tag can play the role of stego key for us.

One of the problems of arithmetic coding is to generate data out of range. Scaling method has been proposed to overcome this problem, which keeps our data in a logical range [7].

In today's digital world, computers' users have been provided with many tools, communication canals and technologies through which much more advanced methods than old techniques can be developed and applied. However, what is certain is that each of these methods will provide a level of security and capacity for users. Therefore, it should not be neglected that there is not a direct relationship between capacity and security, and for increase in each of them the other one should be sacrificed.

It was tried to provide an acceptable level of both in the recommended method. Thus, we did not attempt just to enhance the capacity but we can say with certainty that this method is unique with regard to ensure security against active and inactive attackers while providing a good capacity. In fact, this method improves methods of Satir, who have used Huffman and LZW compression [5, 6], and has higher embedding capacity than their methods and proposed methods to date. Although security has been specifically studied in this paper, diagram of appendix 1 showed 68.9% improvement in capacity.

Furthermore, having no punctuation such as space character or semantic signs such as replacement of some words with their synonyms are considered as other properties of this method and in fact, attackers cannot easily eliminate our message by rewriting and retyping. In addition, this method does not make a noise because it transfers text as a text not an image and finally, results of the experiments conducted on recommended method confirm 82.88% improvement in comparison to all methods proposed to date.

In section 2, the proposed method will be introduced. The research method will be explained in section 3. Finally, experimental results of the proposed method and conclusion will be presented in sections 4 and 5.

2. THE PROPOSED METHOD

Although steganography has a relatively high history, there is still a lot of work to do in order to be able to present safe methods with high embedding capacity. Since researchers' effort for steganography has been in tandem with hackers' effort for violations of steganography, a modern method for steganography has been presented here that increases the embedding capacity of messages compared to previous methods while ensuring security.

In this algorithm, a database of texts and an array of email addresses between the sender and receiver has been shared. Our coded message will be hidden in recipient addresses, and have the stego key role. Decoding at the receiver will be done on the same address, the text in the body of the email and the information has been shared.

2.1. Procedure of Steganographer Embedded

- At the beginning our coded message considers as an array of characters. Then, using our common text base, making up the matrix, its elements, show the difference in position of the characters of coded message with the text have chosen of our database.
- After completion of the matrix to pay its normalization. This is done by dividing the number 26. So that can applied synthetic code of Latin square [11]. The quotients of this division of *REMAIN* matrix and the remainder of the division is called *EXTRA* matrix.
- At this stage to make a coverage text, the greatest numbers of occurrences of binary patterns in *REMAIN* matrix is calculated and its corresponding row, called *REFER* vector and in event is the index that refers to the coverage text. Our coverage text in the body of the email is written.
- Compress the *REFER* vector with computational compression method with scaling, and make results binary and using by that and Latin square tables our new email addresses are created. Thus, using the addresses is the key to our treasure house and coverage text was built in the previous stage, our coded message is transmitted.

2.2. Procedure of Steganographer Extraction

- Performing the above operation vice versa, the message will be decrypted.

3. RESEARCH METHOD

Information hiding techniques in today world are playing a critical role for secure transmission of information in the public channels. Steganography is one of the most important techniques, which hides the message in the coverage file. Over the past decade, many methods have been proposed to steganography, which, according to our coverage file that is text, image, and etc. they are classified.

In addition, the methods of steganography in the text can be classified according to safety or embedded capacity of messages. Of course, if have a high-capacity with low-quality is not acceptable criterion to us. The low quality of steganography in the picture or video is meant to low -resolution, but in the steganography of a text meaning that in the context statistical and grammatical features have not been observed or the visual features are recognizable, in fact, is the security criterion that considers the matter, the capacity will increase to some extent still our relationship remain secure.

Since each data hidden in the cover, because changing it, so security is the criterion that amount of deviation of the cover shown than the original text. Most often security of message embedded where the text of the cover has changed and the amount of changes imposed upon it, are influenced [6, 12].

Theoretically security along with Kerechoffs principles and Shannon's reviews on these principles is described. From Kerechoffs view, security of system should accompany with key confidentiality [13] and, according to Shannon information theory can be used to provide security and in our assumptions attacker should knows the system [14].

To have a quiet secure system, our system should be robust against all kinds of attacks. Attack to the system can be of syntactic and semantic attack and statistical attack or passive and active. In a passive attack, the attacker is only able to analyze the system and is unable to manipulate the channel and information and in active attacks the attacker is able to modify the data [15].

