An Adaptive Steganographic Method Using Additive Noise

Lingyun Xiang[1*], Jiaohua Qin[2], Xiao Yang[1], Qichao Tang[1]

[1] College of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha, Hunan, China.
[2] College of Computer Science and Information Technology, Central South University of Forestry and Technology, Changsha, Hunan, China.

* Corresponding author. Tel.: +86 13874829742; email: xiangly210@163.com

Manuscript submitted July 25, 2015; accepted October 7, 2015.
doi: 10.17706/jcp.11.3.207-215

Abstract: This paper proposed an adaptive multi-ary steganographic method based on additive noise to ensure the security of secret information. In order to minimizing the distortion caused by embedding operations, the characteristics of multi-ary steganography by adding removable noise and multi-ary matrix embedding are analyzed, and then a combination-based coding method is proposed for embedding information by adding the fewest noises. It directly establishes a map between the vector of the secret information and the cover elements with minimum weight. According to the length of the secret information and the estimated embedding capacity, the steganography adaptively divides the secret information and the cover elements into blocks to make full use of the combination-based coding method. The experimental results show that the proposed steganography method obtains higher embedding efficiency and better ability of resisting statistical steganalysis attacks than the general method.

Key words: Steganography, additive noise, matrix embedding, steganalysis, embedding efficiency.

1. Introduction

The purpose of steganography is to implement covert communication by imperceptibly embedding secret information into multimedia object. The object used to carry information is called cover object, while the one embedded information is stego object. However, the embedding operations always cause statistical changes of cover objects, which hint at the existence of secret information. Generally speaking, the smaller difference between the stego and the corresponding cover object, the higher security of the secret information will achieve, as the probability of the successful steganalysis is lower. There exist two main steganographic approaches based on conclusion upon, one is to preserve the statistical characteristics of the stego object under a chosen cover model; the other is to minimize the embedding distortion usually measured by the embedding efficiency [1].

Some works have been done on statistical characteristics preservation. Desoky proposed a novel noiseless steganography paradigm (Nostega) [2], which embedded data into automatically created noiseless unquestionable data without using cover objects, and presented several Nostega-based methods, e.g. list-based steganography [3], which hid information in legitimate items(products, songs, books, etc.) by generating a stego text in a form of popular used textual list. The stego texts have good imperceptibility and high resistance against steganalysis. Reference [4] proposed to select a series of multiple choice questions to automatically generate the stego text for concealing information. The outputting stego texts kept the same linguistic and statistic characteristics as natural texts.

To improve embedding efficiency, the right solution is designing efficient steganographic codes [5]. Crandall [6] firstly proposed a steganographic coding method called matrix embedding, which applied covering codes to steganography. Afterwards, it was used by the famous steganographic algorithm F5 [7]. And there are many more known researches on steganographic methods based on linear codes and convolutional codes. In [8], syndrome coding was implemented to embed information using BCH codes based on a structured matrix or a generator polynomial. Recently, novel syndrome coding schemes using the called syndrome trellis codes (STCs) [9]-[10] was proposed based on dual convolutional codes equipped with the Viterbi algorithm. A complete practical framework with STCs was also described to embed information by minimizing an additive distortion function. Distortion of an embedding change on stego elements could be expressed in different function under different embedding scheme. The state-of-the-art steganographic algorithms can be designed by using STCs with distortion functions.

However, STCs are complex and inflexible. By analyzing the matrix embedding, this paper endeavors to construct a steganographic coding method to maximize the embedding efficiency for a special kind of cover elements. Taking into account that all the cover elements can be unified to a value after eliminating the noise, through adding extra removable noise, secret information can be embedded. Focusing on the additive noise, an adaptive multi-ary steganography is proposed by using a new combination-based coding algorithm. The combination-based coding method has low complexity and provides low distortion. Thereby, the stego object generated by the proposed method can achieve high security. Experiments show that the proposed steganographic method can greatly improve the ability of the additive-noise-based stego objects against statistical detection.

