An Existential Review on Text Watermarking Techniques

Manmeet Kaur
Rayat and Bahra Institute of Engineering & Bio-Technology, Kharar, Punjab

Kamna Mahajan
Assistant Professor
Rayat and Bahra Institute of Engineering & Bio-Technology, Kharar, Punjab

ABSTRACT
Nowadays, with the extensive use of internet all over the world, most of the government, public and private data are increasingly being published on internet. The protection of these online documents is the need of the hour and need to be dealt with urgently. Text watermarking is the technique which helps to protect the authenticity and integrity of text documents by inserting watermarks in the text. Text watermarking is an active area of research from several years. This paper presents a review of various text watermarking techniques described in literature. We also highlight the security issues like copyright protection, tamper detection and data hiding which need to be focussed for ensuring text security.

Keywords
Text watermarking, Copyright Protection, Security, Encryption, Tampering.

1. INTRODUCTION
The advent of the internet has resulted in many new opportunities for creation and delivery of content in digital form. Digital contents mostly comprises of text, image, audio, and video. Besides, making it easier to access information within a very short span of time, it has become difficult to protect copyright of digital contents and to prove the authenticity of the obtained information. Now a day's, text is the most important medium travelling over the internet in addition to image, audio and video. The major content of websites, newspapers, e-books, research papers, legal documents, letters, SMS messages, etc is the text. These text documents sometimes have various threats like illegal copying of important information, redistribution of copyrighted text documents, tampering, forgery, theft and authentication. To overcome these threats various solutions such as authenticity, integrity, confidentiality, and copyright protection are urgently required. One solution to solve all these problems is digital watermarking [1].

Digital watermarking is the process of embedding a unique digital watermark in a digital content to protect it from illegal copying and copyright violations. The process of embedding and extracting a digital watermark to and from a digital text document which uniquely identifies the original copyright owner of that text is called Digital Text Watermarking. There is no perceptible difference between the watermarked and original signal, and the watermark is difficult to remove or alter without damaging the host signal. An ideal text watermarking solution should be one that can be easily implemented, robust and imperceptible. It should also be adaptable to different text formats and should have high information carrying capacity. It should be effectively applied to print/digital proof[5].

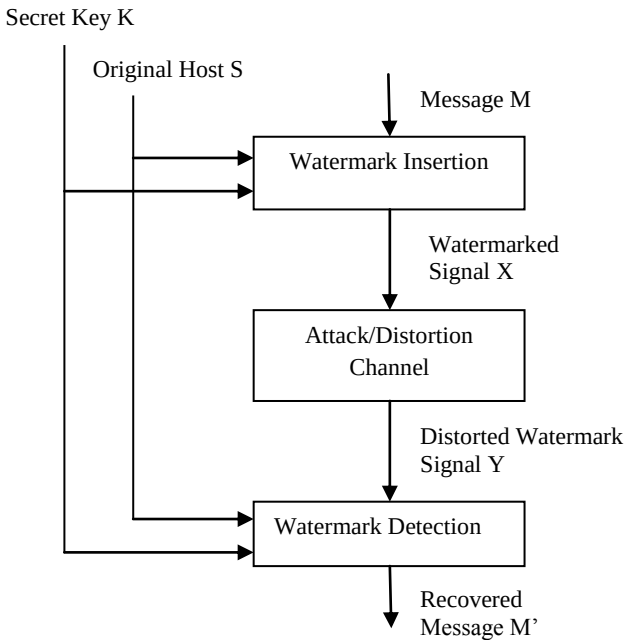

Fig 1: Block diagram of a watermarking system

The digital watermark can be generated in many ways. Following are the categories which are defined and derived by the researchers.

