International Journal of Digital Crime and Forensics
Volume 13 • Issue 6

A Coverless Text Steganography by Encoding the Chinese Characters' Component Structures

Kaixi Wang, College of Computer Science and Technology, Qingdao University, China*
https://orcid.org/0000-0001-8356-540X

Xiangmei Yu, College of Computer Science and Technology, Qingdao University, China

Ziyi Zou, College of Computer Science and Technology, Qingdao University, China

ABSTRACT

The current coverless text steganography methods have a low steganographic capacity, and yet some of them cannot assure a message can be concealed. How to achieve a high steganographic capacity has become the research hotspot in text steganography. This paper proposes a text coverless steganography method by encoding the Chinese characters' component structures. Its main idea is that a binary bit string can be conveyed by the Chinese characters' component structures. The positions of Chinese characters that carry a secret message will be expressed in two systems of the linear remainder equations, whose solutions will be secretly sent to the receiver to extract the secret message. In the method, a single Chinese character can express p bits. The analyses and statistics show that its capacity will be much higher when the same Chinese character is used more than once than existing methods, and it can conceal any message successfully. In addition, this method can also be employed in other languages.

KEYWORDS

Chinese Remainder Theorem, Coverless Steganography, Information Hiding, Steganography by Cover Search

INTRODUCTION

Images, audios, and videos have been widely employed as carriers for steganography (Fridrich, 2009; Lazic & Aarabi, 2006; Mastafa et al., 2017). Compared with these media, a text usually has few redundancies to be a carrier, and this leads to the difficulty of achieving text steganography and small steganographic capacity. But a text is widely used on the Internet, which makes it not prone to be suspected of a carrier and this is an indispensable feature of best steganography carriers (Zielińska et al., 2014). Therefore, how to accomplish text steganography and how to achieve a high capacity have become a research hotspot nowadays.

The traditional text stego-systems (Bennett, 2004; Richard, 2004) are mainly implemented by modifying a text by leveraging its redundancy in text formats, the characters' appearances, text syntax or semantics, and so on. Due to the aforementioned issues, the emerging trends in text steganography is

DOI: 10.4018/IJDCF.20211101.oa4 *Corresponding Author

to generate a new text or select a text to be a carrier, which is usually called "coverless text information hiding" (Chen et al., 2015; Ali, 2018; Wang, 2019), instead of modifying a text. Herein, "coverless" doesn't mean no carrier to convey a secret, but actually, it performs no modification to a carrier. Compared with steganography by modification, the coverless steganography is free of comparison attacks, what's more, the steganography by selection doesn't suffer from semantic attacks, which will be further studied in the following.

This paper proposes a new coverless text steganography method by encoding Chinese characters' component structures (CCCSs). A Chinese character usually consists of single or multiple components, which are assembled to form different CCCSs. These structures can be grouped according to their usage frequencies and every group will be assigned to express a different Binary Digital String (BDS). The characters in a text are organized in a square matrix which is reshaped to get the Minimal Square Matrix (MSM) who should include all the structure groups. The MSM will be further transformed into a Code Square Matrix(CSM). A secret message is converted into a secret BDS which will be split into binary digital slices (BiDSs) that are elements in the CSM; the row and column numbers of these slices in the CSM are organized to construct two systems of linear remainder equations, which can be solved according to the Chinese remainder theorem, and their solutions are taken as the keys shared with the receiver. The receiver figures out the corresponding row and column numbers by calculating the remainder using the shared key, and gets the slices, then reconstructs the complete BDS, finally converts it to the original secret.

RELATED WORKS

In recent years, many text steganography methods have been proposed. From the perspective of the embedding operation, these existing methods can be divided into two categories: steganography by modification and steganography by non-modification (Zhang et al., 2016). And the former can be further divided into the following three sub-categories: (1) Text Steganography by Format-based Modification (TSFM). TSFM makes full use of the imperceptible changes in a formatted text to represent binary digits. Such subtle changes include: shifting the specific matra towards left or right (Changder et al., 2009) or changing the special strokes of the characters(Xiao et al., 2018); changing the font style (Bhaya et al., 2013), the font color (Alsaadi et al., 2018), the brightness (Ou et al., 2007); widening or shortening the spaces between words or lines, also named open space method (Alattar & Alattar. 2004; Bender et al., 1996); using the different appearance of the punctuation marks in different languages, such as the comma, the period in Chinese or English (Popa, 1998), and so on. When such a stego-text is being transmitted, it is highly susceptible to the attacks such as re-formatting, retyping or saving in another format, that is, these methods have poor robustness. (2) Text Steganography by Modification based on Invisible Characters (TSMIC). TSMIC takes full advantage of the invisible characters defined in Unicode and ASCII codes, such as "control character", "escape character", to conceal a secret message. For instance, Cui et al. (2016) employed the communication-specific character SOH, which looks the same as blanks, to accomplish steganography, in which a blank indicates "0", and SOH is "1". Lee and Tsai (2008) used "typed space", " ", " ", "&#32", " ", " ", " ", " " to represent "000", "001", "010", "011", "100", "101", "110", "111" separately. If an attacker adds blanks or some invisible characters to a stego-text generated by this method, it is difficult for the receiver to extract the secret correctly. This method is not suitable for a paper file yet. (3) Text Steganography by Semantics-based Modification (TSSM). TSSM conceals a secret message via semantic substitutions in a text. The classical method is synonym-based steganography (Gan et al., 2007) who chose words from a shared synonym corpus to hide the secret. In addition, borrowed characters or word variants can also be employed. All these methods might replace the original words, which could lead to bad readability or change the style of the writing. And once the original text is got, the concealment traces could be easily found by comparison. Therefore, these methods might have poor imperceptibility and its steganography

capacity would be very small because of such poor imperceptibility, or the small size or quantity of the synonym corpus. Moreover, the modification in the above-mentioned methods could be detected easily by the corresponding steganalysis tools.

