Alternative Shift Algorithm for Digital Watermarking on Text

Nikita Pandey[*], Sayani Nandy[*], Shelly Sinha Choudhury[**]

[*] Department of Information Technology, Narula Institute of Technology, Kolkata, India.
[**] Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata, India.

Abstract- Various techniques have been employed till date to ensure the achievement of three basic needs of security: authenticity, integrity and confidentiality. With sharing of documents becoming a mundane phenomenon and the risk of copyright infringement becoming even more predominant with the easy availability of digital media, the need of better ways to ensure integrity and authentication was never more palpable. These growing issues led to the evolution of digital watermarking. In this paper the proposed algorithm generates a watermark which modifies the inter-word space based on the content of the document. The value being too small remains unnoticeable to the human eyes which increases the robustness. The watermark can be extracted to ensure the authenticity and the integrity of the document. The content of the document remains unchanged which ensures one of the advantages of this algorithm.

Index Terms- Digital Watermarking, text, inter-word shift, copyright protection, authentication.

I. INTRODUCTION

The Internet is a series of computers which are connected by either electronics, wireless or optical medium. The Internet has provided us with the option to share information between computers distance between them now being a trivial issue. Data are shared, transferred over long distance as signals. Digital information such as websites, e-paper, e-books, document files, social networking sites mostly contains plain text. With the growth of information sharing there has also been an increase in information duplicity and illegal distribution. Thus with this expansion of technology there is a growing need for information security or protection.

In real life, data or any possession that can be replicated are usually protected by a sign or a symbol. For example Banknote is protected with the various techniques like the use of security thread (the silver security band), latent image, micro lettering (visible only under microscope) and watermark etc. A virtuoso protects his creation by signing it or using a thumb-print or symbol(s).

In virtual world, data - image, audio, video and text or an amalgamation of the two or more are prone to plagiarism, falsification, fabrication, distortion, fraud, infringement, and piracy and even stealing. Out of this plain text is most easily tampered as compared to other digital data. The characters in a text file can easily be read by a character reader and be regenerated, text being most sensitive. Human vision is restricted to detect such changes and redundancy in a plain text. The already existing copyright rules are inefficient to detect such changes and limit the illegal activities. This growing need of security led to new technologies. Digital watermarking is one the methods to authenticate and protect the data. Digital watermark is very similar to steganography and it is based on hiding signal in the document and the signal can be an image, pattern, text or simply some text hidden in the data to be secured. But unlike steganography it does not secure the access to the information, the information is very much shared and distributed and digital watermark holds the authenticity, integrity and identification of the owner(s) of the data. [1]-[2]. The signal which is used as a watermark does not need to be related with the content of the document. The watermark is separately embedded to ensure security without disrupting the contents.

A digital watermarking generally involves three steps: generating and embedding of the watermark, attacking and detecting (Figure 1). The watermarking can be done by various algorithms, the algorithm works on the document, generating and embedding the watermark. In attacking the attacker chooses to change some content or add some content in the document or may even try to remove the watermark. This is detected in the extracting stage where the algorithm is applied on the attacked document to extracted and check for any modification in the watermark.

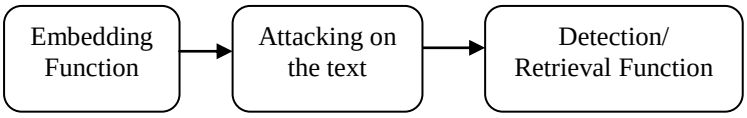

Figure 1: Steps involved in watermarking

A digital watermark may be visible or invisible. The visible watermark is easy to identify as it is noticeable on the screen whereas an invisible is embedded in the document and it's done by changing the bit, inserting noise, masking or some transformation. Generally a watermark should be secure, robust and it should not affect the quality of the signal. The watermark is characterized by robustness and perceptibility. The digital watermarking used for text can be either fragile, that is it should not resist any tampering or it can be Semi-fragile, if it is meant to resist any modification. Robust- this is meant to resist a list of transformations. Perceptibility: imperceptible-this is invisible and perceptible being the visible one. The watermarks also classified on the capacity or length of the data embedded and the algorithm used for embedding the watermark.

The type of watermark used and its use vary in different field. Watermark is not limited to copyright protection but has also found application in ensuring authentication and determining credibility for instance if an image has been compressed watermark is applied to ensure that the content has not been rigged with. Watermark is one essential part of this digital world and every day the use, technique and application is improving.

