International Journal of Frontiers in Sociology
ISSN 2706-6827 Vol. 2, Issue 1: 1-9, DOI: 10.25236/IJFS.2020.020101

Research and Implementation of Text Watermarking Technology Based on PDF Document Structure

Weijuan Zhao[1,*], Hu Guan[2],Shuwu Zhang[3]

[1] *College of Information and Communication, Communication University of China,Beijing 10033, China*
[2]*Digital Content Technology and Service Research Center, Institute of Automation,Chinese Academy of Sciences, Beijing 100080, China*
[*]*575869686@qq.com*

ABSTRACT. *Aiming at the problems of illegal copying, malicious tampering, information leakage, copyright disputes and other issues caused by the convenience of the Internet, a watermarking algorithm based on the PDF document structure was developed, a watermarking algorithm model based on the PDF document structure was established, and a DES algorithm was used to encrypt the watermark. Based on the research of the physical structure and logical structure of the PDF document, the algorithm of modifying the color attribute value of the text to embed the watermark information is proposed. Experimental results show that the algorithm achieves information hiding and meets the requirements of digital copyright protection, has good security, and can resist multiple attacks.*

KEYWORDS: *PDF document, DES algorithm, watermark, color attribute value，information hiding*

1. Introduction

Digital multimedia is also called electronic publication, such as digital voice, image, music, video and so on. It can be quickly, accurately acquired, transmitted, and stored, but it also brings more challenges and security issues. Such as illegal copying, malicious tampering, copyright disputes, information leakage, etc. Pirates use these features of digital products to be easily copied, processed, and disseminated to undermine the legitimate rights of manufacturers and users for personal gain. Therefore, how to make full use of the convenience of the network and effectively protect intellectual property rights has been highly valued by people. Under this background, digital watermarking technology is proposed.

Digital watermarking is an important and effective way to protect the copyright of digital media. It embeds some identification information (that is, digital

ISSN 2706-6827 Vol. 2, Issue 1: 1-9, DOI: 10.25236/IJFS.2020.020101

watermarking) into digital multimedia, but it does not affect the use value of the original carrier, and it is not easy for people to perceive the system (such as Visual or auditory system). Due to the special nature of the text carrier, there is little or no redundant space for embedding watermarks, leading to rare and breakthrough research results on text digital watermarking, most of which are still in the theoretical research stage. Therefore, how to improve the performance of text digital watermarking in terms of capacity, robustness, and concealment, and to develop high-performance text digital watermarking and apply it to digital copyright protection systems are particularly urgent. It is important to study text digital watermarking technology. Research significance and practical value.

2. Analysis of Research and Development of Text Watermarking Technology

2.1 Text watermarking technology overview

Digital watermark (refers to the identification information embedded in the carrier data, such as numbers, serial numbers, text, images, etc.As a type of digital watermark, the carrier of the digital watermark is a text document. In general, text is divided into three major categories [1]: unformatted text, formatted document files, and document images that describe content in a pixel matrix manner. Text digital watermarking technology usually combines the characteristics of text structure and content to modify the document format or content in a specific way to embed specific information (digital watermark) representing the identity of the copyright owner into the text document Without affecting the value of the document.

Text digital watermarking algorithms are mainly divided into:

a) Text watermarking algorithm based on document format;

b) Text watermarking algorithm based on natural language processing technology;

c) Text watermarking algorithm based on traditional binary image watermarking method;

d) Text watermarking algorithm based on document structure;

e) Other algorithms.

The structure of the document is very compact with almost no redundant information. Therefore, the development of text watermarking algorithms is very slow. The research on text watermarking algorithms is very difficult, and the literatures on text digital watermarking at home and abroad are less than those of other carrier types.

2.2 A Review of Existing Algorithms for Text Watermarking

a) Text watermarking algorithm based on document format

ISSN 2706-6827 Vol. 2, Issue 1: 1-9, DOI: 10.25236/IJFS.2020.020101

Brassil [2-3], etc. proposed three well-known text watermarking encoding techniques and implemented them in Postscript documents. They use the characteristics of the document to complete the embedded watermark by slightly adjusting the document format, including: line spacing encoding, word spacing encoding, and feature encoding. There is basically no formatting information available in unformatted text, and the hiding method is generally invisible encoding.