To solve these issues, many researchers proposed new steganography methods with no modification to text, which is named as Coverless Text Steganography (CTS). The CTS methods include: steganography by cover search, steganography by cover generation, and steganography by cover index. These methods are summarized as follows: (1) Steganography by Cover Search (SCS). The SCS methods will search for a suitable webpage as a carrier on the Internet (Shi et al., 2016). This webpage must contain all the characters of the secret message. The positions of those characters on the webpage will be encoded, then embed the result into the URL of the webpage. (2) Steganography by Cover Generation (SCG). Chang and Clark (2010) employed the N-gram model to generate anew text as the carrier. NiceText system is another typical example (Chapman & Davida,1997), in which a secret message would convert into other sentences based on a large code dictionary. Luo and Huang (2017) employed the RNN Encoder-Decoder structure to generate Chinese poetry. High-quality text covers can also be generated to hide a secret bitstream based on the RNN (Yang et al., 2019). (3) Steganography by Cover Index (SCI). Zhou et al. constructed a text corpus from the Internet that includes novels, news, essays, etc., then split each text into many words and built an index for each word, whose structure is expressed as "label + keyword". The embedding process is to find a set of texts whose word index corresponds to the index of the words in a secret message, and the qualified texts will be sent to the receiver. To sum up, The SCS has a contradiction between the high steganography capacity and good concealment. For the (SCG), because the NLP technology is not very perfect when a secret message is long, the imperceptibility is poor, what's more, some semantic problems will occur, including the syntax errors and bad readability, especially when a long text is generated. For the SCI, a text can only conceal a few characters, moreover, there may be no suitable text.

In short, the current CTS methods have a low steganography capacity or bad imperceptibility, and even some secret messages cannot be concealed successfully by some aforementioned methods. In order to solve these problems, this paper proposes a new CST method by encoding the Chinese characters' component structures. The method does not modify a text in any way, and easily achieves the steganography successfully in any case, and has a relatively high steganographic capacity.

ANALYSIS AND CODING OF CHINESE CHARACTERS' COMPONENT STRUCTURES

Chinese characters' component structures refer to the composition of the Chinese characters' components (Fu, 1991). A single-component character has only one Chinese character structural component and it is directly constituted by strokes, and single-component characters are defined as "Single structure"; a compound Chinese character has more than one component, and it can be further divided into eleven categories. All these structures are shown in Table 1.

Table 1. The Chinese characters' component structures classification

Structures	Examples	Structures	Examples
Single	"一" "不"	Upper-Left-Surround	"庙" "房"
Left-Right	"解" "结"	Lower-Left-Surround	"建" "连"
Upper-Lower	"置" "昌"	Upper-three-edge-Surround	"同" "问"
Left-Middle-Right	"辩" "脚"	Lower-three-edge-Surround	"凶" "函"
Upper-Middle-Lower	"熏" "亵"	Left-three-edge-Surround	"区" "匣"
Upper-Right-Surround	"句" "司"	Whole-Surround	"国" "回"

For the 3500 common characters (Ministry of Education & National Working Committee on Languages and Writing Systems, 2013), the quantities of characters and their usages of each structure are listed in Table 2.

Table 2. The quantity and the usage rate of common characters' component structures

Structure	Quantity	Proportion	Usage rate
Single	317	9.06%	32%
Left-Right	1892	54.06%	38%
Upper-Lower	777	22.20%	18%
Left-Middle-Right	170	4.86%	1.4%
Upper-Middle-Lower	73	2.08%	1.2%
Upper-Right-Surround	20	0.57%	0.22%
Upper-Left-Surround	114	3.26%	1.5%
Lower-Left-Surround	85	2.42%	3.61%
Upper-three-edge-Surround	27	0.77%	0.93%
Lower-three-edge-Surround	2	0.06%	0.0087%
Left-three-edge-Surround	9	0.26%	0.14%
Whole-Surround	14	0.4%	1.05%

The main idea of the paper is to encode these structures to express different BiDSs. To ensure that a secret message can be concealed successfully, the selected text must contain all these structures. However, Table 2 shows there are only a few Chinese characters with some structure, so it would be difficult to find a text containing a character with it, which could lead to the low efficiency of steganography. Therefore, it is necessary to regroup these structures to make the appearance of each structure as even as possible. This will make it easy to find a text that contains characters with all the regrouped structures and make its length as short as possible.

According to the numbers and the usage listed in Table 2, the regrouping is performed as follows: (1) the Single-component structure is split into the Single-odd structure and the Single-even structure based on the parity of the stroke number of a Chinese character; (2) the Upper-Lower structure is divided into the Upper-Lower-odd structure and the Upper-Lower-even structure. (3) For the Left-Right structure, the statistics show that the characters with 1:1 ratio in the size of the left and right

parts account for 11.56%, and the remaining characters could be split into the Left-Right-odd structure (14.23%) and the Left-Right-even structure (13.10%); (4) all other 9 classes with low usage rates are merged into another category named "Others". Now, new eight types are defined and their numbers and usage are shown in Table 3. Other Classifications may be applicable but are not listed here.

Table 3. The quantity and the usage after recombination

Structure	Quantity	Proportion	Usage rate
Single-odd	157	4.48%	16.20%
Single-even	160	4.57%	16.70%
Left-Right-1: 1	260	7.42%	11.56%
Left-Right-odd	837	23.91%	14.23%
Left-Right-even	792	22.62%	13.10%
Upper-Lower-odd	383	10.94%	8.86%
Upper -Lower-even	395	11.28%	9.26%
Others	516	14.74%	10.06%

These new eight different types can be encoded to express different BiDSs as shown in Table 4.

