http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (19): 4946-4952, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.4946.4952

Watermarking Chinese Text Documents for Copyright Protection

[1]Xinmin Zhou, [1]Lina Tan and [2]Li Pan
[1]School of Computer Science and Information Engineering, Hunan University of Commerce, Hunan, 410205, Changsha, China
[2]College of Information and Communication Engineering, Hunan Institute of Science and Technology, Hunan, 414006, Yueyang, China

Abstract: A robust text watermarking algorithm is proposed to solve the problem of copyright protection for Chinese text documents in this study which watermarking structure and embedding strategy are comprehensively considered. Structure types and strokes of Chinese characters can be obtained through Chinese characters mathematic expression and the whole document is divided into two blocks by utilizing structure types of Chinese characters, the position selected to embed is chose by Chinese character's strokes and the watermarking bit in each block. Watermarking bits with a robust structure is generated by chaos encrypting and hamming checkout coding. Positions of the destroyed watermarking bits can be judged exactly and block-checkout and hamming-checkout are used to recover them. The experimental result indicates that the proposed algorithm is robust and transparent.

Key words: Text watermarking, copyright protection, Chinese character mathematical expressions, information security

INTRODUCTION

In the research of digital watermarking, protecting the watermark from being illegally detected or modified and robustness to malicious attacks, are the urgent problems that needs to be solved (Cox and Miller, 2001; Shen *et al*., 2007). Currently, most watermarking research, publications and products are dedicated to images, video, audio; less has been published on text watermarking. For text watermarking, we have to distinguish between methods that hide information in the semantics which means in the meaning and ordering of the words and methods that hide information in the format which means in the layout and the appearance. Atallah *et al*.(2002) implemented watermarking embedding based on TMR (Text meaning representation). This algorithm is more robust, moreover its capacity of watermarking is large too. Since, computer can not understand the meaning of a text fully correctly, it is most difficult to improve this text watermarking technique. Maxemchuk and Low (1997) and Brassil *et al*. (1995, 1999) have proposed three different methods for information embedding into text documents: Line-shift coding; word-shift coding and feature coding. In line-shift coding, single lines of the document are shifted upwards or downwards by very small amounts. The information to be hidden is encoded in the way the lines are shifted. Similarly, words are shifted horizontally in order to modify the spaces between consecutive words in word-shift coding. Both methods are applicable to the format file of a document or to the bitmap of a page image. The third method, feature coding, slightly modifies features such as the pixel of characters, the length of the end lines in characters etc. Among the three presented methods, line-shift coding is the most robust in the presence of noise but also most easily defeated. Although the described methods can theoretically be defeated, it requires interactive human intervention and is expensive in practice.

The literature materials indicate that research on robust text watermarking focus on embedding strategy mostly, but there are two parts to building a strong watermark: the watermarking structure and the embedding strategy. In order for a watermark to be more robust and secure, these two components must be designed correctly (Cox *et al*., 1997). Public watermarking technique requires neither the secret original nor the embedded watermarking during the extracting, so it remains the most challenging problem at present (Kutter and Petitcolas, 1999). In this study, an efficient robust text watermarking algorithm is proposed based on feature coding and the research basis of (Zhou *et al*., 2008; Sun *et al*., 2004). A watermarking structure of self-recovery generated by

Corresponding Author: Xinmin Zhou, School of Computer Science and Information Engineering, Hunan University of Commerce, Hunan, Changsha, 410205, China

Hamming-checkout coding and a better embedding strategy that embed watermarking bits symmetrically in two text blocks, respectively are presented in this method. The algorithm can locate positions of some destroyed watermarking bits and some destroyed watermarking bits can be recovered by block-checkout and hamming-checkout. At the same time, the extracting needs neither the secret original nor the embedded watermarking.

STRUCTURE KNOWLEDGE OF CHINESE CHARACTER AND ITS APPLICATIONS

A Chinese character can be expressed into a mathematical expression in which the components of Chinese characters act as operands and the position relationships between two components act as operators which satisfy some certain operation laws, just like general math expressions. Six spatial relations of two components are needed to be defined as the operators in the mathematical expression. These six operators are lr, ud, ld, lu, ru and we which represent, respectively the spatial relation of left-right, up-down, left-down, left-upper, right-upper and whole-enclosed defined strictly in (Sun *et al*., 2002). Some of the selected basic components and their serial numbers are shown in Fig. 1 and the intuitive explanation of six operators according to the component positions is shown in Fig. 2. We can acquire the structure types and the strokes of Chinese characters by the mathematical expression of Chinese characters.