II. DIGITAL WATERMARK: APPLICATION

1. Copyright Protection

Protection of the intellectual property through digital watermark technique is one of the basic motives of watermarking. [3] It is very essential to protect the document and maintain the ownership against illegal use or duplicity. There are cases in which the owner has a license and the product is reproduced and redistributed without the consent of the owner(s) by an authorized person. This is common in video, audio and even in image, text, what we commonly known as piracy. The unauthorized person makes profit by illicit access to the content. In this case digital watermarking comes into play by embedding a mapping relation between the owner and the protected creation. There may also be cases when an end user may claim any protected content as his own. Digital watermark is very effective in these cases as the quality of the duplicate work though lower than the original one, is usually of low or little significance.

2. Tamper proofing

Tamper proofing in its etymology means to hinder, suspend or detect unauthorized access to any protected device or content. In general the most common use is the burglar alarm in which the alarm goes on tampering. In case of digital protection on unauthorized access the watermark must be fragile and should be able to sense any plausible tampering and disable the content i.e. the data and its functionality.

3. Multimedia authentication and broadcasting

With the wide distribution and sharing of multimedia content it very essential to identify the owner and also track its distribution to recognize any illegal use. Every multimedia is embedded with information (owner details, transaction id, serial number etc). This can be perceivable or unperceivable to hums and is used to track the use of the multimedia content and ensure proper broadcasting.

4. Fingerprinting

Fingerprinting algorithm is used to detect illegal duplication and the source of it. This method can be seen as a high performance hash function. A hash index is created with every distribution, re-distribution based on the content. The original hash value is fetched and compared at each level of distribution to find the source of illegal circulation.

5. Application in medical science

It is very essential to restore the content of a medical image [4] and to detect any tampering. Medical diagnosis of disease, treatment and proper steps are taken on the medical image taken with proper equipments. Several times the medical image may need to transfer between health professionals. When the image is taken it is watermark patient details or the authorization group to protect the privacy and detect the tampering. Any tampering with the image can be revealed by its watermark

6. Surveillance camera/ Monitoring system

Sometimes the video or image in a system is a very essential part of evidence. In any protected zone or place of interest we see surveillance camera or CCTV. The video becomes very essential part of evidence and thus there is a need of protection. Every frame is embedded with a watermark, and in case of evidence it can be found out whether part of the video/image has been changed or not. [5]

III. PREVIOUS WORKS

Document involving plain text transferring and sharing over the Internet has become popular and has become essential over the last two decades. Text watermarking is one of the most difficult types of watermarking. There are several algorithms on text digital watermarking and each algorithm is efficient and improved in its own way. The algorithms are based on image, semantic, syntax, structure. Text watermarking incorporating text and image is most common and widely used. The first one to propose such a method using text image was Brassil, et al [6]. Later Maxemchuk, et al [7]-[8] and Low, et al. [9] further analyzed this work and both worked individually on these methods.

Ding Huang and Hong Yan [10] were the first to propose a work based on inter-word space statistics and without changing the content of the text document. The work is efficient and still used today for references. Their work involved changing the space between the words over a number of lines and forming a pattern in the form of a sine wave. The characteristics of their work included that sine wave varies gradually and such variation cannot be noticed by the human eyes. Sine wave's periodic symmetry makes it easy and reliable for decoding. Young-Won Kim, [11] et al proposed their work based on word classification and inter-word space statistics.

Works including text-image based to fully protect the document were proposed [12] and where a signature was embedded to ensure the authenticity [13]. Recently a work has been proposed which not only protects the document but ensure the authenticity by encrypting the text which involves cryptography. [14] Algorithms based on secret message being embedded in the text document also have been put forwarded. [15]

Xingming,et al's works included noun-verb basis for text watermarking [16], which uses nouns and verbs in a sentence parsed with a parser grammar using semantic. There are also work which based on the synonym, the watermark is the certain words which are replaced with their synonyms [17] There are also works based on the occurrence of punctuation marks and double letter words and the frequency of them. Some of them change the content of the text document while others simply work on encrypting. We propose a method that is based on the statistics of the occurrence of words in English text and changing

the attribute of the adjoining words to certain words. Hence we propose the algorithm which ensures integrity and authenticity.

IV. PROPOSED WORK

We propose a fragile digital watermark with a new concept to safeguard the text documents. This algorithm is entirely based on the frequency of occurrence of words in a text document in English. We use the analysis of the Oxford English Corpus which lists the 100 most occurring words used in English language. The list is being given in Table .1 with only the first 20 words and we use the first 10 words skipping the single letter words. The study which listed the words claims that the first 25 words consist of about thirty-three percent of all printed material in English. Our algorithm uses the document content to be protected to generate the watermark.