- **Visible and Invisible**: In this category watermark message is hidden in digital text document then it is called invisible watermarking. And in case of visible watermark message is noticeable to spectator [2].
- **Readable and Detectable**: In detectable category determining whether the digital text document carries watermark or not. And in readable category a digital text document contain watermark which is readable to the user of text document.
- **Robust, fragile, semi fragile**: The digital watermarking for copyright protection is called robust watermarking and watermark message is detectable but not erased. In fragile category the watermark message is detectable, may alter, may erase and Is used for integrity authentication. And the digital watermarking for content authentication is called semi-fragile watermarking.[3]
- **Blind, Non Blind, Zero Watermarking**: In blind watermark category while extracting watermark there is no need of watermarked text document. In non blind watermark category while extracting

watermark there is necessary of watermarked text document. And in zero watermarking text is not modified to embed watermark, rather the characteristics of text are used to generate watermark key. [4].

- **Simple and multiple**: In simple category watermark is inserted in digital text document only once. And in Multiple category watermark is inserted multiple times but without affecting precedent watermark message or signal [1].

This paper focuses on various techniques used for text watermarking. We also review the work of various researchers available in literature. This paper has been divided into various sections. Section2 enlightens about various domains of text watermarking techniques. Section3 reviews the literature work. In section 4 we conclude the paper.

2. TECHNIQUES FOR TEXT WATERMARKING

A text, being the simplest mode of communication and information exchange, brings various challenges when it comes to copyright protection. Any transformation on text should preserve the meaning, fluency, grammaticality, writing style and value of the text. Short documents have low capacity for watermark embedding and are relatively difficult to protect. Text watermarking algorithms are also dependent on text size, its language, rules, grammar, conventions and writing styles. [6]

Various text watermaking techniques have been proposed, which include: image based, syntactic structure based, semantic based, natural language processing based, noun-verb based, word and sentence based, zero watermarking algorithm, etc.

The description of the work done in each category is as follows:

Text Image Based Watermarking:

In this 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.

The present text image watermarking algorithms proposed by researchers chiefly rely on line-shifting and word-shifting techniques especially on imperceptible modification of spacing of words, spacing of letters, shifting of baselines, modifying the serifs, kerns etc.,[5] The fist method is the line-shifting 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 is the word-shift coding algorithm, in which the words within text are moved horizontally thus expanding spaces to embed the watermark. This algorithm can operate both in non-blind and blind modes. The third method is the feature coding algorithm in which certain text features are altered to encode watermark bits in the text.

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. In this, first a syntactic tree is built and transformations are applied to it to embed the watermark preserving all inherent properties of the text. The natural language processing algorithms are used to analyze the syntax and semantic structure of the text, while making changes to incorporate the watermark bits [4].

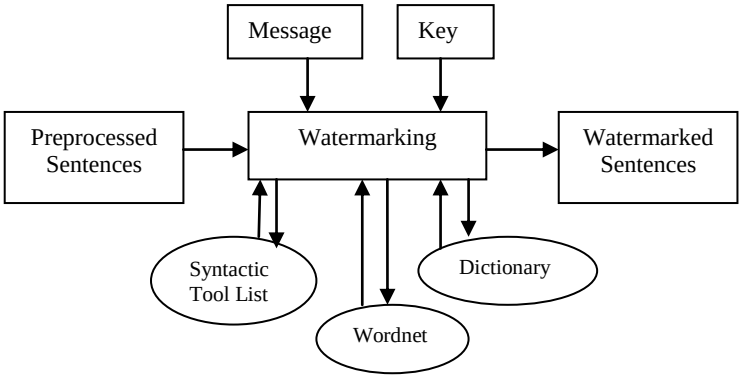

Fig 2: Syntactic sentence level watermarking [6]

Semantic Approach:

The semantic watermarking schemes focus on using the semantic structure of text to embed the watermark. Text contents, like verbs, nouns, prepositions, words spelling, acronyms, sentence structure, grammar rules, etc. are exploited to insert watermark in the text. They are language dependent and have limited applicability.

Figure 3 shows the parse tree for noun-verb based transformation. In noun-verb based technique for text watermarking ,the nouns and verbs in a sentence parsed with a grammar parser using semantic networks.