In this study, Structure types of Chinese characters are used to divide the whole document into two blocks. In each block, not only watermarking bit but also Chinese character's strokes are chose to determine the position selected to embed watermarking. According to the general table of single character frequency of Chinese characters, we have calculated that the frequency of the Chinese characters of left-right body (lr) and mixed body (we, lu, ld and ru) accounts for 50.08% and the frequency of the Chinese characters of up-down body (ud) and single body (basic component) accounts for 49.71%. Based on the above statistical results, a hosted document can be divided into two blocks comparatively uniformly which is called first block and second block, respectively. Further Statistical results show that the respective frequency of the Chinese characters with odd strokes and even strokes in first block account for 21.65 and 28.43%, the corresponding frequency in second block account for 24.58 and 25.12%. Therefore, selecting the parity of Chinese characters strokes as the controlling parameter is advisable while embedding. In this study, the hosted document is divided into two blocks, the first block includes the Chinese characters of left-right body and mixed body; the second block includes the Chinese characters of up-down body and single body.

Obviously, the stroke numbers and structure types of all the 20902 CJK Chinese characters in UNICODE 3.0 can be all obtained. In order to describe the statistical results intuitively, the statistical charts of stroke numbers and structure types of Chinese characters have been shown in following Fig. 3 and 4, respectively.

PRINCIPLE and WATERMARKING ALGORITHM

Through deep analysis on the structures and strokes of Chinese characters, we find that it gives a few advantages to Chinese text digital watermarking.