Table 4. The codes of structures

Structures	codes
Single-odd	000
Single-even	001
Left- Righ-1: 1	010
Left-Right-odd	011
Left-Right-even	100
Upper-Lower-odd	101
Upper-Lower-even	110
Others	111

THE COVERLESS TEXT STEGANOGRAPHY BASED ON CHINESE CHARACTERS' COMPONENT STRUCTURES

The general steganography procedure is described as follows:

Firstly, according to the appearance frequencies of the Chinese characters, their structures will be divided into 2^p categories, e.g. 4, 8, and each Chinese character structure can represent p-bit BiDSs. For example, 2^2 categories Chinese characters structures can be used to represent "00", "01", "10", "11".

Secondly, a text is selected only if it contains all these 2^p -category characters. Every character in the text is regarded as an element of a square matrix, thus the text can be converted into a CSM according to the structures' encoding.

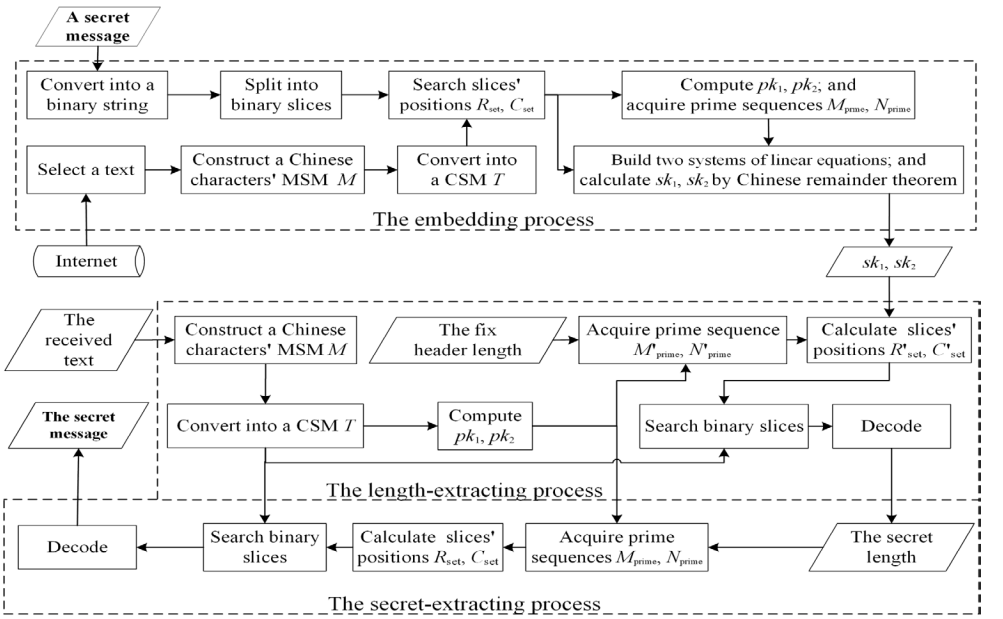
Thirdly, the BDS corresponding to the secret, denoted by S_{secret}, is acquired by encoding and/or encryption, and its length, denoted by L_{secret}, will be added to S_{secret} as the header, which is L_{secret}-bit length, thus a package S_{msg} is formed and its length L_{msg} is the sum of L_{header} and L_{secret}. Finally, S_{msg} will be split into many *p*-bit slices that are elements in the CSM; if the last slice is less than *p*-bit, one or more digit "0" is appended to construct a *p*-bit slice. All these BiDSs of the secret message are lined up into the BiDS Sequence $\left\{SEG_1, SEG_2, ..., SEG_{Pnum}\right\}$, herein, $Pnum = \frac{L_{msg}}{p}$.

Fourthly, two prime sequences $M_{Prime} = \left(m_1, m_2, ..., m_{Pnum}\right)$ and $N_{Prime} = \left(n_1, n_2, ..., n_{Pnum}\right)$ will be calculated out based on the first appearance of the 2^p -category BiDSs in the CSM, which will be described in detail later. The corresponding row number sequence $R = \left(r_1, r_2, ..., r_{Pnum}\right)$ and the corresponding column number sequence $C = \left(c_1, c_2, ..., c_{Pnum}\right)$ of the BiDS sequence $\left\{SEG_1, SEG_2, ..., SEG_{Pnum}\right\}$ in the CSM will be figured out. Two systems of the linear remainder equations will be constructed by taking R, C as the remainder and M_{Prime}, N_{Prime} as the divisor respectively, and its solution, i.e. sk_1, sk_2 will be calculated by the Chinese remainder theorem respectively.

The receiver will figure out the corresponding positions R, C according to sk_1, sk_2 and the same prime number sequences M_{Prime} and N_{Prime}. The secret BDS will be reconstructed and further converted into the original secret.

The above process is illustrated in Figure 1.

Figure 1. The overall steganography procedure



In the following, taking $p = 3$ as an example to analyze the Chinese characters and describe this method in detail.

Selection of a Text Carrier

No matter how long a text is, as long as it contains all eight types of structures defined in Table 4, it can conceal any secret message successfully. Therefore, the primary condition that a text could be a carrier is that it must include all eight types of structures. In fact, only the foremost part of a text, which contains all those eight types of structures, will be referred. Generally, the square matrix, constructed with the foremost part of a text that contains all those 8 kinds of structures, will be smaller than that based on the whole text. Therefore, in order to reduce the computation of two linear systems, the length of this part should be as short as possible.

