Research on Double Watermarking Algorithm Based on PDF Document Structure

Weijuan Zhao
Communication University of China
School of Information Engineering
Beijing, China
575869686@qq.com

Hu Guan
Institute of Automation, Chinese Academy of Sciences
Digital Content Technology and Service Research Center
Beijing, China
Hu.guan@ia.ac.cn

Ying Huang
Institute of Automation, Chinese Academy of Sciences
Beijing, China
ying.huang@ia.ac.cn

Shuwu Zhang
Institute of Automation, Chinese Academy of Sciences
Digital Content Technology and Service Research Center
Beijing, China
shuwu.zhang@ia.ac.cn

Abstract—**The transmission of PDF files is becoming more and more popular on the Internet. In order to ensure that electronic publications are published and distributed on the Internet, copyright transactions occur, or the original watermarked documents are still fully extracted when the watermark information is extracted, a double watermarking algorithm based on PDF documents is designed: robustness against attacks watermarking algorithm and watermarking algorithm to verify the integrity of content. The experimental results demonstrate that the algorithm has good invisibility and robustness.**

Keywords-PDFdocument;Watermarking;copyright; Robustnesst

I. INTRODUCTION

Electronic publications are spread quickly and easily in the Internet era, but at the same time there are infringements such as tampering, deletion, copying, etc. by those without copyright consent. Starting from the requirements of protecting the copyright of electronic products and the security of copyright protection information such as secret communications, digital watermark copyright protection technology came into being. Digital watermark technology can be used as an effective solution to protect electronic products. In recent years, a large number of scholars have conducted research on digital watermark technology.

Digital watermarking technology refers to embedding some text or pictures with special logos about the author's information into the carrier of text, image, audio, video, etc., without affecting the carrier's page display, video playback, audio quality, etc. Use value, and the embedded information is not easy to be modified and the invisibility of the information is not easy to be found. Digital watermarking technology can be divided into text watermarking, image watermarking, audio watermarking, video watermarking, etc. for different application scenarios. Embedding watermark information into digitized text is an important means of copyright protection and infringement certification.

II. PDF DOCUMENT STRUCTURE

PDF is a file format used to represent documents in a way that is independent of the application software, hardware and operating system used to create the document, and the output device on which the document is to be displayed or printed[1]. PDF documents consist of multiple sets of objects that together describe the appearance of one or more pages, and may also contain other interactive elements and higher-level application data. The PDF file contains the objects that make up the PDF document and related structural information, all of which is represented as a single sequence of bytes. Document pages (and other visual elements) can contain any combination of text, graphics, and images. The PDF content stream describes the appearance of the page and contains a series of graphical objects to be drawn on the page.

The composition of the PDF document can be divided into-object, file structure, document structure, content flow, as shown in Figure 1.

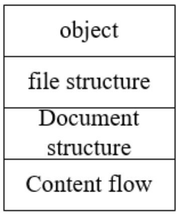


Figure 1. PDF document composition.

978-1-7281-8138-7/20/$31.00 ©2020 IEEE
DOI 10.1109/ICCST50977.2020.00064

III. WATERMARK TECHNOLOGY OVERVIEW

A. The concept of text watermarking)

Digital watermarking technology refers to embedding some special information, namely digital watermarks (numbers, text, pictures, etc.) directly into digital media such as multimedia and software, or indirectly embedding watermarks by modifying the content and form structure of some special areas of the carrier. It is a technology that can reverse extract the special information embedded in the carrier content or structure. Text watermarks can be divided into watermarks based on different text document types such as PDF, WORD, EPUB, etc. according to their carriers [2].

B. Text watermark encryption

A complete data encryption system should include the original data, the key used for encryption and decryption, the encryption algorithm, the decryption algorithm, and the encrypted data[3]. The security of the encryption system mainly depends on the key guarantee, and has little relationship with the encryption algorithm adopted, so the storage of the key is the key to the security of the entire encryption system.

Encryption is to convert data into concealed data by a certain law. The message to be hidden is called plaintext M, the encryption function is E, the key is K, and the encrypted ciphertext is C, as in (1):

$$E_K(M) = C \tag{1}$$

The process of turning ciphertext into plaintext is called decryption, and the decryption function is D, as in (2):

$$D_K(C) = M \tag{2}$$

Decryption and encryption are reversible, the relationship of formula (3) should be satisfied:

$$D_K\big(E_K(M)\big) = M \tag{3}$$

C. Existing algorithms for text watermarking

In summary, text is divided into three categories: ① unformatted text; ② formatted document files, common are Word, WPS, PDF, PostScript, etc.; ③ image documents formed by pixel matrix [3].

