International Journal of Advanced Research Foundation
Website: www.ijarf.com, Volume 2, Issue 6, June 2015)

A Novel Approach of Zero Watermarking for Text Documents

Pankaj Bhambri
Department of information Technology,
Guru Nanak Dev Engineering College,
Ludhiana, Punjab, India
pkbhambri@gmail.com

Pradeep Kaur
Research Scholar,
Punjab Technical University,
Jalandhar, Punjab, India

***Abstract*— With widespread use of Internet and other communication technologies, it has become extremely easy to reproduce, communicate, and distribute digital contents. As a result, authentication and copyright protection issues have arisen. Text is the most extensively used medium travelling over the Internet besides image, audio, and video. The major part of books, newspapers, web pages, advertisement, research papers, legal documents, letters, novels, poetry, and many other documents is simply the plain text. Copyright protection of plain text is a significant issue which cannot be condoned. In this thesis, we have proposed a zero-watermarking approach towards text watermarking. We propose a zero text watermarking algorithm based on occurrence frequency of vowel ASCII characters and words for copyright protection of plain text. The embedding algorithm makes use of frequency vowel ASCII characters and words to generate a specialized author key. The extraction algorithm uses this key to extract watermark, hence identify the original copyright owner. Experimental results illustrate the effectiveness of the proposed algorithm on text encountering meaning preserving attacks performed by five independent attackers and the results are also compared with the recent work on text watermarking.**

***Index Terms*— Copyright protection, Digital watermarking, Document authentication, Watermark embedding and extraction**

I. INTRODUCTION *(HEADING 1)*

Unlike analog media that are becoming obsolete by now, digital media can be accessed, stored, copied, and distributed more easily and in no time. Advancement in digital media and technologies have brought unlimited benefits to mankind, but they also create problems for parties wishing to prevent unauthorized copying and distribution of valuable digital contents such as copyrighted, commercial, secret, and sensitive data. Security of digital contents has gained tremendous importance in current digital era. Internet has become an essential part of our daily life for the transfer of different forms of data such as emails, news papers, articles, websites, images, audios, videos, commercials, and opinion blogs. Most of the information over the Internet is in the form of text and the copyright protection of text is one of the major concerns of its creator\author. Text is the most essential and dominant part of legal documents, reports, and journals; but its protection has been seriously ignored. The threats of electronic publishing like illegal copying and re-distribution of copyrighted material, plagiarism and other forms of copyright violations need to be explicitly addressed, particularly for plain text.

II. DIGITAL WATERMARKING

A digital watermark is a piece of information which is embedded in the digital media and hidden in the digital content in such a way that it is inseparable from its data. This piece of information known as watermark, a tag, or label into multimedia object such that the watermark can be detected or extracted later to make an assertion about the object. The object may be an image, audio, video, or text. Watermarking is the process of inserting a digital signal or pattern (indicative of the owner of the content) into digital content. The signal, known as a watermark, can be used later to identify the owner of the work, to authenticate the content, and to trace illegal copies of the work.

There are two types of digital watermarking: visible (perceptible) and invisible (imperceptible) In visible watermarking, watermarks are embedded in such a way that they are visible when the content is viewed. Invisible watermarks cannot be seen but recovering of watermark is possible with an appropriate decoding algorithm. Invisible watermarks are more robust than visible watermarking. Watermarking can again be robust or fragile. Robust watermarking is a technique in which modification to the watermarked content will not affect the watermark in any way. But in the case of fragile watermarking, watermark gets destroyed when watermarked content is modified or tampered with.

Watermarking can also be classified based on the type of document to be watermarked. The classifications are: Image Watermarking, Video Watermarking, Audio Watermarking and Text Watermarking. Text watermarking solutions are not robust against random tampering attacks such as insertion, deletion attacks. In this paper, we propose a zero text watermarking algorithm which is resistant towards random tampering attacks. The important issues that arise in the study of digital watermarking techniques are capacity, robustness, transparency and security. Cryptography only provides security by encryption and decryption. However,encryption cannot protect the content after decryption. Unlike

International Journal of Advanced Research Foundation
Website: www.ijarf.com, Volume 2, Issue 6, June 2015)

cryptography, watermarks can protect content even after they are decoded. Also cryptography cannot prevent illegal replication of the digital content.It is only about protecting the content of the messages. But watermarks not only protect the content but also provide many other applications like copyright protection, copy protection, ID card security etc.

Text watermarking is an emerging area of research. Text watermarking algorithms developed so far can be classified in to Image based methods, Syntactic methods, Semantic methods and Structural methods Categories.