On the other hand, a 3-bit BiDS may match multiple elements in the CSM because the square matrix may contain the Chinese characters with the same structure. And the element will be chosen to represent the BiDS when its row number r_p is minimum, if there is more than one element whose r_p is equal, and the column number c_p will be selected in the similar way. This means the code in the CSM will be reused.

Let N is the length of the foremost part of a text that contains all those 8 kinds of structures. Texts are crawled from the Internet and organized in different topics, such as news, science, novels, e-commerce. The statistics for the minimum and maximum N are listed in Table 5.

As illustrated in Table 5, the minimum N is 9, and 86 at most. There are a large number of texts that have more than 86 Chinese characters on the Internet. That is, a large number of qualified texts can be used for the proposed steganography.

Table 5. The statistics for the minimum and maximum N

Field	Text number	Minimum N	Maximum N
News	26	10	43
Novel	19	9	73
E-commerce	27	16	86
Technology	22	11	67
Science	22	9	67
Prose	21	15	42

The Code Matrix of Chinese Characters' Component Structures

As described above, the first N characters containing the eight types of structures in Table 4 will be stored in the MSM M from the top-left to bottom-right in the same order as in the text. The size of M is $n * n$, where $n = \sqrt{N}$. The matrix M is illustrated as follows.

$$\mathrm{M} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \tag{1}$$

where, a_{ij} is one of the first N characters in the text, and $i, j \in [1, n]$.

The matrix M is then converted into the CSM T in the light of the mapping defined in Table 4 as follows.

$$T = \begin{vmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{vmatrix} \tag{2}$$

where $b_{i,j} \in \{000, 001, 010, 011, 100, 101, 110, 111\}$ and every binary slice in the set is defined in Table 4.

The Preprocessing of the Secret Message

In general, a secret is usually converted into a BDS before being embedded into a carrier. This can be done in a variety of ways, such as encryption, encoding. For example, a message can be simply transformed into a BDS by using its corresponding machine internal code or ASCII. Besides the transformation, some other preprocessing should be carried out. The specific procedures can be summarized as follows.

(1) Transform a secret into the BDS S_{secret}.
(2) Compute the length of S_{secret}, denoted by L_{secret}.
(3) Convert the value L_{secret} to a BDS indicated as L_s and it will be added as the header whose length is 16 bits, then a package S_{msg} is formed, and its whole length is denoted as L_{msg}.
(4) S_{msg} will be divided into many BiDSs $\{SEG_1, SEG_2, ..., SEG_{Pnum}\}$. Each slice includes 3-bit BDS and the last slice may be appended by "0" until it reaches 3 bits.

The General Solution to the Chinese Remainder Theorem

The following system of the linear remainder equations S can be calculated according to the Chinese remainder theorem.

$$S:\begin{cases} x \equiv d_1 \left(mod\, c_1\right) \\ x \equiv d_2 \left(mod\, c_2\right) \\ \quad \cdots \\ x \equiv d_n \left(mod\, c_n\right) \end{cases} \tag{3}$$

For S, the assumption is that the integers in $C\left(c_1, c_2, ..., c_n\right)$ are relatively-prime, then for any integer sequence $D\left(d_1, d_2, ..., d_n\right)$, the system definitely has a general solution:

$$x \equiv \sum_{i=1}^{n} \left(M_i M_i^{-1} d_i\right) \quad \left(\mathrm{mod}\ M\right) \tag{4}$$

where $M = \prod_{i=1}^{n} c_i$, $M_i = \frac{M}{c_i}$ and $M_i M_i^{-1} \equiv 1 \quad \left(\mathrm{mod}\ c_i\right)$. .

The Embedding Process

Input: A secret;

Output: sk_1, sk_2;

Step 1: Search a text on the Internet according to the subsection titled "Section of a Text Carrier".

Step 2: Organize the text into the CSM T according to the subsection titled "The Code Matrix Of Chinese Characters' Component Structures".

Step 3: Turn the secret into the BiDS sequence $\left\{SEG_1, SEG_2, ..., SEG_{Pnum}\right\}$ according to the subsection titled "The Preprocessing of the Secret Message".

Step 4: Find out the elements of $\left\{SEG_1, SEG_2, ..., SEG_{Pnum}\right\}$ in CSM T, and get the corresponding row numbers $R_{set} = \left(r_1, r_2, ..., r_{Pnum}\right)$ and column numbers $C_{set} = \left(c_1, c_2, ..., c_{Pnum}\right)$ respectively.

Step 5: Locate the first appearance of every 3-bit code in Table 4 in CSM T to get a coordinate sequence $\left\{\left(r_1, c_1\right), \left(r_2, c_2\right), \left(r_3, c_3\right), \left(r_4, c_4\right), \left(r_5, c_5\right), \left(r_6, c_6\right), \left(r_7, c_7\right), \left(r_8, c_8\right)\right\}$. Let $R_{set}^{'} = \left\{r_1, r_2, r_3, r_4, r_5, r_6, r_7, r_8\right\}$ and $C_{set}^{'} = \left\{c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8\right\}$, then $r_{max} = \max\left(R_{set}^{'}\right)$, and $r_{min} = \min\left(R_{set}^{'}\right)$, $c_{max} = \max\left(C_{set}^{'}\right)$ and $c_{min} = \min\left(C_{set}^{'}\right)$

Compute the prime key pair according to the followingormula:

$$pk_1 = r_{max} - r_{min}, \tag{5}$$

$$pk_2 = c_{max} - c_{min}, \tag{6}$$

Step 7: Calculate the prime sequence $M_{Prime} = \left(m_1, m_2, ..., m_{Pnum}\right)$. Herein, m_1 is the first prime number that is no less than r_{max}, and $m_p - m_{p-1} \geq pk_1$, where $p \in \left[2, P_{num}\right]$. Similarly, the prime sequence $N_{Prime} = \left(n_1, n_2, ..., n_{Pnum}\right)$ will be got based on c_{max} and pk_2.