In recent years, scholars have proposed different watermarking algorithms for different types of carrier text. The existing text watermarking algorithms are mainly divided into five categories:

a) watermarking algorithms based on document format.

Line spacing coding is based on fine-tuning the line spacing of the document to embed watermark information [4].Word-spacing coding refers to embedding watermark information by changing the distance between certain characters in a line of character information. The character distance of the adjacent position of the changed character is not changed for reference [5]. Space encoding is also called invisible encoding. Generally, watermark information will be hidden by replacing or adding some regular spaces [6].

b) watermarking algorithms based on document structure.

Liu proposed a watermark embedding method based on the structure of PDF documents by analyzing the physical and logical structure of PDF files[7]. It reads the tail information of PDF files to find the cross-reference table and then obtains the object information. By forging legal page objects and modifying the discard Page objects to embed watermark information. Zhong analyzed the physical structure of the PDF document to find an end-of-row identifier in the cross-reference table that did not affect the output of the document content. By modifying the end-of-row identifier to \r\n, the end of the row was modified when the embedded information was "1" The identifier is \n, when the embedded information is "0", it is not modified, so as to protect the copyright[8].

c) based on natural language processing Watermarking algorithm.

Sun XM embed watermark information based on some Chinese characters in the text content as a left-right structure[9].

d) Watermarking algorithm based on traditional binary image processing.

Wu M used the idea of traditional binary image processing to perform parity classification on the black and white pixels after blocking to perform watermark hiding[10].

e) Other algorithms.

IV. DESIGN OF DOUBLE WATERMARK INFORMATION ALGORITHM BASED ON PDF DOCUMENT STRUCTURE

In this section, based on the in-depth study of the existing text watermarking algorithm and analysis of the PDF structure, this paper proposes two watermarking algorithms that are robust to watermarking and verify the integrity of the PDF document content. By replacing the end of the file in the PDF document structure The ID values of the first string and the second string form a PDF document containing dual watermark information.

A. Watermark information generation method

The original information for text watermark processing is generally text or characters with special properties representing the copyright owner's information. The DES encryption symmetric algorithm can meet its encryption requirements, which is high in efficiency, fast, and time-consuming. The watermark information generation algorithm process is as follows, the process is shown in Figure 2:

Input: Original watermark information.

Output: Encrypted hexadecimal number.

Step 1: Enter the original watermark information and key.

Step 2: DES encryption.

Step 3: Output the encrypted hexadecimal sequence.

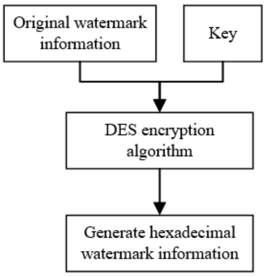

Original watermark information
Key
DES encryption algorithm
Generate hexadecimal watermark information

Figure 2. Watermark generation flow chart.

B. *Robust watermark information embedding method resistant to attack*

The anti-attack watermark embedding algorithm based on the structure of the PDF document is to embed the watermark according to the watermark method of modifying the first string of ID values in the physical structure of the PDF, and generate verification watermark information containing the copyright owner of the document. The algorithm design of the embedded watermark information is as follows，and the flow is shown in Figure 3.

Input: watermark information, key, original PDF document.

Output: PDF document after embedding robust anti-attack watermark information.

Step 1: perform DES encryption preprocessing on the watermark information with the key to generate hexadecimal watermark information.

Step 2: Read PDF.

Step 3: Analyze the PDF document and read the PDF content stream.

Step 4: Read the end information of the PDF file, and replace the first string of IDs in the end information of the PDF file with a hexadecimal watermark sequence.

Step 5: Save the modified physical structure and regenerate the PDF file for output.

