International Journal of Security and Its Applications
Vol. 10, No. 9 (2016), pp.309-320
http://dx.doi.org/10.14257/ijsia.2016.10.9.30

Coverless Multi-keywords Information Hiding Method Based on Text

Zhili Zhou, Yan Mu, Ching-Nung Yang and Ningsheng Zhao

School of Computer and Software & Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science and Technology
zhou_zhili@163.com, my_muyan@163.com, cnyang@mail.ndhu.edu.tw, 151610880@qq.com

Abstract

As a new information hiding method, coverless information hiding has become a hot issue in the field of information security. The existing coverless information hiding method can hide only one Chinese character in each natural text. However, the problem of the method is that the hiding capacity is too small. To address this problem, a new method named coverless multi-keywords information hiding method based on text is proposed in this paper. The main idea of the method is that both the keywords and their number will be hidden in the texts. Experimental results show that the proposed method can improve the capacity of the existing coverless information hiding method based on text.

Keywords*: coverless information hiding; natural text; multi-keywords; hiding capacity*

1. Introduction

Recently, because of the importance on the privacy, copyright and others, information hiding has attracted extensive attention [1, 2]. More and more people have devoted themselves to the study of information hiding. The objective of information hiding is to embed the secret imperceptibility into the cover. As a consequence, it is difficult for illegal users to intercept the secret message through the public cover [3]. In the receiving end, the embedded secret message can be extracted for the purposes of secret communication and copyright protection. Information hiding has been widely applied in many applications, such as military, privacy, and property right protection.

Information hiding method can be divided into different categories according to the kinds of covers, such as information hiding based on text, information hiding based on image, information hiding based on video and information hiding based on audio [4]. During the past two decades, the conventional information hiding methods based on the text mainly change the letter-spacing to embed secret message. Some methods have been proposed by changing the line-spacing, letter-spacing, character height and character width [5-10]. Also, there are some typical image steganography methods include LSB matching [11-13], histogram-based methods [14], quantization table [15], and so on.

From the above, we can find that the existing information hiding methods adopt various technologies for text steganography. It is worth noting that they have a common ground. That is all of them modify the designated cover [16]. The modification traces caused by these methods will be left in the cover. As a consequence, it is probable that the traces would be detected triumphantly by steganalysis tools [17, 18]. A novel hiding method named coverless information hiding [16, 19], in which the secret message can be hidden without any modification, is effective to resist all of the steganalysis tools, so that it can be used to address this problem.

Coverless information hiding is a new concept, which was first proposed to resist all of the existing steganalysis tools. According to the reference [16, 19], coverless information

ISSN: 1738-9976 IJSIA

hiding is different from the traditional information hiding method that requires no other carriers. The main idea of coverless information hiding is to find the natural digital content which already contains the secret message. Since the natural digital content is used for secret communication, coverless information hiding will embed the secret message into the cover without any modification. As a result, it can resist all steganalysis technologies. From the above concept, the text is one of most common carries in our daily life, and thus text is adopted to hide the secret message in this paper. As we know, any natural texts contain a large number of Chinese characters. These Chinese characters may already contain the words needed to be hidden with a certain probability [16]. The method proposed in [19] can hide only one Chinese character in each text. However, an obvious problem in this method is that the hiding capacity is too low. How to improve the capacity of coverless information hiding has become a problem, which needs to be solved urgently.

To solve the problem of the existing method, this paper proposes a novel method, called coverless multi-keywords information hiding method based on text [20, 21]. The method can hide several Chinese characters in each text and outperform the existing work. To improve the capacity, this method hides both the secret message and the number of keywords in the same text simultaneously. The text will have two parts of information needed to be extracted, so that the receiver will extract the number of keywords and the secret message from the received text.

This paper is organized as follows. Section 2 summarizes the previous work and the motivation. Our proposed method is discussed in Section 3. Section 4 explains the experimental data and presents the results of our experimental research. Finally, Section 5 concludes the paper with brief words.