b) Text watermarking algorithm based on natural language processing technology

At present, natural language text digital watermarking methods are mainly divided into two categories: one is based on syntactic structure, and the other is based on semantics. Sun XM, Wang Bingxi, et al. [4-5] constructed a syntactic tree and a grammatical tree by understanding the content of the full text, and then performed certain operations on the syntactic tree and the grammatical tree-grafting, cutting, and equivalent substitution. Embed watermark signal. Bennett K [6] proposed the use of syntax to embed watermark signals, but its watermark capacity is seriously insufficient. Victor Raskin [7] proposed an information hiding method using the "" structure in the text. Without changing the original meaning of the sentence, the article uses the addition and subtraction of certain "" characters to embed the watermark signal.

c) Text watermarking algorithm based on traditional binary image watermarking method

The basic idea of a text watermarking algorithm based on a binary image is an algorithm that treats the image of the binarized text as a special binary image, then divides the binary image into blocks, and embeds the watermark signal in each block. The most representative one is the method proposed by Wu [8], which realizes embedding watermark signals by modifying the parity of the number of black and white pixels in the block. Many scholars have improved on the basis of Wu and achieved some results. For example, Zhao Xingyang [9] proposed a document watermarking algorithm based on the adjustment of the character's step edges. It believes that the fine-tuning of the step edges of characters in a binary text image cannot be perceived by human vision.

d) Text watermarking algorithm based on document structure

Liu Youji [10] and others proposed a new large-capacity information hiding algorithm based on the structure of PDF documents. The secret information was preprocessed and disguised as a legitimate PDF object. It was embedded into the carrier file in the form of a file stream operation. Satisfying the embedded information does not affect the output of the file in the reader, editor, and printer. Information hiding and detection of PDF documents is realized. Zhong Zhengyan [11] proposed a digital watermarking algorithm based on PDF document structure. This algorithm utilizes the feature that row unidentifiers are not displayed in the document, and achieves indirect embedding of watermark information by equivalently replacing row unidentifiers of cross-reference tables with fixed formats in PDF documents.

ISSN 2706-6827 Vol. 2, Issue 1: 1-9, DOI: 10.25236/IJFS.2020.020101

e) Other algorithms

In addition to the above four types of algorithms, there are other types of algorithms. For example, Qingcheng Li, Paulo Borges [12-13] gives a way to describe Chinese characters with mathematical expressions, and can embed and detect/extract watermark signals without changing the content of the document. This is a brand new .The development direction of Chinese document watermarking algorithms enables mature mathematical tools to directly process text, thereby avoiding processing images.

3. Digital Watermarking Algorithm Based on PDF Document Structure

3.1 PDF document structure analysis

a) PDF document physical structure

The physical structure of a PDF document includes four parts: the file header, the file body (the page object), the cross-reference table (the index directory), and the end of the file. A PDF document reader starts reading a PDF document from the end of the file. According to the end of the file, it finds the cross-reference table and the root directory of the entire PDF document, thereby displaying the document. The physical structure of a PDF document is shown in Figure 1.

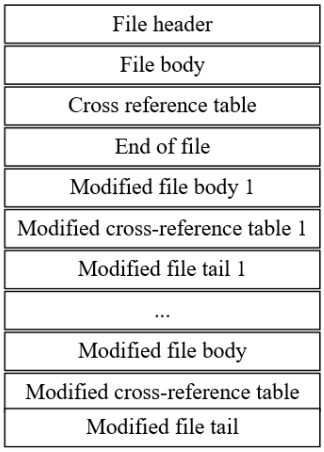


Figure. 1 PDF document structure analysis

File header : Indicates the version number of the PDF specification to which this file complies, which appears on the first line of the PDF file.

ISSN 2706-6827 Vol. 2, Issue 1: 1-9, DOI: 10.25236/IJFS.2020.020101

File body : Also called object collection, the most important part of a PDF file, all objects used in the file, including text / image / music / video / font / hyperlink / encryption information / document structure information, etc., are defined here.

Cross-reference table : An address index table of an indirect object set up for random access to the indirect object. The object address is actually stored in an offset and index manner.

End of file : The address of the cross-reference table is declared, that is, the root object of the file body is specified, so that the position of each object body in the PDF file can be found and random access can be achieved. It also saves security information such as PDF file encryption.

b) PDF document logical structure

The logic of PDF is generally a tree structure, the root node is the catalog dictionary, through which to parse pages, directories, link information, and so on....

A PDF document can be regarded as a tree structure describing all the objects in the body part. Its root node is the dictionary of the root directory of the document. Most of the objects in the tree structure are dictionaries. For example, each page of a document is represented by a page object, which is a page content and other attributes that include references, such as its thumbnail and any annotations associated with it. The individual page objects are linked together by a page tree, which is in turn positioned by indirect references in the document root. The relationship between father nodes, child nodes, and sibling nodes in this hierarchy is realized by dictionary objects, that is, the value of a dictionary entry refers to other dictionary objects indirectly.