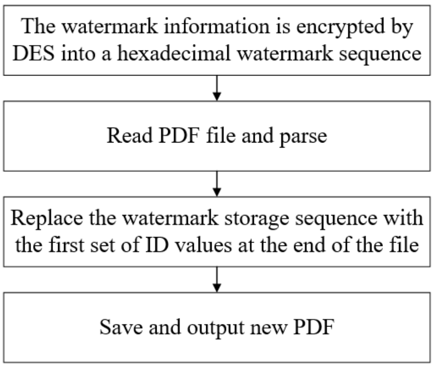

The watermark information is encrypted by DES into a hexadecimal watermark sequence
Read PDF file and parse
Replace the watermark storage sequence with the first set of ID values at the end of the file
Save and output new PDF

Figure 3. Robust watermark extraction.

C. *Robust watermarking information extraction method against attack*

Extract the watermark information embedded in the first string of ID values in the PDF physical structure for copyright verification. The robust watermark extraction steps based on the PDF document structure's resistance to attack are as follows, and the flow is shown in Figure 4.

Input: Key, PDF document with watermark information.

Output: Verify the watermark information of the copyright owner.

Step 1: read the PDF document.

Step 2: Analyze the PDF document to obtain the end-of-file information.

Step 3: Extract the first string sequence of ID values in the tail of the PDF file to form a watermark sequence.

Step 4: Use the key to decrypt the watermark sequence extracted in the third step and output the watermark information.

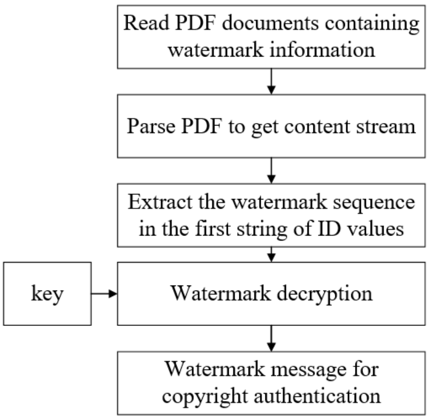

Read PDF documents containing watermark information
Parse PDF to get content stream
Extract the watermark sequence in the first string of ID values
key
Watermark decryption
Watermark message for copyright authentication

Figure 4. Watermark extraction to verify content integrity.

D. *Watermark information embedding method for verifying content integrity*

The watermark embedding algorithm for verifying content integrity based on the structure of the PDF document is to embed repeated watermarks according to the watermarking method of modifying the first string ID value and the second string ID value in the PDF physical structure to generate watermark information that can verify the integrity of the content. The algorithm design steps for embedding watermark information are as follows, and the flow chart is shown in Figure 5.

Input: watermark information, key, original PDF document.

Output: PDF document after embedding robust anti-attack watermark information.

Step 1: perform DES encryption preprocessing on the watermark information with the key to generate hexadecimal watermark information.

Step 2: Read PDF.

Step 3: Analyze the PDF document and read the PDF content stream.

Step 4: Read the end information of the PDF file, and replace the first string of IDs in the end information of the PDF file with a hexadecimal watermark sequence.

Step 5: Replace the second string of IDs in the end information of the PDF file with a hexadecimal watermark sequence.

Step 6: Save the modified physical structure and regenerate the PDF file for output.

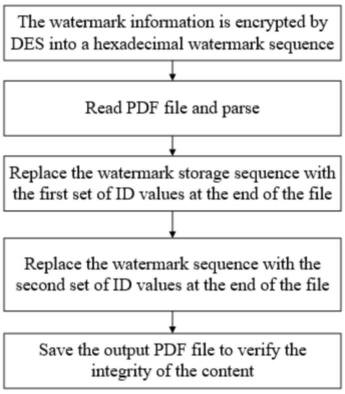

The watermark information is encrypted by DES into a hexadecimal watermark sequence
Read PDF file and parse
Replace the watermark storage sequence with the first set of ID values at the end of the file
Replace the watermark sequence with the second set of ID values at the end of the file
Save the output PDF file to verify the integrity of the content

Figure 5. Flow chart of watermark embedding algorithm.

E. Watermark information extraction method for verifying content integrity

The watermark information embedded in the first string ID value of PDF physical structure is extracted for copyright verification, and the watermark information embedded in the second string ID value of PDF physical structure is extracted to compare with the watermark information extracted from the first string ID value to verify the content integrity. The double watermark extraction steps based on PDF document structure are as follows, and the flow is shown in Figure 6.

Input: key, PDF document with watermark information.

Output: verify whether the watermark information and document of copyright owner have been changed.

Step 1: read PDF document.

The second step: analyze the PDF document to get the tail information.

