Text Watermarking Algorithm for Word Document

Bharati D. Patil
Assistant Professor
ME (CSE), BE (COMP)
R.C.P.I.T, Shirpur

Shailendra M. Pardeshi
Assistant Professor
ME (CSE), BE (COMP)
R.C.P.I.T, Shirpur

Puja D.Saraf
Assistant Professor
MTech (IT), BE (COMP)
R.C.P.I.T, Shirpur

ABSTRACT
In this paper, a novel method of robust watermarking for word documents for protecting copyright and controlling dissemination is proposed. A word document consists of lots of objects and word objects are arranged in a hierarchical order. Every object has its own properties and functions. Some of the properties are special. They can't be modified via the interface of word application but via programming only and their change will not cause any differences on the page. These special properties are suitable for hiding information. Watermarking which includes both the author and the legal user's information is embedded in these special attributes of Word objects circularly after encryption, grouping and packing into the message. As the experimental results shown, this watermarking scheme performances excellent on robustness. After all kinds of attacks, including adjusting the features, deletion, insertion and replacement, and even massive attack, the watermarking can still be extracted from word document. It has a promising prospect of wide use fields of the copyright protection and text delivery on the internet and it applies to both English and Chinese language.

Keywords
Watermarking, Secret Key, Watermark Embedding, Watermark Extraction.

1. INTRODUCTION
The digital watermarking is a new branch of information security technology and it is used for copyright protecting and integrity authentication of digital products. It was put forward in 1990's and developed fast. Watermarking is information embedded into the host digital multimedia including images, videos, audios, text and so on [1]. According to the types of the host media, digital watermarking can be classified into four categories: image watermarking, video watermarking, audio watermarking and text watermarking. The former three are similar in principle and most of the research work has been done are about them. Text watermarking is new and it develops slowly because texts are too simple to add other information. However, as computer technologies and internet developed, in enterprise, organizations and governments, digital documents such as word documents, excel documents and pdf etc. play an important role in normal running. Text watermarking is in urgent need. Comparing with fragile text watermarking, robust text water marking gains more attention and many works has been done. Brassil、Maxemchuk and low[3,4] proposed methods which modifies the spacing between lines and words for hiding. information. Someone proposed to change the features of font such as font color, font size and similar fonts for coding [2]. Whereas, both of the methods are not robust enough. These scheme based on semantics of nature language are developing quickly. It can apply to both simple text and formatted text but it's very complex and it will take a long time before it's put to use. Robustness is the most important thing of watermarking methods, but most existing schemes can't achieve the aim. In this paper, a scheme for word document, which hides information in the special attributes of word objects and performances excellent on robustness, is proposed This paper is organized as follows: Section one briefly introduces the mind of the watermarking scheme. Section two and Section three discusses the watermark-embedding algorithm and watermark-detecting algorithm. Section four given the conclusion.

2. ROBUST WATERMARKING SCHEME

2.1 Word Object Model
A word document consists of lots of objects. Word objects are arranged in a hierarchical order and the two main classes at he top of the hierarchy are the Application and Document classes. The application object means the entire application, each document object represents a single Word document, and the Paragraph object corresponds to a single paragraph, and so on. Each of these objects has many methods and attributes that allow users to manipulate and interact with it. Some of the properties are special. They can't be modified by manipulating the word application. If the properties need to be adjusted, users should program with computer language such as VB or C++. These attributes are called special properties in this paper. They are suitable for embedding watermarking for two reasons. First, more information can be hidden in a character for one character has several special properties. Second, any common instructions of the word application will not affect the watermarking.

2.2 Special Properties of Word Object
After lots of test, the special attributes of word object are found as Table I . The Disable Character Space Grid the property of Range object and it's also the property of Font object. In fact, the two properties are the same thing. It is Boolean; one bit can be embedded here. No Proofing property is Boolean too; one bit can be embedded here. Language ID Far East property can choose seven constants in the constants collection Wd Language ID, which includes 172 constants, as its legal value. Any four constants in the seven can be chosen for coding, thus, two bits can be embedded in this property. By the same token, 168 constants in the collection Wd Language ID can be the legal value of Language ID Other property and any 128 constants in them can be chosen for coding, thus 7 bits can be embedded in Language ID Other property. The valuation of Kerning property can be 1 to 4, so, two bits can be embedded here. In a word, 11 bits can be embedded in a Range object and 2 bits can be embedded in a Font object.

Table 1: Special Properties of Word Objects

Objects	Properties	Specification
Range	Disable Character Space Grid	Boolean,Read/Write
	LanguageIDFarEast	WdLanguageID, Read/Write
	LanguageIDOther	WdLanguageID, Read/Write
	NoProofing	Boolean,Read/Write
Font	Kerming	Single,Read/Write
Variable	Name	Document variables are used to preserve macro settings in between macro sessions.
	Value	

Range object represents a contiguous area in a document. Each Range object is defined by a starting and ending character position. A minimum Range object can contain only one character. Font is a object, meanwhile, it is a property of Range object. If a word document has n characters, it has n Range objects and n corresponding Font objects. That is to say, at most 13n bits information can be embedded in a document which has n characters. Variable represents a variable stored as part of a document. Document variables are used to preserve macro settings between macro sessions. The Variable object is a property of document object.