2. Previous Work and Motivation

Coverless information hiding method based on Chinese mathematical expression is a new method, which was proposed by Xianyi Chen, Zhili Zhou, and *et al.* to resist steganalysis tools [19]. Different from the traditional information hiding technology, this method does not require any other carriers. It is a secret message-driven method, in which natural texts containing the secret message are retrieved from the large-scale text database and then used as stego-texts for secret communication. Firstly, this method splits the secret message into several Chinese characters, and designs a list of location tags, which are used for indicating the positions of the hidden characters. Secondly, it chooses a location tag for each character. Finally, for each character, the texts that contain the combination of the character and its corresponding local tag are searched from the text database. Since natural texts are used as stego-texts for secret communication, the secret message can be hidden without any modification. Consequently, this method can resist all kinds of existing steganalysis tools, and has a positive importance in the field of information hiding. However, this method can hide only one Chinese character in each text. A natural text usually contains a large amount of Chinese characters, but they are not be used in this method. Therefore, the capacity can be enhanced by fully utilizing these characters.

Our proposed method is largely inspired by this method. Our method splits the secret message into several Chinese characters, which are called as "keywords". Since a natural text contains a lot of keywords, it is possible that multi-keywords can be hidden by finding the text that contains these keywords. The main problem is that how can the sender tell the number of keywords in each text to the receiver without any additional information. We propose a novel coverless information hiding method, which can send the number of keywords to the receiver without any additional information. Firstly, we embed several secret keywords into the text by finding the natural texts which already contain them. Secondly, to hide the number of keywords, a mapping relation will be created to map the number of keywords to some certain Chinese keywords. Then, by the

same manner, the number of keywords is hidden in the text that contains the corresponding keyword. Finally, we will obtain texts which contain both the information of the secret message and the number of keywords in each text. As a result, each text can be hidden more than one Chinese keyword, and the receiver can also know how many keywords are needed to be extracted from each text.

3. Coverless Multi-keywords Information Hiding Method Based on Text

In the proposed coverless multi-keywords information hiding method, there are three preprocessing steps, which are the construction of text database, the construction of inverted index structure, and the design of scheme to select location tags. A large-scale text database ensures that the texts containing the secret message can be found. An index structure is necessary to reduce the search time. Location tags are designed to locate the secret message in texts.

To improve the capacity of the coverless information hiding method, for several given keywords, we can search for the texts which contain these keywords, simultaneously. When the size of text database is large enough, several Chinese keywords can be hidden in one text. However, there is a problem that the number of keywords in each text would not be known. To address the issue, the coverless multi-keywords information hiding method is proposed. The main idea of this method is to hide the number of keywords in the same text by the same manner. Consequently, the receiver knows the number of keywords in each text, and the secret message can be extracted exactly from the texts.

3.1. Construction of Text Database

Natural texts are mainly collected by the following ways: fetching the news from the normal news web sites (such as Netease, Sina, and Baidu); downloading short articles from the writing platform (such as www.jianshu.com); collecting the various novels from free website and ancient literature. Through these ways, a large-scale text database whose size is 10.2G has been constructed, and the size of each text in this database is about 1KB. According to the main idea of coverless information hiding, to better meet the requirement of the method, the text database can also be expanded constantly.

3.2. Construction of Inverted Index Structure by Text Indexing

For a given keyword from a Chinese text, if we search for the texts which have the keyword exhaustively, it will cost huge time.

To reduce the huge cost of search time, an inverted index structure has been constructed [22]. And the Chinese character mathematical expression [23, 24] is used in the text indexing. When we index the text database, we will divide the Chinese keyword into various components by using the Chinese mathematical expression. We just adopt the first appearance of the components, and select the Chinese keyword behind these components as useful message. Figure 1 shows the index structure. The first part of the structure represents the serial number of the component which is named location tag. The second part shows Chinese keywords behind the tag, and the last part is the texts which contain the Chinese keyword. According to this structure, if we have new natural texts to expand the text database, the index files can be updated in time.