2. The Proposed Method

Here, the proposed steganographic method based on additive removable noise and noise combination-based coding will be introduced in detail.

2.1. Multi-ary Steganography by Adding Removable Noise

Currently, the most popular steganography is based on LSB for images. It is a binary steganography. The case that only two equal states of a cover element can be exchanged for steganography is binary. If each cover element in a cover object has multiple equal states to be used for embedding information, then it can be regarded as a multi-ary steganographic method. There are a lot of multimedia object in various types can be used for steganography as their diversity. For example, the method using single and double spaces to represent different values is a binary steganography; if the number of substituted synonyms is three, it is a ternary steganography; when the method is based on modifying the least two bit of the font color, it is a quaternary steganography.

Although a cover element has various states of being replaced each other, generally, there always exist dominant states in the view of the whole cover object, i.e. some states occupy a large proportion in a cover object, while the number of others are small, which can be regarded as negligible noise. Removing this kind of noise by replacing them with their equal dominant states would not observably change the appearance and content of the original object. From another point of view, by adding this kind of noise, the original state of the cover element can be altered to embed information while maintaining good visual imperceptibility. In this paper, this kind of method is generally named additive noise-based steganographic method.

Taking additive noise-based text steganographic methods as examples, they mainly include: the space-based [12], the tag case-based in HTML document [13], the font format-based [14], and so on. The space-based method has multiple ways of hiding information, the simplest one is to add different number of extra spaces to represent different values, for example, taking one space character to represent ‘0’ while two space characters to represent ‘1’. The extra space characters can be imperceptibly deleted or added. The tag case-based method alters the cases of the letters in tags of HTML documents to hide information as the fact that letters are always case-insensitive, for example, defining the uppercase letter as ‘0’ while the lowercase letter as ‘1’. The font format-based mainly makes imperceptible modification of font format in the document to hide information, for example, replacing the least several bits of the font color by the secret information. These replace operations generate new similar font colors without influencing the appearance of the cover document.

These additive noise-based methods have a common point that most frequently appearing states of the data would be changed to other rarely appearing similar states. Despite the rarely similar states are reasonable appearing in a cover object, their amount cannot be large as they are also regarded as noise. Many noises are easily attracted attention to be removed. In the view of steganography, some measures should be adopted to restrict the number of the added noises. More added noises means the secret information is lower security, which will attract more suspicious from the steganalysis. The impact of embedding modifications will be simply measured using the number of added noises and their intensity.

2.2. Multi-ary Matrix Embedding

We denote the multi-ary steganography as q-ary steganography for convenience. Secret information must be transformed into q-ary sequence before being embedded. The embedding process can be regarded as a coding problem in blocks. All cover elements are divided into blocks with the length of *n*. Set the cover elements C in every block is any subset of F_q^n, where F_q^n denotes n dimension vector space in a finite field F_q, each cover element has q number of states to be exchanged. When the embedding rate is less than 1, it is possible that a *n* dimension q-ary C is just required to embed a n-k dimension q-ary secret information. *n* cover elements have q^n possible states, while the *n-k* symbols of secret information just have q^{n-k} possible states, thereby, some different states of the n cover elements will represent the same state of the n-k symbols of secret information. In the embedding process, the number of modifications will be minimized if the states after embedding information are always closest to the original cover object.

Fridrich *et al.* [1] proposed and proved the matrix embedding using pixels as the cover elements can be extended to any type of cover data. Matrix embedding is based on linear block code and can be described as follows:

Let an (n,k) linear block code with a parity check matrix $\mathbf{H}$ and covering radius R. The embedding scheme below can communicate $n-k$ symbols in n pixels with pixel symbols $\mathbf{x}$ at most R changes.