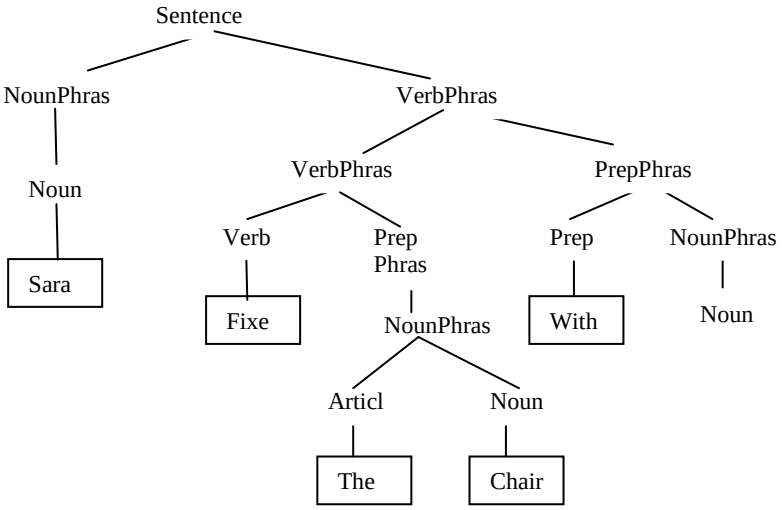

Fig 3: Noun verb based tree of sentence "Sarah fixed the chair with glue" [6]

The semantic techniques for digital watermarking use natural language processing algorithms to analyze text meaning and to perform transformation. The synonym based techniques are not resilient to the random synonym substitution attacks. Also, synonyms may not always be giving the exact meaning of the word hence destroying the value of the text. The sensitive nature of some documents e.g. legal documents, poetry, and quotes do not allow us to make semantic transformations randomly because in these forms of the text, a simple

transformation sometimes destroys both the semantic connotation and the value of text.[6]

Structural Approach:

Structural approach is the most recently used approach for watermarking text documents. In this the structure of the text is used to watermark the text. A text watermarking algorithm for copyright protection of text using occurrences of double letters (aa-zz) in text to embed the watermark has recently been proposed [7]. This algorithm is a combination of encryption, steganography and watermarking, and provides a robust solution for text watermarking problem.[7]

The structural approach is not applicable for all types of text documents because it uses an alphabetical watermark; hence another method is used that uses an image watermark. The image can be a logo, fingerprint or digital signature of the original author.

3. LITERATURE REVIEW

Mali et al. [1] proposed a novel watermarking technique for English language text documents, which was based on grammatical and suitable encryption methods. The grammatical rules like conjuctions, pronouns and modal verbs were used to generate encrypted watermark message.

A novel semi-fragile text watermarking scheme for content authentication of Chinese text documents was proposed by **Xinmin Zhou et.al [3]**. In this, the host document was divided into two blocks by using Chinese mathematical expressions and two watermarks were generated by using the chaotic encrypting algorithm to encrypt the hashing value of the text document with two different keys respectively, and they were embedded into the two text blocks. These two different keys were generated by computing the frequency and stroke numbers of Chinese characters of two different text blocks. When the content of the watermarked text document was modified, the extracted results of two text blocks did not match, and the authentication result of the text document would turn out to be *false*.

Text watermarking algorithm that could be used to protect all open textual digital contents from forgery was developed by **Zunera Jalil et al. [4]**. The algorithm embeds the watermark image logically in text and that image was later extracted and used to prove ownership. The performance of the algorithm was tested for random tampering attack in dispersed form on 12 text samples.

A new method was proposed by **Suganya Ranganathan et al.[5]**, that combined the best of both image based text watermarking techniques and language based watermarking techniques for an efficient copyright protection mechanism.

In another paper **Zunera Jalil and Anwar M. Mirza [6]** presented a review of some of recent research in watermarking techniques for plain text documents. The reviewed approaches were classified into three categories, the image based approach, the syntactic approach and the semantic approach. The paper also discussed the main contributions, advantages and drawbacks of different methods used for text watermarking in the past.

A natural language watermarking for Korean based on morphological division and syntactic displacement was presented by **Mi-Young Kim et al. [9]**. By using the characteristics that a word in the agglutinative language usually consists of several morphemes, they divided the word that had two content morphemes into two new words, and a new function morpheme was inserted for the first new word that did not have a function morpheme.