Fig. 1: Basic components and their serial numbersw

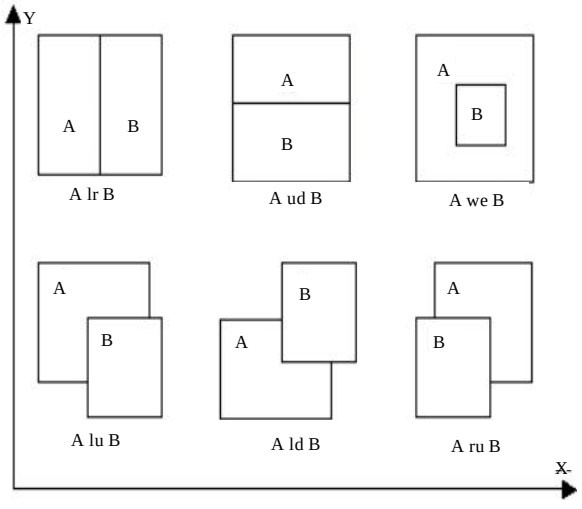

Fig. 2: Intuitive explanation of the 6 operators

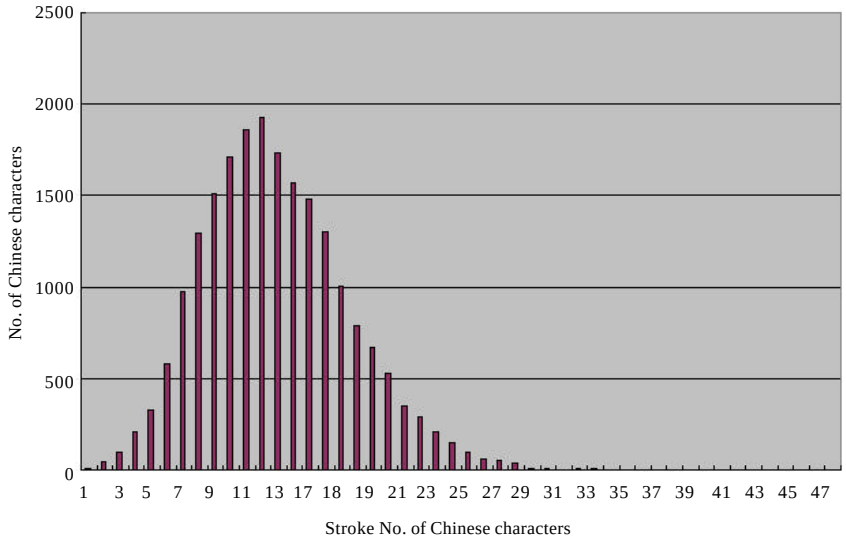

Fig. 3: Statistical chart of stroke numbers of Chinese characters

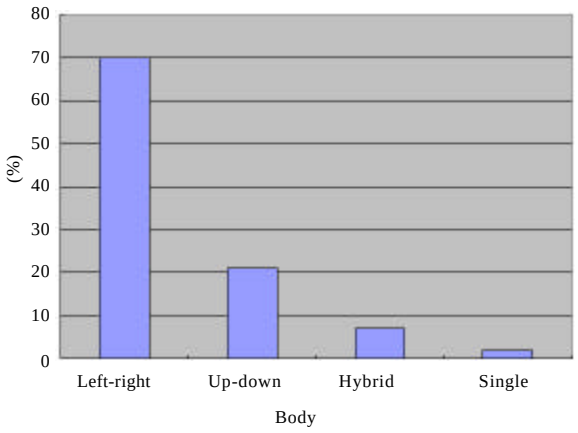

Fig. 4: Statistical chart of structure types of Chinese characters

Moreover, the preferable encoding density of feature coding provides conveniences to encoding of error correction. In this section, we will present the principle and algorithm of the watermarking technique.

Principle: In watermarking algorithm, hamming coding and chaotic encrypting are utilized for designing an advisable watermarking structure to improve the robustness and security of watermarking. Get the UNICODE of watermarking information (marked as M_0) and convert it into binary bits then M' = E(M_1) and M" = H(M'). Given binary bits S, the length of S is a whole number divided by 8. Every 8 bits of S are encoded by hamming encoding and become into 12 bits after hamming encoding. This encoding process is defined as function H (x). If (M_i = 1 and S(C_i) mod 2 = 1) or (M_i = 0 and S(C_i) mod 2 = 0), then C_i is a candidate of watermarking embedding positions. Let C_i hiding a watermarking bit, if S(C_i) mod 2 =1, then char = 1, else char = 0. In order to make the following watermarking algorithm easily understood, some definitions are expounded as follows:

Definition 1: Let C be any Chinese character, C = A_1 OP_1 A_2 OP_2 A_3... OP_{n-1} A_n, let A_i (i = 1, 2,... n) denote the number of the basic component of Chinese characters, OP_i (i = 1, 2,... n-1) denote the operator. If S[A_i] denotes the stroke number of the basic component A_i and S[C] denotes the stroke number of the Chinese character C, then, S[C] = ΣS[A_i]. Here S[A_i] is obtained from the database document of stroke numbers of the basic components.

Definition 2: Optical bistable model X_{n+1}= A $\sin^2(X_n-X_B)$ is used as chaotic encrypting algorithm and then make iterative computations, if X_i>2×A/3, let the chaotic sequence value (denoted by C_i) be equal to 1, else let C_i = 0, here take A = 4, X_B = 2.5, the equation is in chaos. As different X_0 will results in complete different chaotic sequence values, X_0 is regarded as the key of the encrypting algorithm. Plaintext is marked as P_i and ciphertext is marked as W_i, then W_i = P_iXOR C_i, P_i and W_i are equal to 0 or1. This encrypting process is defined as function E(x).

Definition 3: Given the jth bit of watermark w_j and underline[16] = {2, 3, 4, 6, 7, 9, 10, 11, 20, 23, 25, 26, 27, 39, 43, 55}, in which every different element of the array represents the value of different type underline. w_j is embedded by setting a corresponding underline to a candidate and hiding it simultaneously, the value of underline type is equal to underline[i], i = ((j-1) mod 16)+1.

Definition 4: Let M" = M_0"M_1"...M_{n-1}", n is a whole number divided by 12. Block M" into k groups and k = n/12. Given the bits of group i: $G_i = M_{12i-12}$"M_{12i-11}"... M_{12i-1}", j = (i-1), then converting j into 4 bits of binary bits marked as B_i. Here define A+B = AB, then $G'_i = G_i+B_i$ Watermarking bits M is generated according to above rules from M". This process of generating watermarking bits is defined as function G(x).

Definition 5: Let M = $M_0M_1...M_{n-1}$, the watermarking bit M_j is embedded by setting underlines to Chinese characters which using the value corresponding to the array element underline[i] and hiding it simultaneously, i = 1 mod 16, iϵ[0, 15], jϵ[0, n-1], n represents the length of watermarking bits.