$$Emb(\mathbf{x},\mathbf{m}) = \mathbf{x} + \mathbf{e}_{\mathrm{L}}(\mathbf{m} - \mathbf{Hx}) = \mathbf{y} \tag{1}$$

$$Ext(\mathbf{y}\ ) = \mathbf{Hy} \tag{2}$$

where $\mathbf{x} \in F_{\mathrm{q}}^n$ is a n-dimension vector representing the cover elements in a block, $m \in F_{\mathrm{q}}^{n-k}$ is a $n-k$ dimension vector representing the $n-k$ symbols of secret information. e_L is a coset leader of the coset for the syndrome $\mathrm{m}\text{-}H\mathbf{x}$. $\mathbf{y} \in F_q^n$ is a n-dimension vector corresponding to $\mathbf{x}$ having be embedded information. Note that $Ext(Emb(\mathbf{x},\mathbf{m})) = \mathbf{Hx} + \mathbf{He}_{\mathrm{L}}(\mathbf{m} - \mathbf{Hx}) = \mathbf{Hx} + (\mathbf{m} - \mathbf{Hx}) = \mathbf{m}$.

Embedding efficiency is defined as the expected number of embedded random message bits per one embedding change. From the Matrix embedding, we can find that the embedding efficiency is closely related to the covering radius of the used linear block code. Moreover, the expected number of embedding changes for random secret information in each block is equal to the average weight of all coset leaders, and the average distance to the code, too. Unfortunately, [1] have demonstrated that a code with the smallest average distance to code does not necessarily have the smallest covering radius.

Assuming the maximum weight of coset leaders is $R_{\max}$, the number of the coset leaders with weight i for code (n,k) is at most $C_n^i(q-1)^i$, where C_n^i represents the combinations of i elements from n, and the state of each element can be 0 to $q-1$. In best case, all the code words with weight $R_{\max}-1$ are coset leaders, then the total number of coset leaders N_c is:

$$N_c = C_n^0 + C_n^1(q-1)\ldots + C_n^{R_{\max}-1}(q-1)^{R_{\max}-1} + tC_n^{R_{\max}}(q-1)^{R_{\max}} = q^{n-k} \quad (3)$$

where $0<t\le 1$. N_c equals to the total number of all the states of the *n-k* symbols of secret information. It is possible that not all code words whose weight is less than $R_{\max}$ must be a coset leader, in other words, some code words are possible to have the shorter distance to code than the coset leader. In this time, the average distance R_a to the code will be minimum, which can be calculated as follows:

$$R_a \ge \frac{\sum_{i=1}^{R_{\max}-1} iC_n^i(q-1)^i + R_{\max}tC_n^{R_{\max}}(q-1)^{R_{\max}}}{N_c} = \frac{\sum_{i=1}^{R_{\max}-1} iC_n^i(q-1)^i + R_{\max}(q^{n-k} - \sum_{i=0}^{R_{\max}-1} C_n^i(q-1)^i)}{q^{n-k}} \quad (4)$$

Thus, the theoretical upper bound of the embedding efficiency $e_{\max}$ is:

$$e_{\max} = \frac{(n-k)\log_2 q}{\min(R_a)} = \frac{(n-k)q^{n-k}\log_2 q}{\sum_{i=1}^{R_{\max}-1} iC_n^i(q-1)^i + R_{\max}(q^{n-k} - \sum_{i=0}^{R_{\max}-1} C_n^i(q-1)^i)} \quad (5)$$

2.3. Combination-Based Coding Method for Steganography with Additive Noise

From the above analysis, the embedding efficiency of matrix embedding with a random linear block code (n,k) cannot reach the theoretical upper bound calculated in (5). And we have to using a parity check matrix to find coset leaders for steganography. If it randomly creates a random parity check matrix instead of careful construction, the obtained embedding efficiency may be non-ideal. Moreover, it is difficult of constructing linear block codes with arbitrary parameters n and k. In order to resolve the above problems, we proposed a combination-based coding method for embedding information by the additive noises.

Considering the particularity of the removable noise in cover object, it can build a map between n q-ary cover elements and $n-k$ q-ary secret information to obtain the maximal embedding efficiency of the matrix embedding based on linear block code (n,k). In other words, it builds a relationship between each combination of locating the minimum noises in the block with the secret information. This method has low complexity, little computation and low distortion.