In order to combat active attacks the digital signature is required. In order to provide security against passive attacks, Cachin proposed relative entropy to accurate calculation. (According to the formula: 1) [15, 16].

$$D\left(P_c \| P_s\right) = \int P_c . \log \frac{P_c}{P_s} \leq \varepsilon \quad (1)$$

In this theory having complete security in the steganography system means ε=0. It should be noted that such a system does not exist in practice, and one of our hypotheses in this system is that the cover text and stego text, are independent identically distributed (i.i.d) random variables. Solutions to correct this condition, is calculating the relative entropy close to zero [12].

Where *Pc* is the probability distribution of cover text where *Ps* is probability distribution of stego text and *D* is relative entropy.

However, many efforts have been carried in the text of steganography to date, different techniques work for different languages, with different applications and different degree of security has been designed. In Table 1, some previous studies have compared with proposed method in term of safety. Among of the positive features of arithmetic coding techniques is production of safer coverage text than methods that have been proposed to date.

4. RESULTS AND DISCUSSION

In this section, the algorithms given in section 2 have implemented by content programming language and repeated the applications for 100 coded messages. The length of coded message at least is 10 bits and maximum is 100 bits. Here, each of the ten coded messages have been classified as *Li*. Per replicates for each *Li* the value of relative entropy that is same $\boldsymbol{\varepsilon}$, calculated according to the formula 1 and the results are shown in Figure 2. In this diagram the horizontal axis is length of coded message and vertical axis show values of ε and *Li* for different messages**.** According to the experiments the value of ε for this algorithm is equal to the average 0.1712. According to this diagram, increasing length of message, our security level reduced that according to the formula 1 and probability distributions of *Pc* and *Ps* is fully justified. Thus, shorter messages are more secure. According to above calculation the amount of the security of proposed method against active attacks is 82.88 %. This method against active attacks is not secure and if the active attacker manipulate coverage text or stego key (changing email address that placed at send section or change the email text), the only way to deal with the attacks is using digital signature.

Table 1. An overview of the various methods of steganography [3, 5, 6, 17, 18, 19]

Method	Year	Definition	Advantages	Disadvantages
Mimic [17]	1992	Using reverse Huffman code randomly generates a stream of data.	1. Resistant against statistical attacks 2. Correct grammar 3. Following context-free grammar and syntax, Van Vijnardon to increase output	1. unclear of its output in term of lexical, syntactic, and then arouses suspicion
TEXTO [20]	1995	To convert ASCII data into English sentences.	1. The appropriate binary data 2. Simple replacement 3. Correct grammar 4. high capacity	1. Having incoherent stego cover and unclear output, so it does raise suspicions. 2. Markup and related words anywhere of list.
NICETEXT [21], Winstein[1] Murphy[22] Nakagawa [23]	1997 To 2007	Generating a coded message by replacing synonyms	1. Having a cover text logically and linguistically valid 2. Lack of loss data by re-typing	1. The frequent use of the same part of the text, it does raise suspicions. 2.Changing the meaning of the text
L-R scheme [24]	2004	Left and right combinations of Chinese characters uses as cipher text. It was improved and all characters were used and its disadvantages resolved.	1. Improve capacity than the other method - that used for Chinese	1. Special for Chinese language 2. The need for text steganography to maintain files consistency 3. limited capacity
Wang and Chang [25]	2009	In the forums put coded messages into visual icons	1. Use a private encryption key 2. high security	1. The transmitter and receiver before must be shared a set of video icon 2. Our capacity is dependent on the number of shared visual icons
Translation based methods [3]	2009	Coded message hides on the (noise), spelling, grammar error	1. The use of machine translation and synonymous substitution to improve the bit rate	1. For all languages cannot be used 2. Sensitive to the amount of noise 3. Outlet is impossible to read and incoherent
Listega [3]	2009	Use a text list with logic items to hidden data	1. Easy coding and extraction 2. Protection of authenticity of coverage text 3. Use synthetic codes to enhance security 4. The accuracy of linguistic and logical 5. High capacity at most time 6. Suitable for all languages	1. The low security level for easy extraction procedures
SPAM method [6]	2009	Sending the Cipher text to multiple recipients via an outlet media (such as Web, TV, etc.) assume the availability of OCR functionality in decoder	1. For SMS messages in any language are used. 2. Publish the coded messages to different receivers at different locations at a time	1. Decoding in decoder without OCR functions can be done with eye so have low security 2. It is required the sender continuously checks transmit channel 3. Messages will be sent visually
Lee and Tsai [18]	2010	Embedding the message in files PDF (manipulation of space)	1. Easy coding and extraction	1. Limited capacity and its dependence on the number of PDF characters 2. Problems with security
Ryabko [26]	2011	This method for limited memory with unknown probabilities	1. It has full security 2. An observer cannot know whether the secret information is sent or not.	1. It has a non-optimal transmission rate. 2. It is used only to i.i.d coverage text.
UniSpaCh [19]	2012	steganography is carried by manipulation of White space	1. Increase in performance of embedded 2. Improve capacity	1. If frequent used in the document, affects appearance features of the document 2. It has simple extraction procedure and is the known method, the security level is high.
LZW compression [5]	2012	Steganography by helping LZW lossless compression method	1. High capacity 2. Enhance the security of stego key and synthetic code and LZW code 3. Complex extraction procedure 4. Use email template to be invisible	1. If our data is changed, we need to change our dictionaries. 2. Its code is not optimal and prefix 3.LZW has (n) time complexity.
Huffman compression [6]	2012	Steganography by helping Huffman lossless compression method	1. Higher capacity than LZW method due to higher rates of compression 2. It has all the advantages of LZW method 3. it has an optimal prefix code	1. If the number of character of encoded message is low or probability one of them to be high, Huffman would be a good bit rate, so our capacity is reduced 2. Huffman code has (n log n) time complexity
Proposed Method	2013	Steganography by helping Arithmetic lossless compression method	1. Elevate capacity than Huffman and LZW method, especially for short messages 2. It has all the advantages of Huffman compression method. 3. In the adaptive condition just need a table	1. For messages with too much length causes the overflow