Step 3: extract the first string of string sequence of ID value in PDF file tail to form watermark sequence 1

Step 4: extract the second string sequence of ID value in PDF file tail to form watermark sequence 2.

Step 5: decrypt the watermark sequence extracted in the third step with the key and output it.

Step 6: compare whether the watermark sequences extracted in step 4 and step 3 are consistent. If they are consistent, the output document is complete; if not, the output document is changed.

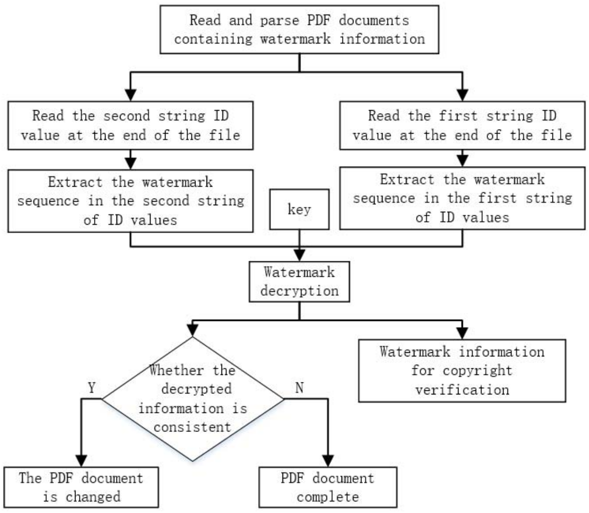

Read and parse PDF documents containing watermark information
Read the second string ID value at the end of the file
Read the first string ID value at the end of the file
Extract the watermark sequence in the second string of ID values
key
Extract the watermark sequence in the first string of ID values
Watermark decryption
Whether the decrypted information is consistent
Y
N
Watermark information for copyright verification
The PDF document is changed
PDF document complete

Figure 6. Flow chart of watermark extraction algorithm.

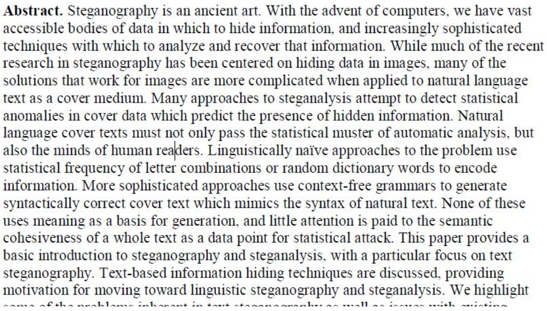
V. SIMULATION EXPERIMENT AND RESULT ANALYSIS

The experimental test environment is: processor Inter(R) Core(TM) i7-6500 CPU @ 2.50GHz 2.60 GHz; memory 8G; operating system Windows 10; software platform Microsoft Visual Studio 2010, PDF document editor Adobe Acrobat DC, PDF document Reader Adobe Reader XI.

A. Experimental test

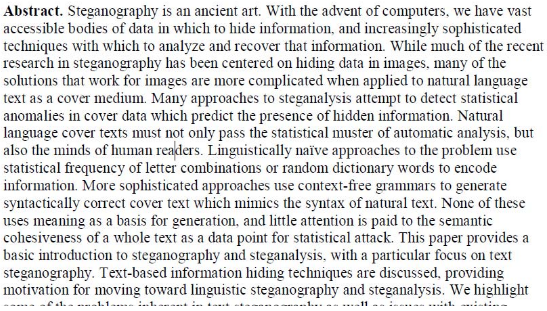
The experiment selected a PDF document named "bennet" downloaded from IEEE as the test object. The embedded watermark information is: The weather is very good today. Figure 7 is a partial screenshot of the original PDF, Figure 8 is a screenshot of the PDF after embedding robust anti-attack watermark information, and The extracted watermark information is shown in Figure 9.

Figure 10 is a screenshot of the PDF after embedding watermark information to verify the integrity of the content，and The extracted watermark information is shown in Figure 11.