Let $\mathbf{x}=[x_1,x_2,\ldots,x_n]^T$ denote n q-ary cover element, $\mathbf{m}=[m_1,m_2,\ldots,m_{n-k}]^T$ denote $n-k$ q-ary secret information. For steganography based on additive noise, the state of $\mathbf{x}$ can be initialized to only one case, i.e., $\mathbf{x}=[0,0,\ldots,0]^T$, but secret information $\mathbf{m}$ have q^{n-k} possibilities. In order to minimize the average number of embedding changes to reduce the amount of added noise, the q^{n-k} possible states of $\mathbf{x}$ with smallest weight should be chose for carrying the q^{n-k} secret information. Assuming that the maximum changes of each cover element is $R_{\max}$, which corresponds to the weight of $\mathbf{x}$, then the q^{n-k} states of $\mathbf{x}$ with the smallest weight consist of the all n-dimension vectors whose weight is 0,1,2, …, $R_{\max}-1$ and part of vectors whose weight is $R_{\max}$. Without considering the denoising operations on the cover object before embedding information, it is easy to find that the embedding efficiency of this coding method, named combination-based coding, is equal to the theoretical upper bound of matrix embedding; and it actually represents the average amount of noise introduced into each block. The state of the cover element is altered by adding noise.

Let coding function $f(\mathbf{m})=\mathbf{x}_s$, where $\mathbf{x}_s$ denotes the stego state of $\mathbf{x}$ after embedding information; inverse function $f^{-1}(\mathbf{x}_s)=\mathbf{m}$. According to the inverse function, the secret information $\mathbf{m}$ can be extracted from stego data $\mathbf{x}_s$.

The process of building the coding function can be described as: construct a one-to-one mapping of q^{n-k} states of $\mathbf{m}$ and all q^{n-k} number of *n* dimension vectors whose weight is less than $R_{\max}$ and part *n* dimension vectors whose weight is $R_{\max}$. Therefore, firstly, all the *n* dimension vectors be used to embed information should be ordered in a certain order. Each *n* dimension vector will be assigned to a *n-k* dimension vector, which is the value of secret information $\mathbf{m}$.

The order of the used *n* dimension vector is described as follows: $[0,0,\ldots,0]^T$, all the n dimension vectors whose weight is 1 (in an ascending order), all the n dimension vectors whose weight is 2 (in an ascending order),…, all the *n* dimension vectors whose weight is $R_{\max}-1$ (in an ascending order), $q^{n-k}-\sum_{i=0}^{R_{\max}-1} C_n^i (q-1)^i$ number of *n* dimension vectors whose weight is $R_{\max}$ (in an ascending order). The order of *n-k* dimension secret information can be an ascending order.

As the number of the used *n* dimension vectors is the same as that of the *n-k* dimension secret information. Therefore, we can construct a one to one mapping between them in the above orders. Table 1 shows an example of building map between secret information and cover elements, when $n=4, k=1, q=2$. The number of the secret information's states is 8, thus 8 states of cover elements with smallest weight are chose to carry secret information. The secret information are arranged in ascending order.

Table 1. The Map between Secret Information and Cover Element

Secret information	Cover element
000	0000
001	0001
010	0010
011	0100
100	1000
101	0011
110	0101
111	0110

It is worthwhile to note that one of the q^{n-k} secret information can be adjusted to be coded to another state of cover element, which is determined the actual used coding function. The coding function can be easily constructed, and the value of n and k can be arbitrarily set. The position of the n dimension q-ary cover elements with nonzero value will be added noise for embedding secret information. The number of the nonzero positions in the stego elements will measure the distortion caused by embedding modifications. Since we choose the combinations of the minimum noise to carry the secret information, the distortion will be minimized.