Watermarking extracting can be regarded as the reverse of watermarking embedding simply. During embedding and extracting algorithm, some definitions are presented as follow: H_t denotes the string of characters in the text document to be used as the hosted medium of watermarking; H'_t denotes the string of characters in the hosted text document in which the watermarking bits have been embedded.

Embedding algorithm:

Input: A hosted Chinese text document H_t, watermarking information M_0 and the key X_0
Output: Watermarked hosted Chinese text document H'_t in which the watermarking information M_0 is embedded

Begin
S_1: Divide the hosted document H_t into two blocks
S_2: Generate the bit stream M from M_0 using the key X_0
S_3: Embed bit stream M in the first block of the hosted document H_t from front to back:
for k: =1 to n do // n is the length of M
{get C_i
embed M_k}
S_4: Embed bit stream M in the second block of the hosted document H_t from back to front, the steps is the same with S_3.
End.

Extracting algorithm:

Input: A watermarked hosted text document H'_t and the key X_0
Output: Watermarking information M_0
Begin
S_1: Divide watermarked text document H'_t into two blocks
S_2: Extract bit stream M in the first block of the hosted document H'_t from front to back:
for k:=1 to m do
// m is the number of given text block
{get the watermarking embedding positions C_i
extract M_k}
S_3: Extract bit stream M in the second block of the hosted document H'_t from back to front, the steps is the same with S_2
S_4: If the extracting result is complete, then goto S_6, else recover the destroyed watermarking bits utilizing block-checkout and hamming-checkout
S_5: If the recovering result is complete, then goto S_6, else report watermarking extracting is failed
S_6: Map M to watermarking information M_0
End.

EXPERIMENTAL RESULTS

We have implemented the watermarking technique in Microsoft Word2003. The experimental results show that the technique in this study is more robust. In order to prove the validity of the algorithm, let us give some examples of the watermarking system. Here The Transient Days of Zhu Ziqing's famous essays is regarded as the original hosted document in which there are 542 Chinese characters, 270 Chinese characters in the first block and 272 Chinese characters in the second block. Figure 5 is the blocking result of watermarked text document and Fig. 6 is the extracting result of integrated watermarked text document. Figure 7 is the extracting result of watermarked hosted text which has been attacked by deleting some characters and sentences. Figure 8 is the extracting result by illegal key. The extracting result using the legal key