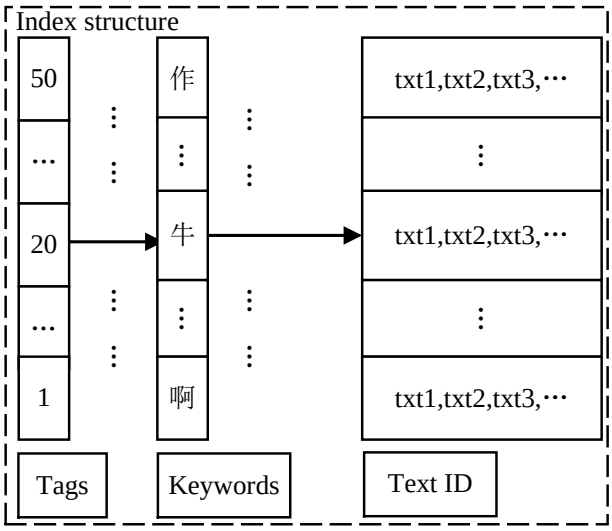

Figure 1. The Design of Index Structure

3.3. The Design of Scheme to Select Location Tags

In terms of the selection of location tags, the randomness, universalness and distinguishability of location tags should be taken into account [19].

The index files are analyzed statistically to select the components which satisfy the above requirements. According to the analysis of statistical results, 50 components are selected as the location label of the secret message. Figure 2 illustrates the Chinese keyword coverage of each component. The horizontal axis represents the serial number for each component, and the vertical axis represents the Chinese keyword coverage. 50 components can satisfy the requirements mentioned above better, because the coverage of different Chinese keywords behind these selected components are higher than others.

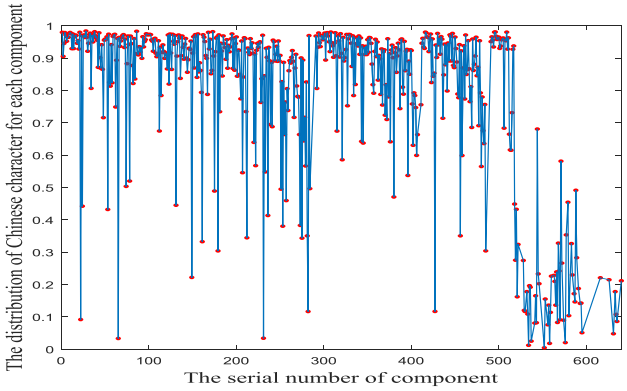

Figure 2. The Distribution of Keywords for Each Component

To describe the randomness of location tags, a scheme is designed to get different tags from the 50 tags by the user's identity. Users can use their particular ID to get their unique tags. If the length of secret message l is longer than the number of tags k, the tags will be used around s times. This ensures that one keyword will have one corresponding tag. The equation (1) is used to calculate s.

$$s = \left\lceil \frac{l}{k} \right\rceil \tag{1}$$

3.4. The Process of Information Hiding

Different from the coverless information hiding method based on Chinese mathematical expression [19], our method concludes two hiding operations. The first one is hiding the keywords of secret message in the text, and the second is hiding the number of the keywords.

In our method, a mapping table would be created to map the numbers of keywords to Chinese keywords. According to the statistical results, we found that a text can hide 5 keywords at most. Therefore, to facilitate the informaiton hiding, 5 keywords which are the most common in natural texts are chosen to correspond to the number of keywords. Table 1 displays some examples about the mapping relationships.

Table 1. Mapping Relationship

Number	Keywords
1	一
2	的
3	不
4	是
5	有

In the process of coverless information hiding, two lists of location tags need to be obtained, *i.e.*, L^1 and L^2. $L^1 \neq L^2$. When we hide the secret message, we use the L^1. When we hide the number of keywords, we use the L^2.

The text is considered as "tag + keyword" one by one. Then, we would begin to hide the secret message. For example, we assume to hide the combination of "亻+息". The process is shown as Figure 3.