Build two systems of linear remainder equations. The secret key sk_1 can be calculated via the following system of linear equations S_1, which is constructed via the row numbers R_{set} and the prime sequence M_{Prime} according to the subsection titled "The General Solution To The Chinese Remainder Theorem". Similarity, the secret key sk_2 is figured out according to S_2.

$$S_1:\begin{cases} sk_1 \equiv r_1 \pmod{m_1} \\ sk_1 \equiv r_2 \pmod{m_2} \\ \cdots \\ sk_1 \equiv r_{Pnum} \pmod{m_{Pnum}} \end{cases} \tag{7}$$

$$S_2:\begin{cases} sk_2 \equiv c_1 \pmod{n_1} \\ sk_2 \equiv c_2 \pmod{n_2} \\ \cdots \\ sk_1 \equiv c_{Pnum} \pmod{n_{Pnum}} \end{cases} \tag{8}$$

Step 9: sk_1, sk_2 will be shared with the receiver offline or are taken as the parameter in the URL of the text carrier and sent to the receiver.

The Extracting Process

Input: The received text, and its URL or two keys, i.e. sk_1, sk_2;

Output: The secret message;

Step 1: Get the sk_1, sk_2 directly from the receiver or by parsing the received URL.

Step 2: Construct the text into the CSM T as the sender does according to the subsection titled "The Code Matrix Of Chinese Characters' Component Structures".

Step 3: Compute pk_1 and pk_1 as described in Step 5 & 6 in the embedding process.

Step 4: Calculate two short prime number sequences, i.e., $M'_{Prime} = \left(m_1, m_2, ..., m_6\right)$ and $M'_{Prime} = \left(m_1, m_2, ..., m_6\right)$, to get the length of the secret based on T as described in Step 7 in the embedding process.

Step 5: Figure out the remainder of $sk_1 \,/\, m_i \left(m_i \in M'_{Prime}\right)$ to construct the row number sequence $\left(r_1, r_2, ..., r_6\right)$ of the BiDSs $\left\{SEG_1, SEG_2, ..., SEG_6\right\}$ in T. Similarly, get the column number sequence $\left(c_1, c_2, ..., c_6\right)$.

Step 6: Identify the elements $\left\{SEG_1, SEG_2, ..., SEG_6\right\}$ in T based on the sequences $\left(r_1, r_2, ..., r_6\right)$ and $\left(c_1, c_2, ..., c_6\right)$.

Step7: concatenate the elements in $\left\{SEG_1, SEG_2, ..., SEG_6\right\}$ to form a BDS which is L_s, and convert L_s into a decimal digit, i.e. L_{secret}.

Step 8: Calculate the complete prime sequence $M_{Prime}=\left(m_1,m_2,...,m_{Pnum}\right)$ and $N_{Prime}=\left(n_1,n_2,...,n_{Pnum}\right)$, where $L_{msg}=L_{secret}+16$ and $P_{num}=\frac{L_{msg}}{p}$. By the way, M'_{Prime} and N'_{Prime} are the first 6 elements in the M_{Prime} and N_{Prime} respectively.

Step 9: Acquire the remainder of $sk_1 / m_i \left(m_i \in M_{Prime}\right)$ to constitute the row number sequence $\left(r_1,r_2,...,r_{Pnum}\right)$ of the BiDSs $\left\{SEG_1,SEG_2,...,SEG_{Pnum}\right\}$ in T. Similarly, get the column number sequence $\left(c_1,c_2,...,c_{Pnum}\right)$ based on sk_2 and $N_{Prime}=\left(n_1,n_2,...,n_{Pnum}\right)$.

Step 10: Find out $\left\{SEG_1,SEG_2,...,SEG_{Pnum}\right\}$ in T according to the row number sequence $\left(r_1,r_2,...,r_{Pnum}\right)$ and the column number sequence $\left(c_1,c_2,...,c_{Pnum}\right)$.

Step 11: The secret is obtained by decoding the remaining BDS after removing the first 16 bits of the $\left\{SEG_1,SEG_2,...,SEG_{Pnum}\right\}$.

RESULTS AND ANALYSIS OF EXPERIMENT

The Experiment

Take the following text as a carrier:

故宫博物馆,是一座建立在明清两朝皇宫——紫禁城的基础上的特殊的博物馆。

According to the subsection titled "The Code Matrix Of Chinese Characters' Component Structures", the following results will be figured out: $N=23$, $n=5$.

The corresponding MSM and CSM are illustrated in Tables 6 & 7 respectively.

Table 6. The Chinese Characters' Matrix

	1	2	3	4	5
1	故	宫	博	物	院
2	是	一	座	建	立
3	在	明	清	两	朝
4	皇	宫	紫	禁	城
5	的	基	础		

Table 7. The Code Matrix

	1	2	3	4	5
1	010	101	011	010	011
2	101	000	111	111	000
3	001	010	010	000	010
4	101	101	110	101	011
5	010	101	100		

The subsequent embedding process is illustrated in Table 8.