Fig. 5: Blocking result of watermarked text document

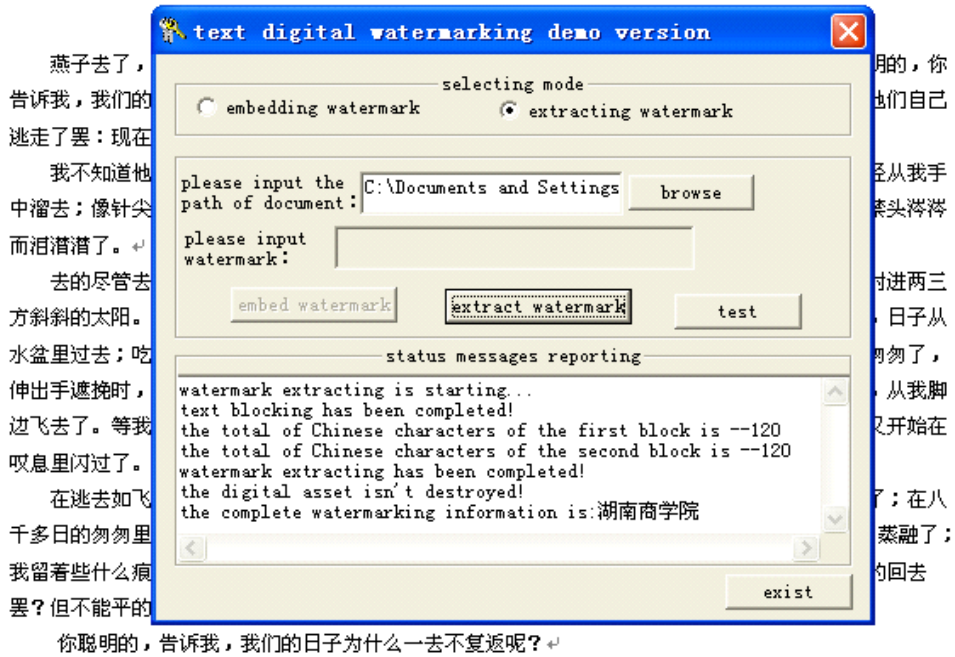

Fig. 6: Extracting result of integrated watermarked text document

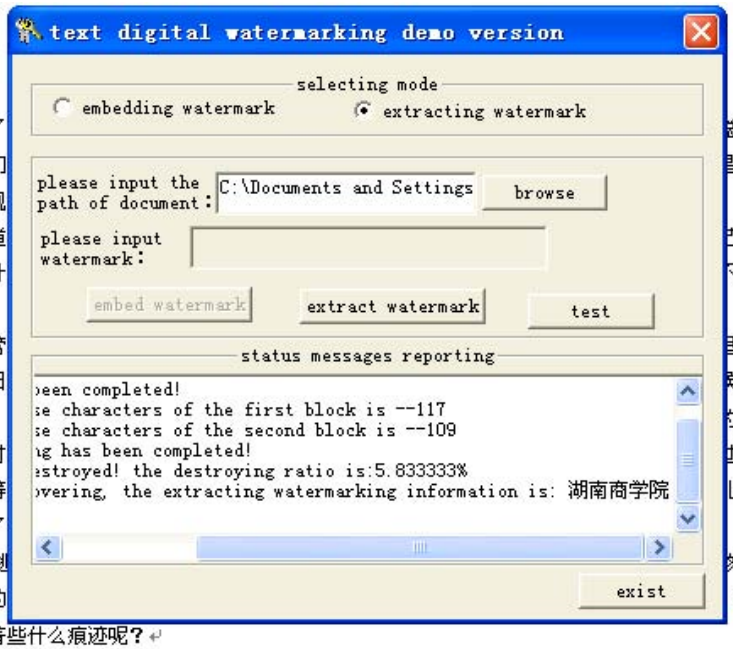

Fig. 7: Extracting result of attacked watermarked text document

Fig. 8: Extracting result by illegal key

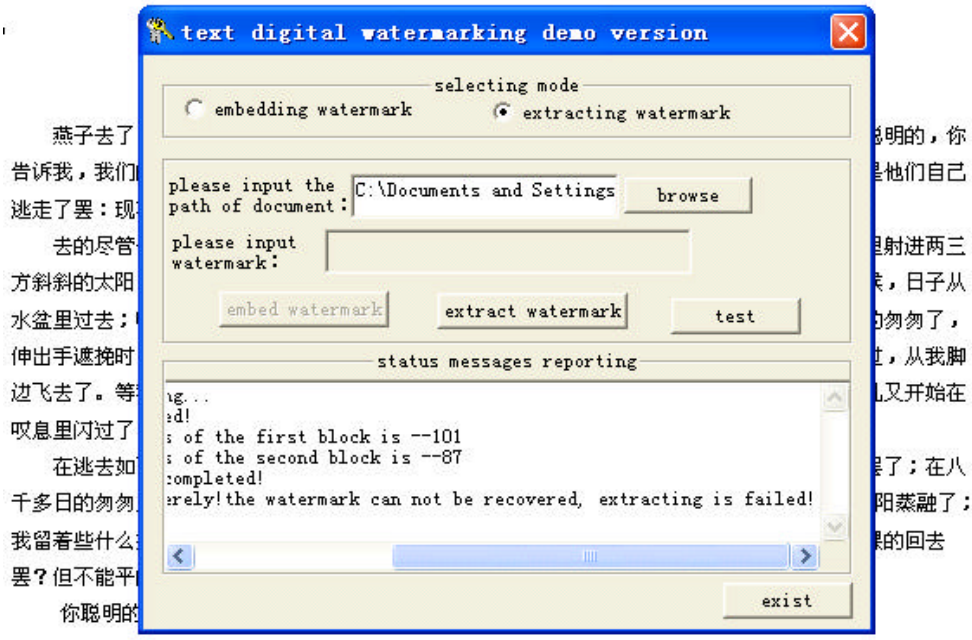

Fig. 9: Extracting result of watermarking bits beyond recovery

X_0 = 3.14156 is Hunan University of Commerce; however the extracting result using illegal key X_0 = 3.141561 is random codes. The extracting result of watermarking bits beyond recovery is showed in Fig.9. From these figures we can see the robustness of this technique is relatively high.

Because the frequency of Chinese characters of odd strokes counts for 46.24% and the frequency of Chinese characters of even strokes counts for 53.55%, the theoretical value of watermarking capacity of the algorithm is the half of the Chinese characters number in the block which Chinese characters number is lower than the other. Experimental results show, the practical value of watermarking capacity is slightly lower than the theoretical value of it in order to enable watermarking bits to distribute every row of documents more uniformly.