In image-based methods of text watermarking, the binary watermarks are embedded in text image. In syntactic methods, the syntactic structure of a text is utilized to embed the watermark. In semantic schemes, the watermark embedding is done by utilizing the semantics of text. Many algorithms are proposed based on these three schemes. Structural schemes of text watermarking are the recently used watermarking approach which uses text structures to embed watermarks. In this scheme, text is not modified when the watermark is embedded in to it. The structural algorithm is not applicable to specialized text such as poetry, legal documents, Web contents, and documents containing mathematical notations with floating point numbers. These types of text watermarking schemes are robust zero watermarking. In this paper we proposed a system in which the image is being secured by watermarking so as to protect the data more easily This paper is includes as follows: Section 2 gives an overview about the proposed work on text watermarking. The propose algorithm for embedding and extraction are discussed in detail .Section 3 gives the details of earlier work done on text watermarking.

III. PROPOSED ALGORITHM

Text watermarking techniques help to protect the text from illegal copying, forgery, and redistribution. It also helps to prevent copyright violations. Besides, watermarking provides authentication and protection of text documents. Text documents face many threats such as copying, tampering, plagiarism, reproduction, and paraphrasing attacks. The best solution to address these problems is digital watermarking, which not only helps in authentication of the digital material but also in its protection. Digital watermarking can be used to identify the owner of the copyright material which may be in the form of audio, video, image, a plain text. The original copyright owner of text inputs his/her watermark. The algorithm generates a unique key that corresponds to input data. This key is used later for extraction of watermark, whenever a copyright conflict arises in future. In the proposed approach,We propose a zero-watermarking approach in which the host text document is not altered to embed watermark, rather the constituents of text are utilized to generate a key in a unique way to protect it. We have used the essential text constituents; vowel and articles in our algorithm.

The main contributions of this work are:

- A novel zero-watermarking approach has been adopted towards text watermarking.
- Watermarking is blended in a unique way which provides a robust text watermarking solution.
- The resilience against attacks on text has been improved and the watermark has been made robust to attacks of varying length and nature.
- Proposed techniques employ a novel approach for text watermarking in which the text is not modified to embed a watermark. The text is declared as a sensitive medium and its protection technique is proposed without changing it.
- The proposed technique provides optimal results using vowel and articles in a unique way.

The watermarking process involves two stages, watermark embedding and watermark extraction. Watermark embedding is done by the original author and extraction done later by the CA to prove ownership and compare the result with the previous algorithm [1]. Embedding Algorithm incorporates the watermark in text. The embedding algorithm logically embeds the watermark in text and generates the author key. Extraction Algorithm is used to extract watermark from the text. It takes the author key obtained from the CA (generated by the embedding algorithm) as input and extracts the watermark from the text. The algorithm is kept with the CA that is used to resolve copyright issues.

IV. PREVIOUS WORK

Text watermarking is an emerging domain for research. A robust and practical solution may open new horizons to the information security world. Many watermarking techniques have been developed since 1993, which includes text watermarking that uses text image, synonyms based, noun verb word and sentence structure based, acronyms based schemes and many others.

- In Image based approach towards digital text watermarking, text document image is used to embed the watermark. Text is difficult to watermark because of its simplicity, sensitiveness, and low capacity for watermark embedding. The initially attempts in text watermarking tried to treat text as image. Watermark was embedded in the layout and appearance of the text image. Brassil, et al. proposed a few methods to watermark text document by using text image. The first method proposed by Brassil was the line-shift coding algorithm which alters the document image by moving lines upward or downward (left or right) depending on binary signal (watermark) to be inserted. The detection algorithm is non-blind in which the original document should be available. The second method was the word-shift coding algorithm which moves the words within text horizontally thus expanding spaces to embed the watermark. The algorithm can operate both in non-blind and blind modes. The third method is the feature coding algorithm which slightly modifies features such as the pixel of characters, the length of the end lines in characters to encode watermark bits in the text.

- In Syntactic Approach, Text is made up of characters, words, and sentences. Sentences have different syntactic structures. Applying syntactic transformations on text structure to embed watermark has also been one of the approaches towards text watermarking in the past. Mikhail. J. Atallah, et al. first proposed the natural language watermarking scheme using the syntactic structure of text where the syntactic tree is built and transformations are applied to it to embed the watermark preserving all inherent properties of the text. They developed techniques for embedding a robust watermark in text by a number of information assurance and security techniques with the advanced and resources of natural language processing.
- In Semantic Watermarking schemes focus on using the semantic structure of text to embed the watermark. Text contents, verbs, nouns, words and their spellings, acronyms, sentence structure, grammar rules, etc. have been exploited to insert watermark in the text but none of these proved to be resilient and degrade the quality of the text to a large extent.