2.4. The Adaptive Multi-ary Steganographic Method Using Combination-Based Coding

The additive noise-based steganography would cause the amount of noise increasing. More noise will arouse suspicion leading to weak security. In order to apply the combination-based coding method into additive-noise-based steganography, the sequence of the cover elements is divided into blocks whose length is *n*, and the sequence of secret information converted into q-ary symbols is divided into blocks with length $n-k$. Thus, two parameters *n* and *k* should be set. In this paper, $n-k$ is assigned to a pre-determined appropriate value, and n is adaptively determined by $n-k$, the secret information and the embedding capacity provided by the cover object .

Based on the above analysis, a novel adaptive steganography algorithm is proposed. The details are as follows.

Step 1: according to the characteristics of the cover object, transform secret information into a q-ary sequence with length N.

Step 2: according to N and the total embedding capacity L, and a pre-determined $n-k$, select a block parameter n subjecting to $\frac{n-1-k}{n} < \frac{N+n-k}{L} \le \frac{n-k}{n}$, n is the size of cover data blocks, $n-k$ is the size of secret information block.

Step 3: build the coding function f according to the combination-based coding method.

Step 4: divide the cover elements and secret information into blocks, for each block of the cover elements $\mathbf{x}$ with length n, and initialize it to $\mathbf{x}=[0,0,\ldots,0]^T$, and the corresponding block of the embedding information $\mathbf{m}$ with length $n-k$, calculate $f(\mathbf{m})=\mathbf{x}_s$, and modify $\mathbf{x}$ to $\mathbf{x}_s$ by adding noise.

Extracting the secret information is directly to obtain the states of the cover data in each block and calculate the secret information $\mathbf{m}$ by $f^{-1}(\mathbf{x}_s)=\mathbf{m}$.

3. Experimental Results and Discussion

3.1. The Analysis of the Embedding Efficiency

In this experiment, just $q=2$ is selected. According to formula 5, the embedding efficiency of the proposed method with arbitrary (n,k) can be calculated. Thus, the actual embedding efficiency of this method is determined when parameters n,k are known.

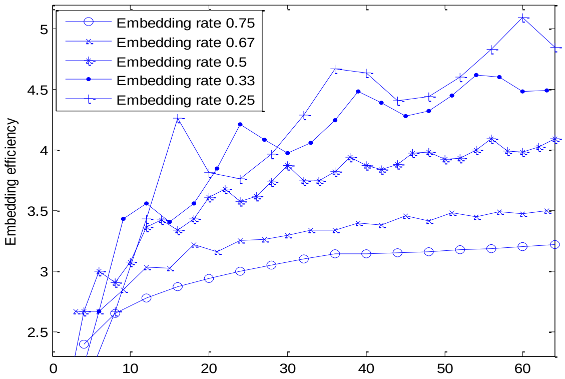


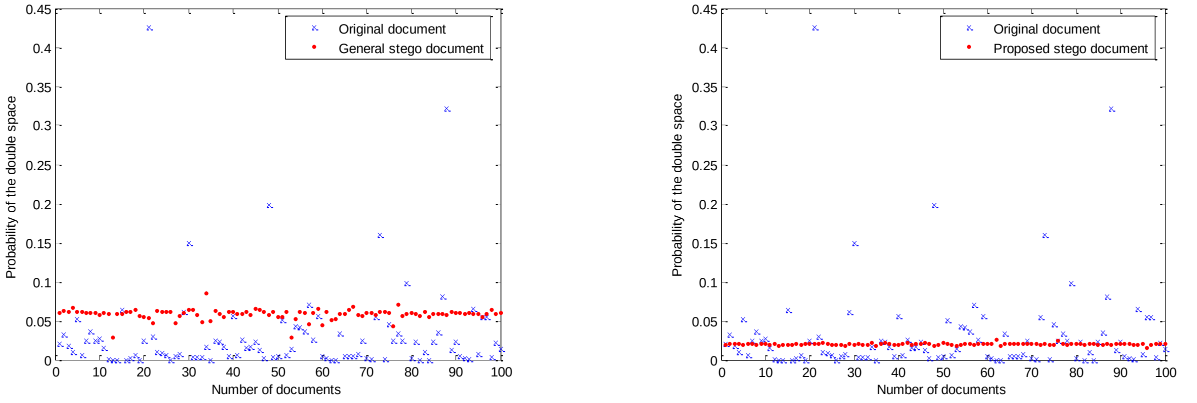
Fig. 1. The embedding efficiency of the noise combination-based coding with different embedding rate.