Table 8. An Example

Step No.	The Steganography Procedure	The Example
step1	The secret message	"行动"
step2	Convert the secret into the binary string	"行": 1101000011010000; "动": 1011011010101111; (the codes in GB 2312-80 characters' dataset)
	concatenate the binary string and package the length of the whole secret binary string as a header in 16 bits	0000000000010000011010000 1101000010110110101011111
	Split the composed binary string into the slices	000, 000, 000, 010, 000, 011, 010, 000, 110, 100, 001, 011, 011, 010, 101, 111
step3	The Chinese characters square matrix The code square matrix	Table 6 Table 7
step4	The row numbers sequence $(r_1,r_2,\ldots,r_p,\ldots,r_{Pnum})$ The column numbers sequence $(c_1,c_2,\ldots,c_p,\ldots,c_{Pnum})$	{2,2,2,1,2,1,1,2,4,5,3,1,1,1,1,2} {2,2,2,1,2,3,1,2,3,3,1,3,3,1,2,3}
step5	pk_1, pk_2	4, 2
step6	$M_{Prime}(m_1,m_2,...,m_{Pnum})$ $N_{Prime}(n_1,n_2,...,n_{Pnum})$	{5,11,17,23,29,37,41,47,53,59,67,71,79,83,89,97} {3,5,7,11,13,17,19,23,29,31,37,41,43,47,53,59}
step7	Construct the system of the linear remainder equations	S1 S2
step8	sk_1, sk_2	46883104274564028 20 47084297252105043 56

Note: Step No. in the first column corresponds to the subsection titled "The Embedding Process".

ANALYSIS OF THE ALGORITHM

The Steganography Success Rate

As long as a text contains all the 8 categories of structures in Table 4, a secret will be concealed successfully. In spite that only a small amount of texts are verified, the statistics in Table 5 still show that a text with more than 86 Chinese characters will almost contain these 8 categories of structures. Table 9 shows that the number of characters in a Chinese text on the Internet is usually more than 86. Therefore, a text with 8 different categories can be easily found on the Internet. Therefore, any secret can be concealed by the proposed method in theory. At best, a text could be eight-character length only if these eight characters have different structures categorized in Table 4.

Table 9. Statistics on the number of Internet texts' characters

Field	The number of Texts	The Maximum number of characters	The Minimum number of characters
News	31	6520	2310
Science	32	3535	2129
Prose	36	2551	1457
Health	27	4520	2131
E-commerce	27	6648	2006
Advertisement	33	5960	1978
Novel	22	2665	1753
Technology	28	3269	2390

Generally, if the Chinese character structures are divided into 2^p categories, any secret can be concealed in a text as long as the characters in this text have all these 2^p categories.

Imperceptibility

In this paper, the text can be sent to the receiver directly or via its corresponding URL, which is a very common way to share some information via the Internet. And the text will not be modified in any way. What's more, a text can be chosen in any topic, so the sender can select a text that is related to the topic that both the sender and the receiver are interested in and is not related to the secret. All these practices make a text look innocent.

The Steganographic Capacity (SC)

The SC is a very important evaluation indicator for steganography. In general, the high SC conflicts with good imperceptibility, so the imperceptibility is traditionally achieved by sacrificing the SC. But the proposed method can improve the SC without deteriorating the imperceptibility. The SC generally refers to the total amount of digital bits concealed in a text. Herein, it is reckoned from two perspectives as follows:

(1) The SC of a single Chinese character ($\varepsilon^{'}$)

In the proposed method, each Chinese character in a text can express *p*-bit binary. For instance, each Chinese character can express 3 bits in the above illustration. In fact, the same slice might be used more than once in transmission, so a character usually represents more than *p*-bits.

(2) The SC of a text (ε)

Just as introduced above, the BiDSs in the CSM can be reused, so the SC of a single Chinese character might be more than p bits. The total SC of a text, i.e., ε has nothing to do with the size of text and is not limited to $\varepsilon^{'}$ and it can be calculated as $\varepsilon = 2^p * p * P_{avg}$. Herein, P_{avg} denotes the average reused times of one BiDS in the CSM T. Once the structure categories are defined, p is a determinate value. Thus the total SC is only determined by P_{avg}. In theory, the P_{avg} can be any large value. But as it increases, two prime sequences, i.e. M_{Prime} and N_{Prime} will become longer, sk_1, sk_2 will become too large, thus the computation of the corresponding line remainder equation systems

becomes time-consuming or needs more expensive processing machines. In this case, a long secret often needs to be segmented and concealed.

Other Performances

Robustness is the ability of anti-attack. The proposed method is not subject to format modification, retyping and file format transformation.

Statistical Undetectability means that the secret existence will not be detected in a stego-cover. The proposed method does not do any modification to a text, so the semantics and statistical steganalysis methods couldn't detect the concealment.

Extensibility means that a method can be used in different languages in this paper. The proposed method can be employed in not only Chinese but also in other languages, such as English, Japanese, Korean, etc. For example, a morpheme is the component of an English word. "book" has one morpheme, "telephony" consists of two morphemes, i.e. "tele" and "phony". The words with one morpheme can be used to express "0" and the words with two morphemes can be used to express "1". Other classifications might be employed but not listed here. Japanese characters and Korean characters can also be divided into different categories according to their structures.

The Performance Comparison With Other Methods

As shown in Table 10, six metrics are compared between the proposed method and other methods including 5 catagories of steganography. Herein, the power of perception is differentiated from the human being and machines, as indicated by the two columns titled imperceptibility and the undetectability respectively in Table 10. The imperceptibility column indicates the human perception, and the other is the indicator from the machine's perspective.