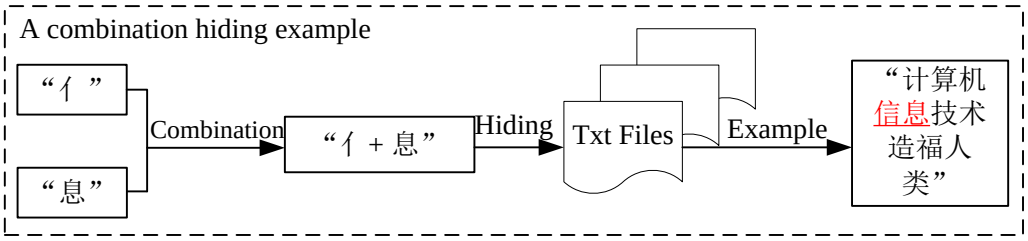

Figure 3. A Combination Hiding Example

If we know the user's ID and the secret message I_1, the process of the method is described as follows:

Hiding the secret message:

1. The secret message I_1 should be segmented into several keywords, denoted as $I_1 = \{W_1, W_2, \cdots, W_i, \cdots, W_l\}$ beforehand. Where, l is the number of keywords of I_1.
2. User's ID and the length of the secret message l are used to get a list of location tags $L_i^1 (i = 1, 2, \cdots, k \times s)$.
3. Get the list of combination of "tag + keyword", denoted as $H_i = L_i^1 + W_i (i = 1, 2, \cdots, l)$.

4. Search for the combination H_i in the index files to obtain all of the texts which contain the combination, and store these texts in a list.
5. Repeat 4). Search for all of the combinations, and get l lists of texts.
6. Compute the intersection of all lists and count the successful times. If the intersection is computed successfully, go on; if not, the intersection of the failed list and its next list are computed until to the end of all text lists. Finally, we will get a list of the texts T_1

Hiding the number of keywords:

7. Above all, a number can be obtained to record the number of keywords in each text, and map the number to a new secret message I_2.
8. Let I_2 as the secret message, and divide I_2 into several keywords, denoted as $I_2 = \{W_1, W_2, \cdots, W_j, \cdots, W_{l'}\}$. Where, l' is number of words of I_2.
9. User's ID and the length of the new secret message are used to get a new list of location tags L_j^2.
10. Get the list of combination of "tag + keyword", denoted as $H'_j = L_j^2 + W_j (j = 1, 2, \cdots, l')$.
11. Each combination of "tag + keyword" will get a list of texts T_2, and compute the intersection of two lists, T_1 and T_2. Then, we will obtain the final texts T_3.
12. Select a text randomly from each list in T_3, and send texts to the receiver.

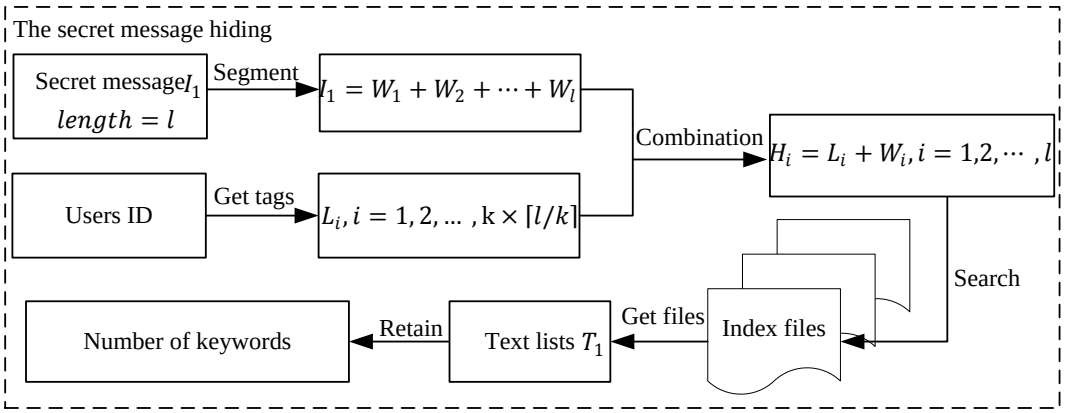

Figure 4. The Process of Secret Message Hiding

For a long secret message, if the process is finished, the texts will contain two types of message. One is the secret message, and the other is the number of keywords. Figure 4 illustrates the process of several keywords hiding using the proposed method.