Table I: Most common words in English by Oxford English Corpus

Rank	Word	Rank	Word	Rank	Word	Rank	Word	Rank	Word
1	the	5	and	9	have	13	not	17	as
2	be	6	a	10	I	14	on	18	you
3	to	7	in	11	it	15	with	19	do
4	of	8	that	12	for	16	he	20	at

We maintain an array which consists of the 10 words and any one of the words is randomly chosen as the keyword K. Next we find the character M at the center of K. Then we determine a value S using the ASCII value of M. Now the document is searched for words matching the keyword K and a note is taken for the preceding and the following word of K. At each occurrence of K in the text, alternatively either the previous or the following word is shifted from its position by S pixels which have been calculated earlier either by performing a left shift or a right shift. In our experiment this value of S is lesser than 1. The words are shifted at a very small pixel and this is almost unperceivable to human eyes. Now the keyword K is sent along with the document in an encrypted form.

The keyword is checked against each word the case being ignored. It is assumed that all the words i.e. that the keyword are in lower case for our experiment. It is because then the ASCII value ranges from 97 to 122. We use the central character of the keyword and using our algorithm we find value of sine 10 to sine 35. This value ranges between 0.34 and 0.94 thus varies gradually and bring unpredictability.

Algorithm:

1. Write 10 most frequent words in an array A
2. Randomly read a keyword K from A
3. Determine M as central character of K
4. Determine S:= sin(ASCII of M – 87)
5. Declare ctr:= 0
6. While 1
7. Read each word T
8. Read the previous word in P and the next word in N
9. Determine if T equals K
10. if ctr is even
11. Then Right shift N by S pixels
12. Else Left shift P by S pixels
13. End if;
14. ctr:= ctr + 1
15. if end of document is reached
16. break
17. end if
18. End while
19. Include the keyword with the document

T = each word of the document
P= previous word corresponding the word matching the keyword K
N=next word corresponding the word matching the keyword K
S=pixel to be shifted
ctr= counter to determine whether left shift or right shift

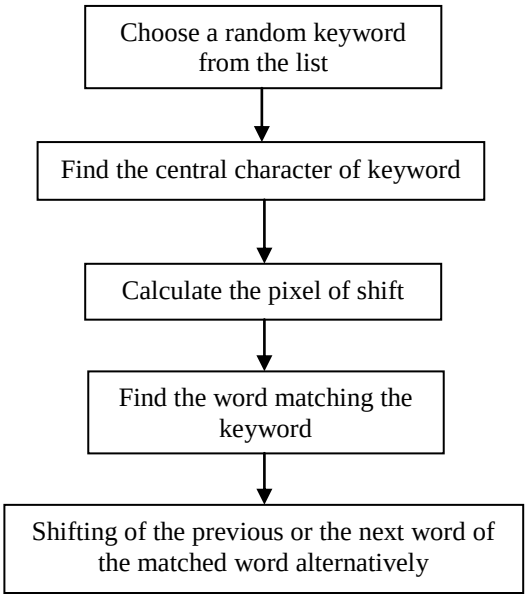

Figure 2: Embedding of digital watermarking

The advantage of this algorithm is manifold. Firstly we are operating on the document only once and the keyword is chosen randomly. This randomness of choosing the word increases protection to the document concerned. Thus if anyone has been tracking some documents of this same algorithm concerned it would be very difficult to determine the keyword used for a document concerned. Moreover choosing the alternative word against the keyword i.e. either preceding or following word along with alternative shifts increases imperceptibility and robustness. If anyone tries to duplicate or change any part/whole of the document the authenticity and integrity of the document can be easily determined. As the keyword is sent in an encrypted form, to determine the ownership or integrity the keyword is decrypted correctly and the same algorithm is used and checked against the document concerned. Any mismatch in the space will prove tampering.

An example showing how our algorithm works:

The random keyword chosen here is: the

A word from the list has been chosen as the keyword over here and this is an example to show how the proposed algorithm works.

Figure 3: The original text

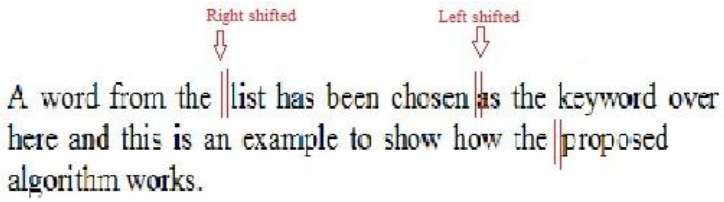

Figure 4: The enlarged image of embedded text

Arrow shows the original position of the word.