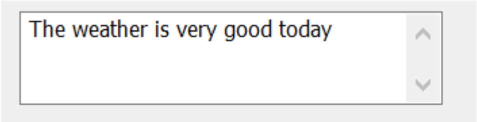
Abstract. Steganography is an ancient art. With the advent of computers, we have vast accessible bodies of data in which to hide information, and increasingly sophisticated techniques with which to analyze and recover that information. While much of the recent research in steganography has been centered on hiding data in images, many of the solutions that work for images are more complicated when applied to natural language text as a cover medium. Many approaches to steganalysis attempt to detect statistical anomalies in cover data which predict the presence of hidden information. Natural language cover texts must not only pass the statistical muster of automatic analysis, but also the minds of human readers. Linguistically naïve approaches to the problem use statistical frequency of letter combinations or random dictionary words to encode information. More sophisticated approaches use context-free grammars to generate syntactically correct cover text which mimics the syntax of natural text. None of these uses meaning as a basis for generation, and little attention is paid to the semantic cohesiveness of a whole text as a data point for statistical attack. This paper provides a basic introduction to steganography and steganalysis, with a particular focus on text steganography. Text-based information hiding techniques are discussed, providing motivation for moving toward linguistic steganography and steganalysis. We highlight

Figure 7. Original PDF document.



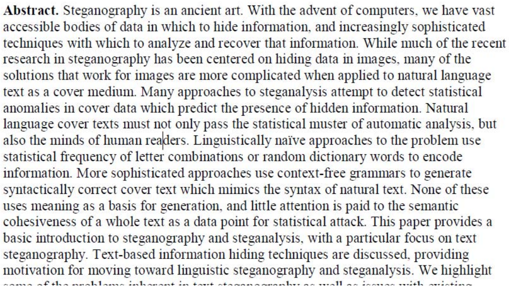
Abstract. Steganography is an ancient art. With the advent of computers, we have vast accessible bodies of data in which to hide information, and increasingly sophisticated techniques with which to analyze and recover that information. While much of the recent research in steganography has been centered on hiding data in images, many of the solutions that work for images are more complicated when applied to natural language text as a cover medium. Many approaches to steganalysis attempt to detect statistical anomalies in cover data which predict the presence of hidden information. Natural language cover texts must not only pass the statistical muster of automatic analysis, but also the minds of human readers. Linguistically naïve approaches to the problem use statistical frequency of letter combinations or random dictionary words to encode information. More sophisticated approaches use context-free grammars to generate syntactically correct cover text which mimics the syntax of natural text. None of these uses meaning as a basis for generation, and little attention is paid to the semantic cohesiveness of a whole text as a data point for statistical attack. This paper provides a basic introduction to steganography and steganalysis, with a particular focus on text steganography. Text-based information hiding techniques are discussed, providing motivation for moving toward linguistic steganography and steganalysis. We highlight

Figure 8. A PDF file with embedded anti-attack watermark information



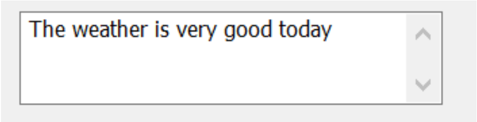
The weather is very good today

Figure 9. Anti attack watermark extraction information display.



Abstract. Steganography is an ancient art. With the advent of computers, we have vast accessible bodies of data in which to hide information, and increasingly sophisticated techniques with which to analyze and recover that information. While much of the recent research in steganography has been centered on hiding data in images, many of the solutions that work for images are more complicated when applied to natural language text as a cover medium. Many approaches to steganalysis attempt to detect statistical anomalies in cover data which predict the presence of hidden information. Natural language cover texts must not only pass the statistical muster of automatic analysis, but also the minds of human readers. Linguistically naïve approaches to the problem use statistical frequency of letter combinations or random dictionary words to encode information. More sophisticated approaches use context-free grammars to generate syntactically correct cover text which mimics the syntax of natural text. None of these uses meaning as a basis for generation, and little attention is paid to the semantic cohesiveness of a whole text as a data point for statistical attack. This paper provides a basic introduction to steganography and steganalysis, with a particular focus on text steganography. Text-based information hiding techniques are discussed, providing motivation for moving toward linguistic steganography and steganalysis. We highlight

Figure 10. Embed a watermarked PDF that verifies the integrity of the content



The weather is very good today

Figure 11. Watermark extraction information display to verify content integrity.

It can be seen from the comparison between FIG. 7 and FIG. 8 and FIG. 7 and FIG. 9 that the PDF document and the original document embedded with the watermark information have no difference in human eye observation and have good invisibility.

B. Algorithm performance analysis

1) Document byte change: The double watermarking algorithm based on the structure of the PDF document is to modify and replace the ID value based on the original structure of the PDF, so the document byte size will not change.