When the process of secret message hiding is finished, we can hide the secret message successfully and get the number of keywords in each text. Then, the number of keywords will be mapped to several keywords. Therefore, we can hide the number of keywords in the same text by the same manner. Figure 5 shows the process of the number of keywords hiding.

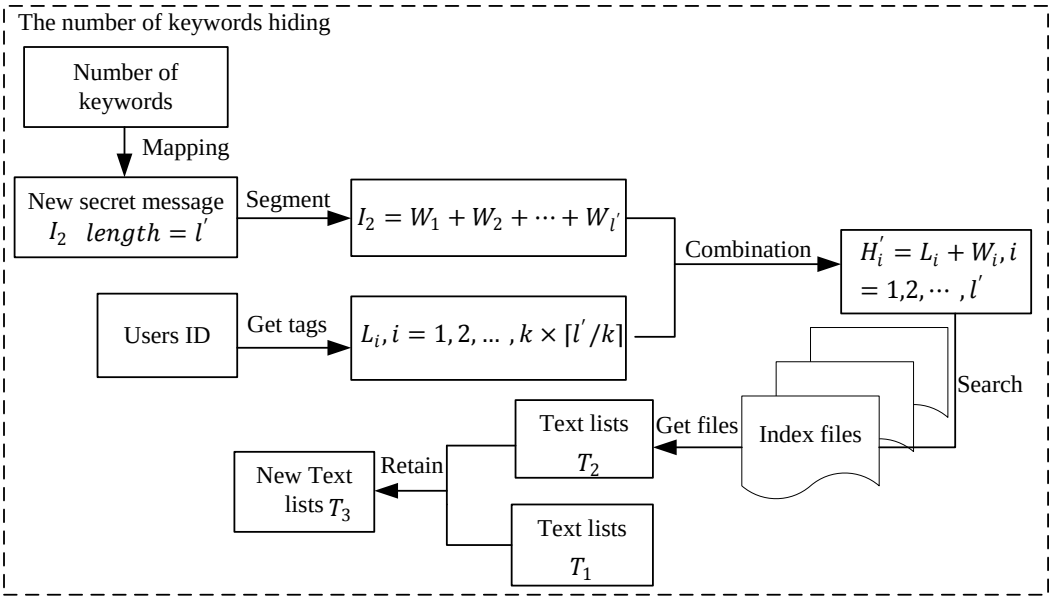

Figure 5. The Process of the Number of Keywords Hiding

3.5. The Process of Information Extraction

In the process of information extraction, each text needs to be extracted twice. Firstly, we can get the number of keywords based on user's ID and texts. Secondly, the secret message would be extracted from the texts according to the number of keywords, texts and user's ID. Through the above operations, the secret message can be extracted.

According to the texts and user's ID, the secret message would be extracted from texts. The process of the secret message extraction is shown as follows:

1) User's ID and the number of texts are used to get the location tags L_j^2.
2) According to the location tags L_j^2 and the texts T_3, the message I_2 which represents the number of keywords in each text will be extracted. Then, we map the message to the number of keywords by using the mapping table.
3) Based on the number of keywords, we can know that how many keywords in each text, and the length of the secret message l.
4) The same to step 1), we take the user's ID and the length of the secret message l to get a list of location tags L_i^1.
5) The location tags L_i^1, the number of keywords in each text and the texts are used to extract the secret message I_1 from these texts.

When the process is finished, the secret message would be extracted from the received texts. Figure 6 shows the process of information extraction.