Jaseena K.U. and Anita John [10] suggested a new text watermarking technique that used combined image and text watermark and encryption to protect the text document. The watermark was logically embedded to the text and the text was encrypted. Later the text was decrypted and then the watermark was extracted to prove authenticity. Experimental results demonstrated the effectiveness of the proposed algorithm.

A novel text watermarking scheme was given by **Yingli Zhang and Huaiqing Qin [11]** which provided good robustness for word document. The scheme embeded the secret signals in the special properties of word object. The watermarking information was encrypted and divided into several groups. Then it was packed into the message before embedding in the word document circularly. All these operations made the scheme performance excellent on robustness when encountered with attacks as compared to the methods based on the features of character font.

In another literature work security issues of zero text watermarking were discussed. To overcome this issue a new technique for plain text was given by **Zhangjie Fu et al.[12].** In this technique a zero knowledge-based watermark detection scheme was proposed, in which homomorphic properties

of asymmetric encryption algorithm in the multiplication operation was used to prevent the owner from cheating by ambiguity attacks. The security of this scheme was improved by using the improved feature extraction algorithm.

Min Du and Quanyou Zhao[13] proposed a text digital watermarking algorithm based on human visual redundancy. According to this paper, the human eye is not sensitive to the slight change for text color, watermarks were embedded by changing the low-4 bits of RGB color components of characters. Experiment showed that the proposed method has good invisibility and robustness to resist deletion, modification attack etc. This algorithm can be used for secret communication of confidential information.

A public zero knowledge watermark detection protocol was proposed by **Zhangjie Fu et al. [14]** to prevent the owner from cheating by ambiguity attacks. Four steps were concluded in this proposed protocol to achieve the goal: robust feature extraction method, watermarks generation method, watermarking embedding method and zero knowledge watermark detection protocol. Any verifier can check that whether the medium contains a watermark claimed by prover or not. The proposed method can satisfy the three requirements of zero knowledge proof of identity: completeness, soundness, zero knowledge. The method ensures the security for watermark verification in the watermarking detection process without revealing any secret information related to watermarking.

For integrity and confidentiality of the document, a technique was proposed by **Leena Goyal et al. [15].** In this technique,watermark was created based on the contents of the document and embeded without changing the contents of the document and also encrypted the text to provide confidentiality. To authenticate and prove the integrity of the document the watermark could be easily extracted and verified for tampering.

Based on Pseudo-Random Number Generator (PRNG) for Cryptography application, a text based watermarking algorithm was proposed by **Chee Hon Lew and Chaw Seng Woo[16]**. RSA key was generated and made public in text watermarking. This technique proved to be robust to resist

deletion, modification attack and can be used in information hiding area using cloud computing.

New adaptive approach based on zero-watermarking for highly-sensitive documents in order to achieve content-originality verification and authentication without physically modifying the cover-text in anyway was presented by **Omar Tayan et al.[17].** The algorithm worked by embedding the watermark logo of the original publisher in an identical-duplicate of the cover-document and a characteristic key was generated. The algorithm can be used to protect all digital textual-content from forgery and illegal content manipulation.

Sukhpreet kaur and Geetanjali babbar[18] presented a zero watermarking technique that used multiple occurrences of letters in a word for generation of watermark.

A system that securely transfers the text messages by hiding them in a digital image file by using the local characteristics within an image was proposed by **Sonia Bajaj and Manshi Shukla[19].** In this paper, Image Steganography with watermarking using LSB and interpolation Techniques were used**.** The proposed system not only hides large volume of data within an image but it also limits the perceivable distortion that might occur in an image while processing it.

Blind watermarking technique was given by **Priyanka Verma et al.[20]** in which watermark was embedded in text document using rotation of letters based on EBCDIC code lookup table. The scheme proved to be better in terms of imperceptibility, robustness and fidelity.

4. CONCLUSION AND FUTURE SCOPE

This paper reflects the work of various researchers on text watermarking. It has been seen that work done so far on text watermarking is limited and specific. Therefore, research work has to be focussed on ensuring robustness, integrity and accuracy of text documents. Hence we conclude that new watermarking algorithms, which are computationally inexpensive and robust, need to be developed to ensure online safety of text documents. In future more work is required to be done in the area of security of text watermaking. More secure methods need to be developed to ensure security of text documents over internet.