2) Algorithm capacity analysis: The watermark embedding bit is a 32-bit hexadecimal string in the PDF file tail, so 128-bit information can be embedded.

3) Watermark invisibility analysis: It can be seen from the comparison between FIG. 7 and FIG. 8 and FIG. 7 and FIG. 9 that the PDF document and the original document embedded with the watermark information have no difference in human eye observation and have good invisibility.

4) Robustness analysis: The watermark information can still be extracted completely after performing text attack and page attack on the PDF with two different algorithms embedded in the watermark information.

VI. CONCLUSION

This paper studies the watermarking algorithm designed for the structure of the text carrier, and compares the same watermarking algorithm based on the document structure. The comparison effect is shown in Table 1.

In this paper, through the characteristics of the existing text watermarking technology and the structure of the PDF document, by replacing the ID value in the end of the PDF file, embedding robust watermarking information resistant to attack and verifying the integrity of the content of the watermarking information. The algorithm uses the ID value originally present in the PDF file as the embedding option without changing the document size. Experiments show that the algorithm can still extract the watermark completely after text attack and page attack, and it has good robustness.

TABLE I. WATERMARK ANTI ATTACK ABILITY OF DIFFERENT ALGORITHMS

Attack type	Algorithm	Literature[26]	Literature[26]	Literature[26]
Text attack	Does not affect watermark extraction	Insertion and deletion attacks affect watermark extraction	Does not affect watermark extraction	Does not affect watermark extraction
Page attack	Does not affect watermark extraction	Page deletion and cropping affect watermark extraction	Influence watermark extraction	Influence watermark extraction
Byte statistics	Can resist	Can resist	Irresistible	Irresistible
Watermark removal	Strong resistance	Strong resistance	Weak resistance	Weak resistance

ACKNOWLEDGMENT

This work was supported by the National Key R&D Program of China (2017YFB1401000) and the Key R&D Program of Shanxi (201903D421007). It was also the research achievement of the Key Laboratory of Digital Rights Services, which is one of the National Science and Standardization Key Labs for Press and Publication Industry.

REFERENCES

[1] Adobe PDF reference version 1.7:http://www.adobe.com.

[2] Liu Minhao, Zhang Ru, Niu Xinxin. Summary of research on text digital watermarking technology [J]. Journal of Southeast University (Natural Science Edition), 2007, 37(zl): 225-230.

[3] Wang Bingxi, Chen Qi, Deng Fengsen. Digital watermarking technology[M]. Xi'an: Xidian University Press, 2003.

[4] Brassil J, Low S,Maxemchuk NF, et al.Electronic marking and identification techniques to discourage document copying[J]. IEEE Journal on Selected Areas in Communications. 1995, 13(8):1495-1504.

[5] Brassil J, Low S, Maxemchunk NF. Copyright protection for the electronic distribution of text documents[J]. Proceedings of the IEEE 1999, 87(7):1181-1196.

[6] Fu Yu,Wang Baobao.Realization and performance of text watermarking with additional space coding method[J].Journal of Changan University (Natural Science Edition),2002(03):85-87.

[7] Liu Youji, Sun Xingpeng, Luo Gang. A new information hiding algorithm based on PDF document structure [J]. Computer Engineering, 2006, 32(17): 230-232.

[8] Zhong Zhengyan, Guo Yanhui. Digital watermarking algorithm based on PDF document structure [J]. Computer Applications, 2012, 32(10): 2776-2778.

[9] Sun XM, Luo G, Huang HJ. Component-based digital watermarking of Chinese texts[C]. Proceedings of the Third International Conference on InformationSecurity. Shanghai, China, 2004,85:76-81.

[10] Wu M,Liu B.Data hiding in binary image for authentication and annotation[J]. IEEE Transactions on Multimedia, 2004, 6(4):528-538.

[11] Xu Zhen. Design and implementation of electronic document copyright protection system [D]. Chengdu: University of Electronic Science and Technology of China. 2013.

[12] Li Gaoyuan. Two new PDF document watermarking algorithms [D]. Xi'an: Xi'an University of Science and Technology, 2016.

[13] Gu Yanchun, Feng Junting. A PDF text digital watermarking algorithm based on space encoding[J]. Journal of Foshan University of Science and Technology, 2015, 33(1): 76-80+87.