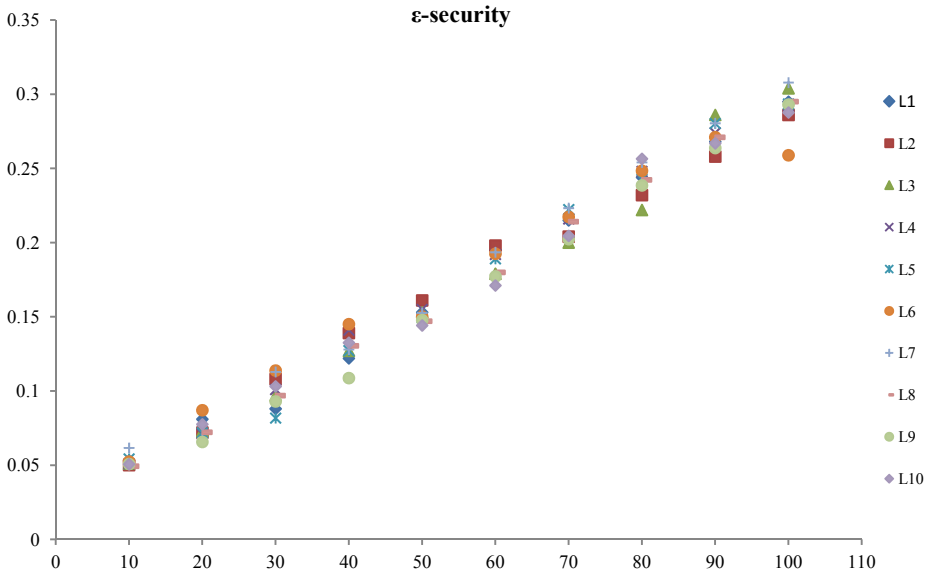

Figure 2. Graph of ε –security versus length of secret message. In this diagram, the horizontal axis shows the length of the secret message and the vertical axis represents the values of ε, and L_i is repeater to display different messages.

5. CONCLUSION

Steganography techniques are an attempt to keep the relationship secret, which is achieved when the least subtle changes to be applied in our coverage text.

In this paper, to solve the problems of the steganography in the text, a method is provided. Its coverage text is resistant to statistical and random attacks. In this way the stego text is transferred as text not image, and syntactically and semantically and visual features are quite correct.

Experimental results of the proposed method show its effective in enhancing the capacity and security. This article has been specializing in security issues, however, this method improve the capacity 68.9% compared to the contemporary approaches (see appendix 1). Also, the amount of security against passive attacks improved 82.88 %. However, at this point due to increase safety, above calculation is required and our algorithm is relatively complex, however, despite the processing high power of today computers, it is perfectly applicable, while increases the security and capacity, unlike the majority modern methods as well. So these processing costs against its benefits are acceptable.