5. REFERENCES

[1] Makarand L. Mali, Nitin N. Patil and J.B.Patil .2013.Implementation of Text Watermarking Technique Using Natural Language Watermarks.International Conference on Communication Systems and Network Technologies.

[2] Ingemar J. Cox and Matt L. Miller 2002.The First 50 Years of Electronic Watermarking.EURASIP Journal on Apllied Signal Processing. Issue 2, pp. 126-132.

[3] Xinmin Zhou, Weidong Zhao, Sichun Wang and Rui Peng, 2009.A Semi-FragileWatermarking Scheme for Content Authentication of Chinese Text Documents.IEEE.

[4] Zunera Jalil, Anwar M. Mirza and Tahir Iqbal,.2010.A Zero-Watermarking Algorithm for Text Documents based on Structural Components.IEEE.

[5] Suganya Ranganathan , Ahamed Johnsha Ali, Kathirvel.K & Mohan Kumar.M, 2010.Combined Text Watermarking. International Journal of Computer Science and Information Technologies. Volume 1, No. 5.

[6] Zunera Jalil and Anwar M. Mirza.2009.A Review of Digital Watermarking Techniques for Text Documents. International Conference on Information and Multimedia Technology.

[7] Zunera Jalil, M. Arfan Jaffar and Anwar M. Mirza.2011.A Novel Text Watermarking Algorithm Using Image Watermark. International Journal of Innovative Computing, Information and Control .Volume 7, No. 3.

[8] Patil Bharati Devidas and Patil Nitin Namdeo.2012.Text Watermarking algorithm using structural approach.IEEE.

[9] Mi-Young Kim, Osmar R. Zaiane, and Randy Goebel .Natural Language Watermarking Based on Syntactic Displacement and Morphological Division.

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

[11] Yingli Zhang and Huaiqing Qin.2010.A Novel Robust Text Watermarking For Word Document. IEEE.

[12] Zhangjie Fu, Xingming Sun, Jiangang Shu and Lu Zhou,.2014.Plain Text Zero Knowledge Watermarking Detection Based on Asymmetric Encryption. Advanced Science and Technology Letters.Vol.48.

[13] Min Du and Quanyou Zhao.2011.Text Watermarking Algorithm based on Human Visual Redundancy. Advanced in Information Sciences and Service Science. Volume 3, Number 5.

[14] Zhangjie Fu, Xingming Sun, Jiangang Shu, Lu Zhou and Jin Wang 2014.Verifiable Text Watermarking Detection to Improve Security. International Journal of Security and Its Applications .Vol.8, No.5.

[15] Leena Goyal, Manoj raman, Prateek Divan and Mukaesh Kumar Vijay2014. A Robust Method for Integrity Protection Of Digital Data in Text Document Watermarking. International Journal for Innovative Research in Science & Technology. Volume 1, Issue 6.

[16] Chee Hon Lew and Chaw Seng Woo.2013.Design and Implementation of Text based Watermarking combined with Pseudo-Random Number Generator(PRNG) for Cryptography Application. Latest Trends in Applied Computational Science.

[17] Omar Tayan, Yasser M. Alginahi and Muhammed N. Kabir.2013.An Adaptive Zero-Watermarking Approach for Authentication and Protection of Sensitive Text Documents. International conference on advances in computer and information technology – ACIT.Vol.1.

[18] Sukhpreet Kaur and Geetanjali Babbar.2013.A Zero-Watermarking algorithm on multiple occurrences of letters for text tampering detection. International Journal on Computer Science and Engineering.Vol. 5, No.05.

[19] Sonia Bajaj and Manshi Shukla.2014.Performance Evaluation of an approach for Secret data transfer using interpolation and LSB substitution with Watermarking. International Journal of Computer Science and Information Technologies. Vol. 5 (5).

[20] Priyanka Verma, Rakhshan Anjum Shaikh and Ketki Deshmukh. 2013. A Novel Approach to Angle based Invisible