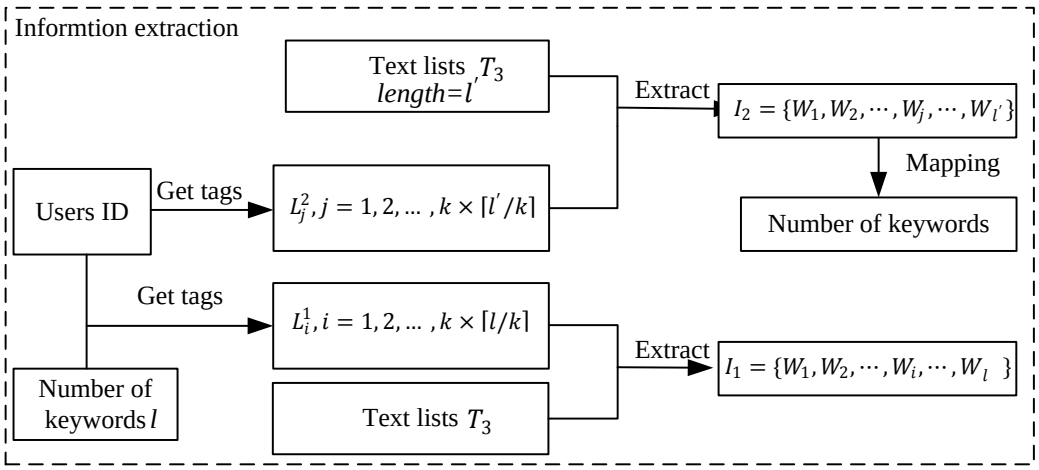

Figure 6. The Process of Information Extraction

4. Experimental Results

In this section, we will measure the hiding capacity of our method. The capacity is defined as the number of keywords hidden in one text. We assume that a secret message contains l keywords, and the message can be hidden successfully by using m texts. The capacity C can be defined by Eq. (2).

$$C = \frac{l}{m} \tag{2}$$

In our experiment, the mean capacity would be measured to get the overall performance of our method on capacity. We would measure the capacity for n secret message $C_i\,(i = 1, 2, \cdots, n)$ and compute the average of the capacity C_i, denoted as $\bar{C}$, which is can be calculated by Eq. (3).

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i \tag{3}$$

In the experiment, the coverless information hiding method based on Chinese mathematical expression, denoted as CME-CIHM [19] is compared with our method. 100 texts are adopted from the Sogou Labs [25] as secret messages, where the number of the keywords in each secret message ranges from 100 to 200. Then, we divide each secret message into several keywords, and hide all of their keywords into the texts by using our method. Thus, we can obtain the capacity for each secret message.

By using to the Eq. (2), the capacity of each secret message can be computed. Figure 7 shows the experimental results. The x-axis represents the serial number of each text, and the corresponding capacity is shown on y-axis. The result of hiding capacity for each secret message can be found from this figure. According to the results of experiment, the mean capacity can be calculated by the Eq. (3). Table 2 shows the results of the mean capacities of different methods.

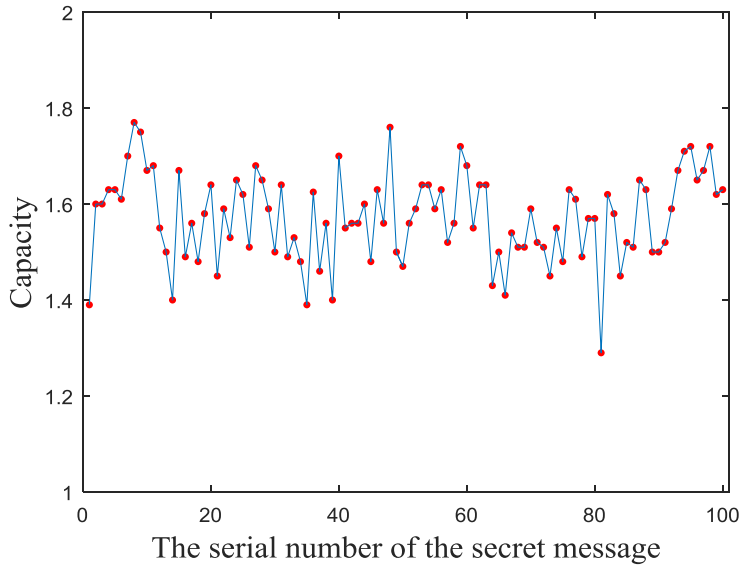

Figure 7. The Capacity for Each Secret Message

From Figure 7, we can see that our method's highest capacity of is 1.77, and the lowest capacity is 1.29. The hiding capacities for all the secret message is greater than 1. As shown in Table 2, the mean capacity of our method is 1.57. Thus, our method achieves higher capacity than CME-CIHM [19]. Consequently, compared with CME-CIHM, our proposed method can effectively improve the hiding capacity.