CONCLUSION

Internet network has brought convenience of the distribution of multimedia information and challenges t the same time. The copyright protection and integrity problem of the digital assets are being threatened seriously. With this background, a novel robust text watermarking algorithm is proposed in this study. The extracting of the proposed algorithm needs neither original hosted text nor primitive watermark, at the same time, watermarking extracted is significant.

The robustness of algorithm is ensured through a better embedding strategy which designed by the structures and strokes of Chinese characters and feature coding of underline-based and an advisable watermarking structure which designed by hamming coding and chaotic encrypting. The characteristics of the algorithm proposed in this study is presented as follows: (1) Watermarking bits are embedded in two text blocks, respectively according to the symmetry theory, (2) Locating accurate, watermarking bits of every group are corresponding to different underlines of every group which record the positions of watermarking bits and (3) Positions of some destroyed watermarking bits can be detected and these destroyed watermarking bits can be recovered by block-checkout and Hamming- checkout to some extent.

Due to its wide prospects in military, commerce and industry, as a challenging research subject, text watermarking has attracted more and more attention of academicians and researchers. In the future, there are still many spaces to improve the watermarking algorithm, such as making the algorithm blind and improving the watermarking security. Our future research will aim at improving watermarking robustness during the process of enhancing the capacity of watermarking.

ACKNOWLEDGMENTS

This study was supported by the Ministry of Education, Humanities and Social Sciences Research Projects (Grant No. 12YJAZH216), the National Social Science Fund Projects (Grant No. 13CJY007) and the Scientific Research Fund of Hunan Provincial Education Department (Grant No. 11A041).

REFERENCES

Atallah, M.J., V. Raskin, C.F. Hempelmann, M. Karahan, R. Sion, U. Topkara and K.E. Triezenberg, 2002. Natural language watermarking and tamperproofing. Proceedings of 5th International Information Hiding Workshop, October 7-9, 2002, The Netherlands, pp: 196-212.

Cox, I.J., J. Kilian, F.T. Leighton and T. Shamoon, 1997. Secure spread spectrum watermarking for multimedia. IEEE Trans. Image Process., 6: 1673-1687.

Cox, I.J. and M.L. Miller, 2001. Electronic watermarking: The first 50 years. Proceedings of the IEEE 4th Workshop on Multimedia Signal Processing, October 3-5, 2001, Cannes, pp: 225-230.

Kutter, M. and F.A.P. Petitcolas, 1999. A fair benchmark for image watermarking systems. Proceedings of the Security and Watermarking of Multimedia Contents, January 25-27, 1999, Society of Photo-Optical Instrumentation Engineers, Sans Jose, California, USA., pp: 226-239.

Brassil, J.T., S. Low, N.F. Maxemchuk and L. O'Gorman, 1995. Electronic marking and identification techniques to discourage document copying. IEEE J. Selected Areas Commun., 13: 1495-1504.

Maxemchuk, N.F. and S. Low, 1997. Marking text documents. Proceedings of the International Conference on Image Processing, October 26-29, 1997, Santa Barbara, CA, USA., pp: 13-16.

Brassil, J.T., S. Low and N.F. Maxemchuk, 1999. Copyright protection for the electronic distribution of text documents. Proc. IEEE, 87: 1181-1196.

Shen, X., H.G. Zhang, D.G. Feng, Z.F. Cao and J.W. Huang, 2007. Survey of information security. Sci. China Ser. F: Inform. Sci., 50: 273-298.

Sun, X.M., G. Luo and H.J. Huang, 2004. Component-based digital watermarking of Chinese texts. Proceedings of the 3rd International Conference on Information Security, November 14-15, 2004, Shanghai, China, pp: 76-81.

Sun, X.M., H.W. Chen, L.H. Yang and Y.Y. Tang, 2002. Mathematical representation of a Chinese character and its applications. Int. J. Pattern Recogn. Artif. Intell., 16: 735-747.

Zhou, X.M., W.D. Zhao, Z.C. Wang, R. Peng and G. Wei, 2008. A robust digital watermarking of Chinese texts based on watermarking structure and embedding strategy. Proceedings of the International Congress on Image and Signal Processing, Volume 5, May 27-30, 2008, Sanya, China, pp: 635-639.