Table 10. The performance comparison

Method		Modify text(Y/N)	Imperce-ptibility	Robust-ness	steganography success rate	Steganographic capacity	Statistical Undetectability
Format-based Modification	(Bhaya et al., 2013)	Y	Middle	Bad	100%	2.67 bits per character	Bad
Semantics-based Modification	(Sui & Luo, 2004)		Bad	Bad	100%	log_2^n bits per Word in a n-word ynonym group	Bad
Steganography by cover search	(Shi et al., 2016)	N	Middle	Good	<=100%	14bit per Character (URL)	Good
Steganography by cover generation	(Yang et al., 2019)		Middle	Middle	100%	Related to the text library	Good
Steganography by cover index	(Zhou et al., 2016)		Good	Good	<=100%	2.5 characters per text	Good
The method in this paper			Good	Good	100%	p bits per character; $8*p*P_{avg}$ bits per text	Good

As shown in Table 10, the SCs of the Format-based Modification method is subjected to the number of the similar font types, at most 2.67 bits per character for 3 similar font types (Bhaya et al., 2013). And it is easily detected by machine and might lead to wrong extraction, cannot resist the retyping attack. In Sui and Luo (2004), the SC is log_2^n bits per word in n-word synonym group, and semantics-based modification usually makes a text badly fluent and the semantics detection algorithm (Zuo et al., 2018) and the syntax detection algorithm (Fu et al., 2015) could figure out whether there

exists a secret in a text. That is, the statistical undetectability in Bhaya et al. (2013) and Xin and Hui (2004) is in general worse. In Shi et al. (2016), the steganographic capacity is relatively high, and the detection methods such as semantics detection algorithm (Zuo et al., 2018) and syntax detection algorithm (Fu et al., 2015) don't work, but its steganography success rate will decrease as a secret becomes longer. In Yang et al. (2019), the steganography success rate appears relatively high, but the secret existence is very easy to be detected by the statistical detection algorithm proposed in Yang and Cao (2010), which doesn't work on the proposed method, because its statistical undetectability is bad. The method proposed in Zhou et al. (2016) has good the imperceptibility and robustness, but its SC is 2.58 characters per text and it is relatively low, and the steganography success rate is also lower than other methods. Compared with these methods, the proposed method has higher, good imperceptibility, robustness and high statistical undetectability than other methods.

CONCLUSION

In order to solve the existing problems such as poor imperceptibility, bad robustness, statistical detectability, low success rate, and low steganographic capacity, this paper proposes a new coverless steganography method by encoding the Chinese characters' component structures. The analysis shows the proposed method can easily conceal any message successfully and has a high SC as long as the computing power is enough. A Chinese character can conceal at least p bits, and for a text carrier, the SC can be reckoned as $\varepsilon = 2^{p} * P_{avg}$, which is higher than the existing text stegnography. Its robustness, imperceptibility, and statistical undetectability are also much better than currently existing methods, and especially, it is also applied in other languages including English.

Chang, C. Y., & Clark, S. (2010). Linguistic Steganography Using Automatically Generated Paraphrases. In Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics. Association for Computational Linguistics.

Changder, S., Debnath, N. C., & Ghosh, D. (2009). A new approach to Hindi text steganography by shifting matra. In *Proceeding of the 2009 International Conference on Advances in Recent Technologies in Communication and Computing* (pp. 199-202). Washington, DC: IEEE Computer Society.

ACKNOWLEDGMENT

This research was supported in part by the NSFC-General Technical Research Foundation Joint Fund of China under Grant U1536113; and the CERNET Innovation Project under Grant NGII20180405.

REFERENCES

Alattar, A. M., & Alattar, O. M. (2004). Watermarking electronic text documents containing justified paragraphs and irregular line spacing. *Proceedings of SPIE - The International Society for Optical Engineering, 5306*, 685-695. doi:10.1117/12.527147

Ali, S. (2018). A state-of-the-art survey of coverless text information hiding. *International Journal of Computer Network and Information Security*, *10*(7), 52–58. doi:10.5815/ijcnis.2018.07.06

Alsaadi, H. I., Al-Anni, M. K., Almuttairi, R. M., Bayat, O., & Ucan, O. N. (2018). Text Steganography in Font Color of MS Excel Sheet. In *Proceedings of the First International Conference on Data Science, E-learning and Information Systems*. New York: ACM. doi:10.1145/3279996.3280006

Bender, W., Gruhl, D., Morimoto, N., & Lu, A. (1996). Techniques for data hiding. *IBM Systems Journal*, *35*(3&4), 313–336. doi:10.1147/sj.353.0313

Bennett, K. (2004). *Linguistic steganography: survey, analysis, and robustness concerns for hiding information in text*. Center for Education and Research in Information Assurance and Security, Purdue University.

Bhaya, W., & Rahma, A. M. (2013). Text steganography based on font type in ms-word documents. *Journal of Computational Science*, *9*(7), 898–904. doi:10.3844/jcssp.2013.898.904

Chapman, M. T., & Davida, G. I. (1997). Hiding the hidden: A software system for concealing ciphertext as innocuous text. In *Proceedings of the First International Conference on Information and Communication Security (*ICICS '97). London, UK: Springer-Verlag. doi:10.1007/BFb0028489

Chen, X., Sun, H., Tobe, Y., Zhou, Z., & Sun, X. (2015). Coverless Information Hiding Method Based on the Chinese Mathematical Expression. In Z. Huang, X. Sun, J. Luo, & J. Wang (Eds), *International Conference on Cloud Computing and Security. (ICCCS 2015, LNCS)* (vol. 9483, pp. 133-143). Springer International Publishing. doi:10.1007/978-3-319-27051-7_12

Cui, G. M., Hong, X., Yuan, X., Zhang, Y. W., & Zhu, E. Z. (2016). Research of information hiding based on invisible character replacement [in Chinese]. *Computer Applications and Software*, *33*(4), 277–280.