The root directory has four leaf nodes, which are the four root nodes of Pages tree, Outlines hierarchy, Articles threads, and Names destinations. The root directory node is a directory dictionary, which is usually located at the end of the PDF document file. The root directory contains references to objects that define document content, bookmarks, indexes, names, and other attribute values. The page tree is the most important part of a PDF document. All the page objects describing the content of the document are leaf nodes of the page tree. The page tree first defines the root node of each page of the PDF document, and then the leaf nodes of these root nodes will describe all the content of the page, including Content stream, Thumbnail image, Annotation, etc. The bookmark tree is a tree structure that defines the bookmarks used in the document, and the leaf nodes are the bookmarks in the document. The nodes of the first level of the index tree are the indexes of the document, and the leaf nodes are the specific indexes, which contain the relationship and position of the indexes. In the name tree, the name objects used in the document are re-listed in the form of a tree structure, which is convenient for searching and other operations.The logical hierarchy of the PDF document is shown in Figure 2.

ISSN 2706-6827 Vol. 2, Issue 1: 1-9, DOI: 10.25236/IJFS.2020.020101

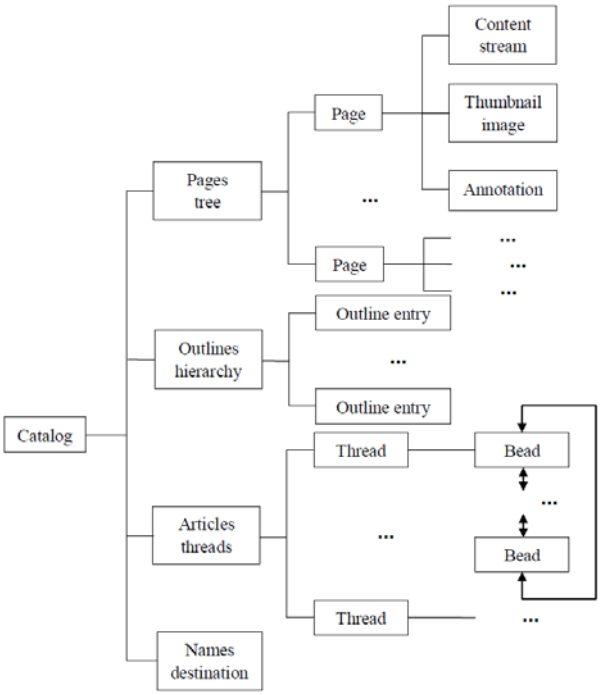


Figure. 2 Logical hierarchy of PDF documents

c) PDF color space basic characteristics

PDF provides powerful features for color image generation of pages. The color function can be divided into two parts: color specification and color conversion. Table 1 summarizes the color families supported by PDF. Gray gray space, only the first bit is used to represent the corresponding gray level. In RGB space, three bits are used to represent the color value. This article is to embed the corresponding watermark data in the three bits in the RGB space, that is, the red, green, and blue components.

Table 1 color space familes

DEVICE	CIE-BASED	SPECIAL
DeviceGray(PDF 1.1)	CalGray(PDF 1.1)	Indexed(PDF 1.1)
DeviceRGB(PDF 1.1)	CalRGB(PDF 1.1)	Pattern(PDF 1.2)
DeviceCMYK(PDF 1.1)	Lab(PDF 1.1)	Separation(PDF 1.2)
	ICCBased(PDF 1.3)	DeviceN(PDF 1.3)

ISSN 2706-6827 Vol. 2, Issue 1: 1-9, DOI: 10.25236/IJFS.2020.020101

d) Watermark information generation method

Read the watermark information, encrypt the watermark information using the DES algorithm, and then encode the watermark ciphertext into a binary data stream. The process is shown in Figure 3.