To enhance the capacity of embedding of message Lossless compression method that has a high bit rate has been used. Among problems of this plan can referred to produce results beyond the scope in using the long message, though the purpose of steganography project for short messages with capacity is high, however tried to deal with this problem with scaling.

In future works , this technique can be combined with Listega steganography method to more complicated and secured this proposed method , because the Listega method produces synthetic code , in addition to be simple has high capacity and protect the authenticity of the text.

APPENDIX 1

Due to the capacity of the embedded messages in a quantity of hidden data in coverage media, the algorithm mentioned in section 2, after the implementation by content language, once for 100 coded messages that noted for both articles Satir [5, 6] and once again have been tested by 100 coded messages. The results have been plotted in Figure 3. So this method improve the capacity 68.9% compared to the contemporary approaches.

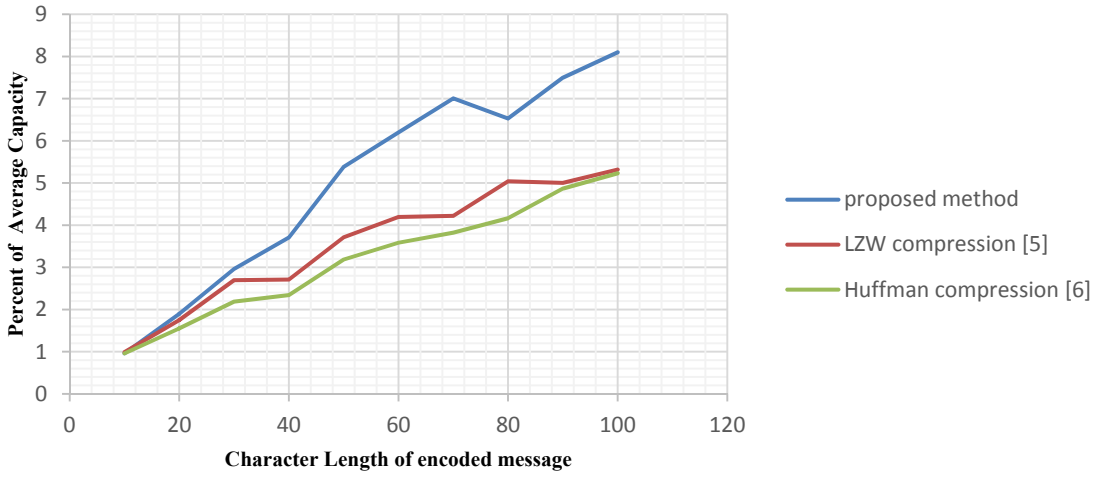

Figure 3. Diagram of percent of capacity versus character length of encoded message in proposed method , LZW method [5] and Huffman method [6].