2.3 Encryption

The first step is to encrypt the watermarking information. After encryption, it is more difficult to attack the watermarking for the adversary. Any cipher can be used here, the more complex the encryption algorithm, the more difficult it becomes to attack the watermarking without access to the key. However the topic of this paper is watermarking, we take a simple encryption called Caesar cipher for example. In cryptography, encryption is the process of transforming information, referred to as plaintext, using an algorithm called cipher to make it unreadable to anyone except those possessing special knowledge, usually referred to as a key. The result of the process is encrypted information, referred to as cipher text. The reverse process of encryption is decryption, to make the encrypted information readable again. Encryption can protect the confidentiality of messages. A Caesar cipher, also known as a Caesar's cipher, the shift cipher, is one of the simplest and most widely known encryption techniques. It is a type of substitution cipher in which each letter in the plaintext is replaced by a letter some fixed number of positions down the alphabet. For example, with a shift of 5, A would be replaced by F, B would become G, and so on. The method is named after Julius Caesar, who used it to communicate with his generals.

2.4 Grouping and Packing

Embedding the watermarking information in the document without grouping, when the document is attacked, the whole watermarking information will be destroyed. After dividing the watermarking information into g groups, when the document is attacked, only 1/g of the watermarking is destroyed and the rest parts can be still extracted. After grouping, pack every group of watermarking information into the message. The message consists of five parts (as Fig 1). The first part of the message is Beginning signal which is set to 10101. It is the symbol of the message's start. The second part of the message is the number of the group and it takes 3 bits. The third part is the size (in byte) of the cipher in the message and it takes 5 bits. Parity bits of the fifth part cipher are saved in the fourth part, which takes m bits.

2.5 Watermarking Information

Robust watermarking is used for copyright protecting and dissemination controlling. The watermarking information comprises two parts. The first part is the information of the author and document and it can be texts or an image. The second part is the information of the original legal user and it can be texts, an image or a series of numbers represented the legal user and saved in a database. When copyright infringements about the document happen, the first part of watermarking extracted from the document can prove who the right author is, and the second part can tell people which legal user give the document to an illegal user.

2.6 Embedding Message Circularly

After packing watermarking information into the messages, embed the message one by one according to the group number. After the last message is embedded in the document, embed the messages again like that until the end of the document. Now, the process of embedding watermarking information is finished.

2.7 Detecting the Watermarking

Detecting the watermarking is the reverse process of embedding watermarking information. Firstly, extract the watermarking signals from the document and save the signals as a series of binary codes. Secondly, find the g messages one by one. Take the first message for example, find 10101001 in the binary codes, and then take the neighboring 5bits as m and save the following m bits as check bits and then m bytes as cipher. Compute the parity bits of the cipher and verify whether the check bits are right or not. If the check bits are right, the first message is found and start to find the next message, otherwise, go on finding the first message until the end of the binary codes. Finally, If one of the g messages is not found, the watermarking information is not exist or it's destroyed. If all the g messages are found, extract the cipher part of every message and the ciphertext of the whole watermarking information is got by joining the g cipher parts according to the group number. Then, after decrypting the ciphertext according to the user's secret key, the plaintext of the watermarking is detected.

3. WATERMARKING ALGORITHM

3.1 Watermark Embedding Algorithm

Input: original document T user key k

Output: the covered document T'

Step1: read the watermarking information and user key

Step2: encrypt the watermarking information with user key

Step3: divide the ciphertext into 5 groups compute the parity bits for every group

Step4: pack the ciphertext and parity bits into the message for every group

Step5: open and read document T

Step6: for g: =1 to 5 do According to the character's position in the document and the special properties chosen for coding embed the g-th message into the document from the position p //p is set to 0 at the beginning g++ End

Step7: to the end of the document if yes: to Step8 and if no: to Step6 p=p+ the size of 5 messages

Step8: saveas T' close T

Step9: output T' and user key k

3.2 Watermark Detection Algorithm

Input: the covered document T' user key k

Output: determine if there is watermarking information in T' ， if yes, output it。

Step1: read user key k

Step2: open and read document T'

Step3: According to the character's position in the document and the special properties chosen for coding, extract the watermarking information as a series of binary codes.

Step4: for g: =1 to 5 do Find the beginning signal 10101 and the group No. taking 3 bits in the binary codes from the position p //p is set to 0 at the beginning Is the g-th message found? Yes: to Step5 No: out put: the watermarking is destroyed or not existing. Close T'. exit. End

Step5: extract the third part of the g-th message size m. According to m, extract the fourth and the fifth part of the g-th message.

Step6: compute the parity bits of the fifth part and verify whether the fourth part check bits are right or not. If the fourth part is right, save the g-th message and g++, p=0, to step 4 If the fourth part is not right, p++, to step 4

Step7: extract the fifth part cipher of every group and join them together according to the group No.

Step8: decrypt the ciphertext with the user key

Step9: close document T'

Step10: output the plaintext of watermarking information

4. CONCLUSION

With the development of internet, more and more texts need to be protected. Word document is one of the most popular text documents in daily life. In this paper, a novel text watermarking scheme with good robustness for word document is proposed. The scheme embeds the secret signals in the special properties of word object. The watermarking information is encrypted, divided into several groups and packed into the message before it is embedded in the word document circularly. All this operations make the scheme performance excellent on robustness when attacks happen, comparing with the methods based on the features of character font. This strategy has a promising prospect of wide use fields of the copyright protection and dissemination controlling and it can apply to both English and Chinese language.

5. REFERENCES

[1] I.j.Cox, M.L.Miller. The First 50 Years of Electronic Watermarking [J]. Journal of Applied Signal Processing.2002, 2:126-132

[2] K. Tanaka, Y. Nakamura, K. Matsui. Embedding secret information into a dithered multilevel image[J]. Proc of 1990 IEEE Military Communications Conference, 1990:216-220.

[3] Lu, Fang Dingyi. A New Chinese Text Digital Watermarking for Copyright Protecting Word Document [In]. 2009 International Conference on Communications and Mobile computing.

[4] Mohan, s. K. and K. F. Hau, Watermarking of Electronic Text Documents[J]. Electronic Commerce Research,2002,2(1):169-187.

[5] Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.