Fig. 1 shows distribution of the embedding efficiency provided by the combination-based coding (n,k) in different embedding rate. Actually, the embedding rate almost equal to $^{n-k}\!/_{n}$. We can see under common circumstances, the larger n, the higher embedding efficiency with the same embedding rate. The lower embedding rate, the greater embedding efficiency growing as n increasing, and when embedding rate gets higher, the changes of embedding efficiency will flatten out. Thus, for the same cover object and secret information, a large parameter n should be chose for reducing the amount of the added noise.

3.2. The Resistance of Steganalysis Attacks

Here, we will verify the resistance of steganalysis attacks of the proposed steganographic method. In experiments, we choose the simple space-based steganography, which embeds information by adding extra space characters. First, we download 432 word documents from Google Scholar, and 99 English documents from Google. For these English documents, stego documents were generated through two ways using nearly 12% embedding rate. One way is use one space to code “0”, double spaces to code “1” to embed information directly. These stego documents were recorded as general stego documents. The other way used the proposed method with $n = 100$, $n - k = 12$ to embed information and the results are recorded as proposed stego documents. In the embedding process, two ways embedded the same secret information into the same document, and the length of secret information ranges from 12 bit to 8280 bit.

Fig. 2 shows the probability distribution of extra spaces in original, general stego and proposed stego documents. In the original natural documents, the probability of double spaces is relatively small, but is unstable; it varies within a certain range. The probability of the general stego document is relatively large, which is easy to take the advantage of the probability to identify this type stego documents from the normal ones. The probability of the proposed stego document is significantly lower than that of the corresponding general stego one, and it is limit in the range of normal document. So, distinguishing normal document and proposed stego documents becomes much more difficult.



(a) Original and general stego documents (b) Original and proposed stego documents

Fig. 2. The probability distribution of double spaces in different texts.

Table 2. The Steganalysis Results of Different Documents

Steganalysis method		Original documents	Stego documents
based on the probability of double spaces	Original documents	430	101
	General stego documents	4	527
	Proposed stego documents	531	0

In experiments, we suppose steganalysis attacks classifying the normal documents and stego ones by the probability of double spaces. Set threshold equal to 0.04. If the probability of double spaces appearing in a detected document is larger than 0.04, and then judging the document is a stego one, otherwise, determining it is a normal one.

Results of steganalysis are illustrated in Table 2. Even if the threshold is chosen to cause normal documents having a high false positive rate, but most general stego documents can still be successfully detected, while nearly no proposed stego document is successfully detected. Combination-based coding method for steganography presented in this paper can greatly improve the resistance of against steganalysis and the security of secret information.

4. Conclusion

Most linear block codes cannot reach the theoretical upper of embedding efficiency no matter how excellent the constructed parity check matrix is. But considering the particularity of additive noise based steganographic method, the matrix embedding is modified. By directly establishing map between the secret information group and the cover element vector with minimum weight, the noise added by embedding information is minimized to make the embedding efficiency achieve the theoretical upper of the linear block codes with same parameters. Embedding efficiency at this time describes the increased degree of noise per each cover block. Finally, example experiments were conducted in English documents by adding space to embed information. Comparing the proposed method with general space-based steganographic method, when the embedding rate is the same, the embedding efficiency of the former is much larger than that of the latter, so that the distribution probability of the extra space in the documents generated by the former is much lower than those of the latter, the former can success to survive from the attack using this distribution probability, while the latter cannot. Experiments show that the proposed method can greatly improve the ability of the additive noise based stego documents against statistical detection.

Acknowledgement

This project is supported by National Natural Science Foundation of China (Nos. 61202439, 61202496), Hunan Provincial Natural Science Foundation of China (Grant No.13JJ2031).