Table 2. The Mean Capacities of Two Different Methods

Methods	Mean Capacity
CME-CIHM (Chen et al., 2015)	1
Our method	1.57

5. Conclusion

In this paper, we propose a novel method, named as coverless multi-keyword information hiding method to improve the capacity of the coverless information hiding for the secret text. To prove the validity of our method, we select 100 texts from the Sogou Labs as the secret message to measure the capacity of the method. From the experimental results, we can see that the method can improve the capacity of coverless information hiding to some extent. Moreover, a mapping relationship is created to map the number into Chinese keywords. However, the improvement is limited. The main reason is that the underutilization of the text when we index the text database. Future work will focus on how to improve the utilization of text by building a more effective index files.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC) (61602253, U1536206, 61232016, U1405254, 61373133, 61502242, 61572258),Jiangsu Basic Research Programs-Natural Science Foundation (BK20150925, BK20151530), Startup Foundation for Introducing Talent of Nanjing University of Information Science and Technology (2014r024), Open Fund of Demostration Base of Internet Application Innovative Open Platform of Department of Education (KJRP1406, KJRP1407), and

Priority Academic Program Development of Jiangsu Higher Education Institutions (PADA) Fund, China.

References

[1] Z. Zhou, C. Yang, B. Chen, X. Sun, Q. Liu and Q. Wu, "Effective and efficient image copy detection with resistance to arbitrary rotation", J. IEICE Transactions on Information and Systems, vol. 99, no. 6, **(2016)**, pp. 1531-1540.
[2] Z. Zhou, X. Sun, X. Chen, C. Chang and Z. Fu, "A novel signature based on the combination of global and local signatures for image copy detection", J. Security & Communication Networks, vol. 7, no. 11, **(2013)**, pp. 1702-1711.
[3] Z. Fu, K. Ren, J. Shu, X. Sun and F. Huang, "Enabling personalized search over encrypted outsourced data with efficiency improvement", J. IEEE Transactions on Parallel and Distributed Systems, vol. 27, no. 9, **(2015)**, pp. 2546-2559.
[4] I. Cox and M. Miller, "The first 50 years of electronic watermarking", J. Journal of Applied Signal Processing, vol. 2002, no. 2, **(2002)**, pp. 126-132.
[5] N. Maxemchuk, "Electronic document distribution", J. At & T Technical Journal, vol. 73, no. 5, **(1994)**, pp. 73-80.
[6] J. Brassil, S. Low and N. Maxemchuk, "Copyright protection for the electronic distribution of text documents", J. Proceedings of the IEEE, vol. 87, no. 7, **(1999)**, pp. 1181-1196.
[7] J. Brassil, S. Low, N. Maxemchuk and L. Gorman, "Electronic marking and identification techniques to discourage document copying", J. IEEE Journal on Selected Areas in Communications, vol. 13, no. 8, **(1994)**, pp. 1495-1504.
[8] S. Low, N. Maxemchuk and A. Lapone, "Document identification for copyright protection using centroid detection", J. Communications IEEE Transactions on, vol. 46, no. 3, **(1998)**, pp. 372-383.
[9] S. Low and N. Maxemchuk, "Performance comparison of two text marking methods", J. Journal on Selected Areas in Communications, vol. 16, no. 4, **(1998)**, pp. 561-572.
[10] S. Low, N. Maxemchuk, J. Brassil and L. Gorman, "Document marking and identification using both line and word shifting", 14th Annual Joint Conference of the IEEE Computer and Communications Societies, Boston, MA, USA, **(1995)** April 2-6.
[11] Z. Xia, X. Wang, X. Sun, Q. Liu and N. Xiong, "Steganalysis of LSB matching using differences between nonadjacent pixels", J. Multimedia Tools & Applications, vol. 75, no. 4, **(2014)**, pp. 1-16.
[12] J. Mielikainen, "LSB matching revisited", J. IEEE Signal Processing Letters, vol. 13, no. 5, **(2006)**, pp. 285-287.
[13] Z. Xia, X. Wang, X. Sun and B. Wang, "Steganalysis of least significant bit matching using multi-order differences", J. Security & Communication Networks, vol. 7, no. 8, **(2014)**, pp. 1283-1291.