Atallah et al. were the first to propose the semantic watermarking schemes in 2000. Later, the synonym substitution method was proposed in which watermark is embedded by replacing certain words with their synonyms. Xingming, et al. proposed noun-verb based technique for text watermarking which exploits nouns and verbs in a sentence parsed with a grammar parser using semantic networks. Jalil Z., Farooq M., Zafar H., Sabir M., and Ashraf E.: they proposed algorithm which uses the non vowel characters to watermark the text document. In the algorithm the text document is first analyzed and prepositions are identified and the highest occurring non-vowel ASCII characters are identified and populate list. This list is then used to generate an author key based on watermark provided by the owner. In order to examine the performance of the proposed algorithm, take a sample to perform different types of attacks by five different individuals. The sample text was given to five different individuals to make meaning preserving intelligent attacks. The individuals were selected randomly with different English proficiency and educational background. Five attacked samples were obtained. Attack on the original file can only alter the characteristics of the original file but the whole theme of text remains same. When the attacker intends to violate the copyrights, he/she will perform intelligent attacks in order to retype or reproduce the text and all the attacked samples were varied based on attack volume. Furthermore, in this algorithm we used two two watermark of different length, in order to check the alteration occurred by characteristic of the watermark. Zunera Jalil and Anwar M. Mirza The proposed algorithm embeds the watermark image logically in text, which means watermark is not actually embedded in; rather it is used to generate a key based on text constituents. We have used the essential text constituents; double letters and frequently used words in English text in our algorithm. Double letters are very common in the English language. Almost 7 to 11% words of any text contain double letters. Besides, there are some frequently used words in English language which are utilized in the proposed text watermarking algorithm. The algorithm with a blind zero-watermarking approach gives a robust solution for the text watermarking problem. The image watermark can be extracted later for copyright protection and text authentication.

Jaseena K.U, Anita John: In this paper a new text watermarking algorithm using combined image and text watermark to fully protect the text document is proposed. In this algorithm, the occurrences of double letters existing in text are used to embed the watermark. The original copyright owner of text embeds the watermark in a text and generates an author key using an embedding algorithm. The author key along with the watermark is kept with the Certification Authority (CA), where the original author is registered. Later the watermark is extracted from the text using the watermark key to identify original owner. And also proposed an algorithm in which a text watermarking algorithm based on the occurrence of non-vowel ASCII characters for protection of the text document is proposed. In this algorithm, the occurrence of all non-vowel ASCII characters is analyzed in each partition and maximum occurring non-vowel ASCII character is identified to form MONV (Maximum Occurring Non-Vowel) list. The author key is generated using this MONV list and user given watermark. The original author then registers this author key with a certification authority (CA), a trusted third party. The watermark and this author key are kept with the CA along with time and date. This key is used in the extraction algorithm to identify the original copyright owner. According to this algorithm, the length of generated watermark key is high. Lengthy watermark key has at the same time advantages and disadvantages. The advantage is that since the key length is high, it will be difficult for an attacker to guess the key easily. Thus chance for brute force attack will be reduced. However, the disadvantage is that it will be difficult for CA to maintain key and also transfer of key between owner of text and CA will not be easy.

Zunera Jalil and Anwar M. Mirza, A Zero Text Watermarking Algorithm Using Hybrid Watermarks In this proposed algorithm ,watermarks composed of both image and text, make the text secure and has better robustness, they developed a text watermarking algorithm, which uses combined image-plus-text watermark to watermark the text document. Watermark can later be separately identified to prove the ownership. evaluated the performance of the algorithm for localized and dispersed random tampering attack .The algorithm using text plus image watermarks are more robust, secure and efficient against random tampering attacks.

REFERENCES

[1] Zunera Jalil, Anwar M. Mirza, and Maria Sabir, “Content based Zero-Watermarking Algorithm for Authentication of Text Documents”, International Journal of Computer Science and Information Security, Vol. 7, No. 2, February, 2010.

[2] Zunera Jalil and Anwar M. Mirza, "A Novel Text Watermarking Algorithm based on Double Letters", International Journal of Computer Mathematics, (Indexed by ISI Impact Factor 0.48), *(Under second review).*

[3] Jaseena K.U, Anita John "An Invisible Zero Watermarking Algorithm using Combined Image and Text for Protecting Text Documents", International Journal on Computer Science and Engineering (IJCSE).

[4] Zunera Jalil and Anwar M. Mirza, "A Zero Text Watermarking Algorithm Using Hybrid Watermarks", Journal of Multimedia (JMM), *(Under second review).*

[5] Z. Jalil, M. Farooq, M. Arif and A. M. Mirza, "A Zero Text Watermarking Algorithm using Non-Vowel Alphabets", International Journal of Electrical, Computer, and Systems Engineering (ICCESSE 2010), November 24-26, 2010, Venice, Italy.

[6] Z. Jalil, A.M. Mirza ,"A Review of Digital Watermarking Techniques for Text Documents" IEEE,,2009.

[7] Z. Jalil, A. M. Mirza ,M. Sabir "Content Based Zero Watermarking Algorithm for Authentication of Text Documents" ,International Journal of Computer Science and Information Technology,V 7, 2010.

[8] Z. Jalil, A. M. Mirza, and T. Iqbal, "A Zero-Watermarking Algorithm for Text Documents using Structural Components",International Conference on Information and Emerging Technologies (ICIET 2010), June 2010.