References

[1] Fridrich, J., Lisonek, P., & Soukal, D. (2006). On steganographic embedding efficiency. *Proceedings of 8th International Workshop on Information Hiding* (pp. 282-296). Alexandria USA.

[2] Desoky, A. (2008). Nostega: A novel noiseless steganography paradigm. *Journal of Digital Forensic Practice*, *2(3),* 132–139.

[3] Desoky, A. (2009). Listega list-based steganography methodology. *International Journal of Information Security*, Springer, *8(4),* 247–261.

[4] Xiang, L., Sun, X., Liu, Y., & Yang, H. (2011). A secure steganographic method via multiple choice questions. *Information Technology Journal*. *10(5),* 992-1000.

[5] Zhang, W., & Li, S. (2008). A coding problem in steganography. *Designs, Codes and Cryptography*, *46(1),* 67-81.

[6] Crandall, R. (2007). Some notes on steganography. steganography mailing list. From: http://os.inf.tu-dresden.de/westfeld/crandall.pdf

[7] Westfeld, A. (2001). F5 — A steganographic algorithm. *Proceedings of 4th International Workshop on Information Hiding* (pp. 289-302). Lecture Notes in Computer Science, Pittsburgh, PA, USA.

[8] Schönfeld, D., & Winkler, A. (2006). Embedding with syndrome coding based on BCH codes. *Proceedings of the 8th Workshop on Multimedia and Security* (pp. 214–223). ACM.

[9] Filler, T., Judas, J., & Fridrich, J. (2010). Minimizing embedding impact in steganography using trellis-coded quantization. *Proceedings of Media Forensics and Security III*. SPIE 7451:715405–1.

[10] Filler, T., Judas, J., & Fridrich, J. (2011). Minimizing additive distortion in steganography using syndrome-trellis codes. *IEEE Transactions on Information Forensics and Security*, *6(3)*, 920–935.

[11] Feng, B. W., Lu, W., & Sun, W. (2015). Secure binary image steganography based on minimizing the distortion on the texture. *IEEE Transactions on Information Forensics and Security,10(2),* 243-255.

[12] Por, L. Y., Wong, K., & Chee, K. O. (2012). UniSpaCh: A text-based data hiding method using Unicode space characters. *Journal of Systems and Software*, *85*, 1075-1082.

[13] Yang, Y. J., & Yang, Y. M. (2010). An efficient webpage information hiding method based on tag attributes. *Proceedings of 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery* (pp. 1181-1184). Yantai, China.

[14] Bhaya, W., Rahma, A. M., & Al-Nasrawi, D. (2013). Text steganography basedon fong type in Ms–word documents. *Journal of Computer Science*, *9(7)*, 898-904.

Lingyun Xiang was born on April 15, 1983. She received her BE in computer science and technology, in 2005, and the PhD in computer application, in 2011, Hunan University, Hunan, China.

She is currently a lecturer at College of Computer and Communication Engineering, Changsha University of Science & Technology, Hunan, China. Her current research interests include information security, steganography, steganalysis, machine learning, pattern recognition and computer vision.

Jiaohua Qin received her BS in mathematics from Hunan University of Science and Technology, China, in 1996; and her MS in computer science and technology from National University of Defense Technology, China, in 2001; and her PhD in computer applications at the School of Computers and Communication of Hunan University, China, in 2009.

She is currently a professor at the College of Computer Science and Information Technology, Central South University of Forestry and Technology, China. Her research interests include steganography and steganalysis, image processing, and pattern recognition.

Xiao Yang was born in 1990. She received her BE in communication engineering from Tongda college of Nanjing University of Posts and Telecommunication, China, in 2012; and her MS in communication and information system from Changsha University of Science and Technology, China, in 2015. Her research interests include information security and steganography.

Qichao Tang was born in 1994. He is currently pursuing his BE in network engineering at Changsha University of Science and Technology, China. His research interests include information security and big data analysis.