Fridrich, J. (2009). *Steganography in digital media: principles, algorithms, and applications*. Cambridge University Press. doi:10.1017/CBO9781139192903

Fu, M., Dai, Z. X., & Hu, W. T. (2015). A syntax checking algorithm for information-hiding [in Chinese]. *Kexue Jishu Yu Gongcheng*, *15*(21), 142–145F.

Fu, Y. H. (1991). The structure of Chinese characters [in Chinese]. *Language Planning*, (9), 10–11.

Gan, C., Sun, X., Liu, Y., & Xiang, L. (2007). Improved steganographic algorithm based on synonymy substitution for Chinese text. *Journal of Southwest University (Natural Science Edition)*, *37*(I), 137–140.

Lazic, N., & Aarabi, P. (2006). Communication over an acoustic channel using data hiding techniques. *IEEE Transactions on Multimedia*, *8*(5), 918–924. doi:10.1109/TMM.2006.879880

Lee, I. S., & Tsai, W. H. (2008). Secret communication through web pages using special space codes in html files. *International Journal of Applied Science and Engineering*, *6*(2), 141–149.

Luo, Y., & Huang, Y. (2017). Text Steganography with High Embedding Rate: Using Recurrent Neural Networks to Generate Chinese Classic Poetry. *ACM Workshop on Information Hiding & Multimedia Security*. doi:10.1145/3082031.3083240

Ministry of Education, & National Working Committee on Languages and Writing Systems. (2013). *List of Common Standard Chinese Characters.* Beijing: Ministry of Education of China. Retrieved Oct. 15, 2019, from http://www.gov.cn/zwgk/2013-08/19/content_2469793.htm

Mstafa, R. J., Elleithy, K. M., & Abdelfattah, E. (2017). Video steganography techniques: Taxonomy, challenges, and future directions. In *Proceeding of 2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT)* (pp. 1-6), New York: Farmingdale State College. doi:10.1109/LISAT.2017.8001965

Ou, L., Sun, X., & Liu, Y. (2007). Adaptive algorithm of text information hiding based on character intensity [in Chinese]. *Jisuanji Yingyong Yanjiu*, *24*(5), 130–132.

Richard, B. (2004). *Towards Linguistic Steganography: A Systematic Investigation of Approaches, Systems, and Issues*. Available from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.122.8569

Shi, S. W., Qi, Y. N., & Huang, Y. F. (2016). An Approach to Text Steganography Based on Search in Internet. In *2016 International Computer Symposium*. Syngress. doi:10.1109/ICS.2016.0052

Wang, K., & Gao, Q. (2019). A Coverless Plain Text Steganography Based on Character Features. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 95665–95676. doi:10.1109/ACCESS.2019.2929123

Xiao, C., Zhang, C., & Zheng, C. (2018). FontCode: Embedding Information in Text Documents Using Glyph Perturbation. *ACM Trans. Graph., 37*(2), 15:1-15:16.

Sui, X. G, & Luo, H. (2004). A secure steganography method based on text. *Computer Engineering, 30*(19), 104-105, 191. (in Chinese)

Yang, H., & Cao, X. (2010). Linguistic steganalysis based on meta features and immune mechanism. *Chinese Journal of Electronics*, *19*(4), 661–666.

Yang, Z. L., Guo, X. Q., Chen, Z. M., Huang, Y. F., & Zhang, Y. J. (2019). Rnn-stega: Linguistic steganography based on recurrent neural networks. *IEEE Transactions on Information Forensics and Security*, *14*(5), 1280–1295.

Popa, R. (1998). *An Analysis of Steganographic Techniques*. The Politehnica University of Timisoara. Department of Computer Science and Software Engineering. Retrieved from https://ad.informatik.uni-freiburg.de/mitarbeiter/will/dlib_bookmarks/digital-watermarking/popa/popa.pdf

Zhang, X. P., Qian, Z. X., & Li, S. (2016). Prospect of digital steganography research. [in Chinese]. *Journal of Applied Sciences (Faisalabad)*, *34*(5), 475–489.

Zhou, Z., Mu, Y., Zhao, N., Wu, Q. M. J., & Yang, C. N. (2016). Coverless Information Hiding Method Based on Multi-keywords. In X. Sun, A. Liu, H.-C. Chao, & E. Bertino (Eds.), *Proceedings of the Second International Conference on Cloud Computing and Security (ICCCS 2016) Part I, LNCS* (vol. 10039, pp. 39-47). Nanjing, China: Springer International Publishing AG.

Zielińska, E., Mazurczyk, W., & Szczypiorski, K. (2014). Trends in steganography. *Communications of the ACM*, *57*(3), 86–95.

Zuo, X., Hu, H., Zhang, W. M., & Yu, N. H. (2018). Text Semantic Steganalysis Based on Word Embedding. In X. Sun, Z. Pan, & E. Bertino (Eds.), *Proceedings of 4th International Conference on Cloud Computing and Security (ICCCS 2018), Part IV, LNCS* (vol. 11066, pp. 485-495). Haikou, China: Springer Nature Switzerland AG.

Kaixi Wang received his doctoral degree in telecommunications from Beijing University of Posts and Telecommunications, Beijing, China, in 2008. He is currently an Associate Professor at the College of Computer Science and Technology in Qingdao University. His research interests mainly include network security, content security, next generation networks, telecommunication software, distributed computing.

Xiangmei Yu is currently pursuing for a master's degree in the College of Computer Science and Technology of Qingdao University, China. Her research interests mainly include information security, coverless steganography and natural language processing.

Ziyi Zou received his BS in Software Engineering from Qingdao University in 2018, and is currently pursuing his master's degree in the College of Computer Science and Technology of Qingdao University, China. His research interests mainly include information security, coverless steganography and other content security based on natural languages.