[14] Z. Li, X. Chen, X. Pan and X. Zeng, "Lossless data hiding scheme based on adjacent pixel difference", Proceedings of the International Conference on Computer Engineering and Technology, Singapore, Singapore, **(2009)** January 22-24.
[15] X. Li, and J. Wang, "A steganographic method based upon JPEG and particle swarm optimization algorithm", J. Information Sciences An International Journal, vol. 177, no. 15, **(2007)**, pp. 3099-3109.
[16] Z. Zhou, H. Sun, R. Harit, X. Chen and X. Sun, "Coverless Image Steganography Without Embedding", 1st International Conference on Cloud Computing and Security, Nanjing, China, **(2015)** August 13-15.
[17] H. Shahrokh and N. Mosayeb, "A novel LSB based quantum watermarking", J. International Journal of Theoretical Physics, **(2016)**, pp. 1-14.
[18] Z. Duric, M. Jacobs and S. Jajodia, "Information hiding: steganography and steganalysis", J. Handbook of Statistics, vol. 2005, no. 24, **(2005)**, pp. 171-187.
[19] X. Chen, H. Sun, Y. Tobe, Z. Zhou and X. Sun, "Coverless information hiding method based on the Chinese mathematical expression", 1st International Conference on Cloud Computing and Security, Nanjing, China, **(2015)** August 13-15.
[20] Z. Fu, X. Sun, Q. Liu and J. Shu, "Achieving efficient cloud search services: multi-keyword ranked search over encrypted cloud data supporting parallel computing", J. IEICE Transactions on Communications, vol. E98.B, no. 1, **(2015)**, pp. 190-200.
[21] Z. Xia, X. Wang, X. Sun, Q. Liu and N. Xiong, "A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data", J. IEEE Transactions on Parallel and Distributed Systems, vol. 27, no. 2, **(2015)**, pp. 340-352.
[22] Z. Zhou, Y. Wang, Q. Wu, C. Yang and X. Sun, "Effective and Efficient Global Context Verification for Image Copy Detection", J. IEEE Transactions on Information Forensics and Security, **(2016)**.
[23] Y. Liu, X. Sun, I. Cox and H. Wang, "Natural language information hiding based on Chinese mathematical expression", J. International Journal of Network Security, vol. 8, no. 1, **(2009)**, pp. 10-15.
[24] X. Sun, H. Chen, L. Yang and Y. Tang, "Mathematical representation of a Chinese character and its applications", J. International Journal of Pattern Recognition & Artificial Intelligence, vol. 16, no. 6, **(2011)**, pp. 735-748.
[25] Sogou Labs. [online] http://download.labs.sogou.com/ (Accessed 30 June 2016).

Authors

Zhili Zhou, He received his BS degree Communication Engineering from Hubei 2007, and his MS and PhD degrees in computer Application at the School of Information Science and Engineering from Hunan University, in 2010 and 2014, respectively. He is an Assistant Professor with the Nanjing University of Information Science and Technology. His current research interests include near-duplicate image/video detection, image/video copy detection, coverless information hiding, digital forensics, and image processing.

Yan Mu, He received his BE degree Internet of Things Engineering from Huaiyin Institute of Technology 2015. He is attending Nanjing University of Information Science and Technology to his MS degree Computer Science and Technology. His current research interests include near-duplicate image/video detection, and coverless information hiding,

Ching-Nung Yang, He received the B.S. and M.S. degrees in telecommunication engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1983 and 1985, respectively, and the Ph.D. degree in electrical engineering from National Cheng Kung University, Tainan, Taiwan, in 1997. He is currently a Full Professor with the Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien, Taiwan. His research interests include coding theory, information security, and cryptography.

Ningsheng Zhao, He received his BE degree Network Engineering from Nanjing University of Information Science and technology. He is attending Nanjing University of Information Science and Technology to his MS degree Computer Science and Technology. His current research interests include near-duplicate image/video detection, and coverless information hiding.