REFERENCES

[1] K Winstein. "*Lexical steganography*". 1999, http://alumni.imsa.edu/~keithw/tlex .Accessed 3 Aug 2008.
[2] FH Rehab, FH Nidaa. "Data hiding in Arabic text based on Letters, Diacritics and Extension". *First information Technology conference*. 2009: 21-23.
[3] A Desoky. "Listega: list – based steganography methodology". *International Journal of Information Security*. 2009; 8(4): 247-261.
[4] FDL Rochefoucauld. "Data hiding", *Coding for data and computer communications*, Springer US. 2005: 341-363.
[5] E Satir, H Isik. "A compression – based text Steganography method". *The Journal of Systems and software*. 2012; 85: 2385-2394.
[6] E Satir, H Isik. "A Huffman compression based text Steganography method". *Multimedia Tools and Applications*, Springer US, ISSN 1573-7721. 2012: 1-26.
[7] K Sayood, M Kaufmann. "Introduction to Data Compression". Edited by Adams R. ISBN-13: 978-0-12-620862-7, 3rd Edition, San Francisco, Kaufmann M. Publishers is an imprint of Elsevier, 2006.
[8] J Frederic. "Probabilistic information theory". New York: Mc Graw-Hill. 1968: 476-489.
[9] R Jorma. "Generalized Kraft inequality and arithmetic coding". *IBM Research Laboratory*, Monterey and cottle Roads, San Jose, USA. 1976; 20: 198-203.
[10] R Boris, F Andrei. "Fast and space-Efficient Adaptive Arithmetic Coding". *Cryptography and coding*, lecture Notes in computer science, 1999; 1746: 270-279.
[11] G Lei. "Latin Squares in Experimental Design". Michigan State University, 2005, http://www.mth.msu.edu/~jhall/classes/mth880-50/projects/latin.pdf.
[12] H. Sajedi, M. Jamzad, "Secure steganography based on embedding capacity", Journal of information Security, Springer, vol. 8, pp. 433-445, 2009.
[13] A Kerckhoffs. "La cryptographie militaire (Military cryptography)". *Journal des sciences militaires.* 1883; 9: 5–83, 161–191.
[14] CE Shannon. "Communication theory of secrecy systems". *Bell System technical Journal*. 1954; 28: 656-715.
[15] IJ Cox, T Kalker, G Pakura, M Scheel. "Information transmission and steganography". Springer. 2005; 3710: 15-29.
[16] C Cachin. "*An information-theoretic model for steganography*". In Proc. On the second Workshop on Information Hiding, of Springer Lecture Notes in computer Science. 1998; 1525: 306–318.
[17] Y Liu, X Sun, Y Liu, CT Li. "MIMIC-PPT: Mimicking - Based Steganography for Microsoft Power Point Document". *Information Technology Journal*. ISSN 1812-5638.2008; 7(4):654-660.
[18] IS Lee, WH Tsai. "A new approach to covert communication via PDF files". *Signal Process 90*. 2010; 2: 557-565.
[19] LY Por, K Wong, KO Chee. "UniSpaCh: a text based Data hiding method using Unicode space characters". *Journal of Systems and Software*. 2012; 85:1075-1082.
[20] K Maher. "TEXTO". ftp://ftp.funet.fi/pub/crypt/steganography/texto.tar.gz.
[21] M Chapman, G Davida. "Hiding the hidden: a software system for concealing cipher text as innocuous text". *Lecture Notes in computer science*, Springer Beijing. 1997; 1334: 335-345.
[22] B Murphy, C Vogel. "The syntax of concealment: reliable method for plain text information hiding". In: *Processing of the SPIE International Conference on Security, Steganography and Watermarking of Multimedia Contents.* 2007.
[23] H Nakagawa, K Sampei, T Matsumoto, S Kawaguchi, K Makino, I Murase. "Text information hiding with preserved meaning – a case for Japanese documents". IPSJ Trans. 2001; 42(9): 2339-2350, originally published in

Japanese. A similar paper by the first author in English. http://www.r.dl.itc.u-tokyo.ac.jp/nakagawa/academic-res/finpri02.pdf. Accessed 4 June 2008.
[24] XM Sun, G Luo, HJ Huang. "*Component – Based digital watermarking of Chinese texts*". In: Proceedings of the 3rd International Conference on Information Security, Shanghai, China. 2004: 76-81.
[25] ZH Wang, TD Kieu, CC Chang, MC Li. "*Emoticon – Based text steganography in chat*". In: Proceedings of 2009 Asia – Pacific Conference on computational Intelligence and Industrial Application (PACIIA2009), Wuhan, China, 2009b.2009; 2: 457-460.
[26] B Ryabko, D Ryabko. "Constructing perfect stenographic systems". *Information and Computation*, Elsevier. 2011; 209: 1223-1230.

BIOGRAPHIES OF AUTHORS

Reihane Saniei was born in Tehran, Iran. She received her BSc degrees in Computer Engineering from Tehran Branch, Payamenoor University in February 2009, and is now studying student in software Engineering at Azad University of Qazvin.
Email: saniei_re@yahoo.com.

Karim Faez was born in Semnan, Iran. He received his BSc. degree in Electrical Engineering From Tehran Polytechnic University as the first rank in June 1973, and his MSc. and Ph.D. degrees in Computer Science from University of California at Los Angeles (UCLA) in 1977 and 1980 respectively. Professor Faez was with Iran Telecommunication Research Center (1981-1983) before Joining Amirkabir University of Technology in Iran in March1983, where he holds the rank of Professor in the Electrical Engineering Department. He was the founder of the Computer Engineering Department of Amirkabir University in 1989 and he has served as the first chairman during April 1989-Sept. 1992. Professor Faez was the chairman of planning committee for Computer Engineering and Computer Science of Ministry of Science, Research and Technology (during 1988-1996). Dr. Faez coauthored a book in Logic Circuits published by Amirkabir University Press. He published about 300 articles. He is a member of IEEE, IEICE, and ACM, a member of Editorial Committee of Journal of Iranian Association of Electrical and Electronics Engineers, and International Journal of Communication Engineering.
Emails: kfaez@aut.ac.ir, kfaez@ieee.org, kfaez@m.ieice.org.