V. CONCLUSION

This paper provides an overview on the different features of digital watermarking, the various applications and its use in various fields. Text is the most widely used means of communication and we have proposed an algorithm that ensures the integrity of the document. A new technique is put forward which creates a watermark based on the content of the document and embeds it without changing the content of the document. As the keyword used for the algorithm is randomly chosen from a list of words it improves the robustness of this algorithm. Further the alternative shifting adds to the security and imperceptibility without adding any complexity. To authenticate and prove the integrity of the document, the watermark can be easily extracted and verified for tampering.

REFERENCES

[1] Smitha Rao M.S , Jyothsna A.N, Pinaka Pani.R, "Digital Watermarking: Applications, Techniques and Attacks", International Journal of Computer Applications , Volume 44 No.7, April 2012

[2] B Surekha, Dr GN Swamy, Dr K Srinivasa Rao, "A Multiple Watermarking Technique for Images based on Visual Cryptography", International Journal of Computer Applications Volume 1 No. 11,2010.

[3] Abou Ella HASSANIEN, "A Copyright Protection using Watermarking Algorithm", Informatica, Institute of Mathematics and Informatics, Vilnius, Vol. 17, No. 2, 187–198, 2006

[4] A.Umaamaheshvari, K.Thanuskodi, "Survey of Watermarking Algorithms for Medical Images", International Journal of Engineering Trends and Technology- Volume 3 Issue 3, 2012

[5] Dolley Shukla, Manisha Sharma, "Overview of Scene Change Detection - Application to Watermarking", International Journal of Computer Applications , Volume 47 No.19, June 2012

[6] J. T. Brassil, S. Low, N. F. Maxemchuk, and L. O'Gorman, "Electronic Marking and Identification Techniques to Discourage Document Copying", IEEE Journal on Selected Areas in Communications, vol. 13, no. 8, October 1995, pp. 1495-1504.

[7] N. F. Maxemchuk, "Electronic Document Distribution", AT&T Technical Journal, September 1994, pp. 73-80.

[8] N. F. Maxemchuk and S. Low, "Marking Text Documents," Proceedings of the IEEE International Conference on Image Processing, Washington, DC, Oct. 26-29, 1997, pg. 13-16.

[9] S. H. Low, N. F. Maxemchuk, and A. M. Lapone, "Document Identification for Copyright Protection Using Centroid Detection", IEEE Transactions on Communications, Mar. 1998, vol. 46, no.3, pg 372-381.

[10] D. Huang and H. Yan, "Interword distance changes represented by sine waves for Watermarking text images", IEEE Trans. Circuits and Systems for Video Technology, Vol.11, No.12, pg.1237-1245, Dec 2001.

[11] Young-Won Kim , Kyung-Ae Moon, and Il-Seok Oh, "A Text Watermarking Algorithm based on Word Classification and Inter-word Space Statistics" , Proceedings of the Seventh International Conference on Document Analysis and Recognition, IEEE, 2003

[12] Z. Jalil, A. M. Mirza ,"Text Watermarking Using Combined Image-plus-Text Watermark" , IEEE, 2010

[13] Z. Jalil, A. M. Mirza ,"An Invisible Text Watermarking Algorithm Using Image Watermark", Innovations in Computing Science and Software Engineering, 2010.

[14] Jaseena K.U., Anita John, "Text Watermarking using Combined Image and Text for Authentication and Protection", International Journal of Computer Applications ,Volume 20 No.4, April 2011

[15] Sanjay Jadhav, Viddhulata Mohite, "A Data Hiding Techniques Based on Length of English Text Using DES and Attacks", International Journal of Research in Computer Science , Volume 2 Issue 4 ,2012, pg. 23-29

[16] U. Topkara, M. Topkara, M. J. Atallah, "The Hiding Virtues of Ambiguity: Quantifiably Resilient Watermarking of Natural Language Text through Synonym Substitutions", Proceedings of ACM Multimedia and Security Conference, Geneva, 2006

[17] Xingming Sun, Alex Jessey Asiimwe, "Noun-Verb Based Technique of Text Watermarking Using Recursive Decent Semantic Net Parsers", Proceedings of ICNC (3), Springer Press, August 2005, Vol 3612, pg: 958-961

AUTHORS

First Author – Nikita Pandey, B.Tech, Narula Institute of Technology, Email id: nikita.pandey29@gmail.com
Second Author – Sayani Nandy, B.Tech, Narula Institute of Technology, Email id: sayani.nandy00@gmail.com
Third Author – Shelly Sinha Choudhury, Professor, Jadavpur University, Email id: shelism@rediffmail.com