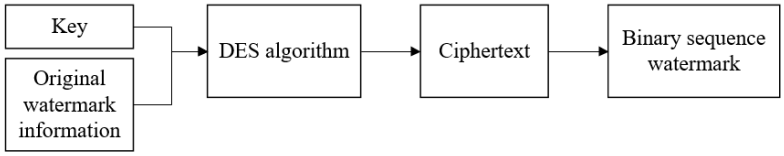


Figure. 3 Watermark information generation

e) Watermark information embedding method

It is necessary to first convert the color space to convert the color space of the text into RGB mode. In this way, the description of the character color in the PDF document is $(a_1, a_2, a_3), a_i \in (0,1)$, 0-1 corresponds to 0-255 levels, because the default least significant digit of ai is one after the decimal point. Bit, the watermark is embedded by modifying the parity of the least significant bit of a_i, that is, $a_i \in (0.0,0.1)$.Select the embedding position uniformly in the embedding point, extract the color attribute values (a_1, a_2, a_3) of the characters that change the position, and modify the three-color part of the color attribute value according to the watermark information. If the watermark information is $W_i = 1$, modify the last digit of the color attribute value to 1, otherwise it is 0.

f) Watermark information extraction method

It is worth the last bit to extract the color attribute of the character at the embedding position. If it is 1, the watermark information $W_i = 1$; otherwise $W_i = 0$.Form watermark information W.

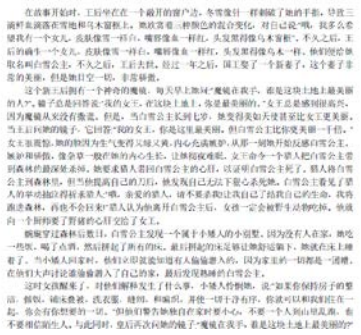
4. Simulation results

a) Watermark generation is shown in Figure 4.

Figure. 4 Watermark information generation

ISSN 2706-6827 Vol. 2, Issue 1: 1-9, DOI: 10.25236/IJFS.2020.020101

b) By modifying the watermark information embedded in the character color, the original image is shown in Figure 5, and the result is shown in Figure 6.



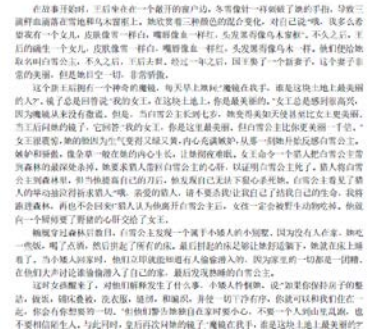


Figure. 5 Original PDF document Figure. 6 Embed watermark PDF document

4. Conclusion

This paper adopts the background and significance of PDF security issues, PDF file formats, encryption standards and currentThere is analysis and research of PDF digital copyright protection system. Data confidentiality and integrity against existing technologiesThe existing defects have been improved accordingly. In terms of protecting data integrity, text-based numbersWord watermarking technology to ensure the integrity of document data and avoid the shortcomings of digital signature technology.

References

[1] Liu Haohao, Zhang Ru. Review of Research on Text Digital Watermarking Technology Journal of Southeast University,2007，37(zl):225-230.

[2]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.

[3]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.

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

[5] Wang Bingxi, Chen Qi, Deng Fengsen.Digital Watermarking Technology [M] .Xi'an: Xi'an University of Electronic Technology Press, 2003.

ISSN 2706-6827 Vol. 2, Issue 1: 1-9, DOI: 10.25236/IJFS.2020.020101

[6]Bennett K. Linguistic steganography: survey, analysis, and robustness concerns for hiding information in text [EB/OL]. Center for Education and Research in Information Assurance and Security, Purdue University, West Lafayette.2004.
[7]Mikhail Atallah J, Victor Raskin, Christian F Hempelmann. Natural language watermarking and tamper proofing[C]. The 5th International Information Hiding Workshop. Berlin,2002: 196-212.
[8]Wu M, Liu B.Data hiding in binary image for authentication and annotation[J]. IEEE Transactions on Multimedia,2004, 6(4):528-538.
[9] Zhao Xingyang, Sun Jiyin, Li Linlin.A Text Watermarking Algorithm Based on Character Step Edge Adjustment [J] .Computer Applications, 2008, 28(12):3175-3178.
[10] Liu Youji, Sun Xingpeng, Luo Gang.A New Information Hiding Algorithm Based on PDF Document Structure [J] .Computer Engineering,2006,32(17):230-232.
[11] Zhong Zhengyan, Guo Yanhui.Digital Watermarking Algorithm Based on PDF Document Structure [J] .Computer Applications,2012,32(10):2776-2778.
[12]Qingcheng Li, Jin Zhang, Zhenhua Zhang, et al. A Chinese text watermarking based on statistic of phrase frequency[C]. 2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP). Harbin, 2008, 8. pp:335-338.
[13]Paulo Borges, Ebroul Izquierdo, Joceli Mayer. Efficient side information encoding for text hardcopy documents[C]. 2007 IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS). London, UK, 2007, 9. pp.552-557.