ISSN:0975-9646

Combined Text Watermarking

Suganya Ranganathan[1] , Ahamed Johnsha Ali[2], Kathirvel.K[3] & Mohan Kumar.M[4]

Department of Computer Applications, Bharathiar University
Nehru College of Management, Coimbatore, Tamilnadu, India

Department of Computer Applications, Karpagam University
Eachanari, Coimbatore, Tamilnadu, India

***Abstract*— There is a vast amount of digital text data exchanged over the internet for the past few years. Such exchanges necessitate a very robust copyright protection mechanism. This paper presents a new idea on how one can effectively combine the advantages of image based text watermarking technique with syntactic water marking technology so as to form a new technique which synthesizes the benefits of both for a robust copyright protection system.**

***Keywords*— Copyright Protection, Data Hiding, Lexical Watermarking, Imaged Text Watermarking, Steganography, Watermarking.**

I. INTRODUCTION

Digital watermarking is a method of hiding copy right protection data in a digital medium whether it is audio, video or text. There are two types of watermarking namely Visible and Invisible watermarking. Visible watermarking is usually in the form of logos, pictures or other text matter which identifies the ownership of the media. Invisible watermarking, on the other hand, is in the form of embedding data that is imperceptible. This type of watermarking is used as a deterrent against copyright violators.

Digital media usually consist of images, audio, video and text. Each one of these digital types requires a suitable individualistic method for watermarking. Of these, image, audio and video watermarking techniques have been extensively researched into and there have been a number of different algorithms and software applications developed, that deal with this kind of text watermarking as per the survey by Young-Won Kim and Ii-Seok Oh because of its inherent qualities.

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.

This paper is organized as follows: Section 2 discusses the present watermarking techniques, its merits and demerits. In Section 3 and 4 proposed approach and discussion will be made. Section 5 marks the conclusion portion of this paper.

II. EXISTING TECHNIQUES

Watermarking of text comes under 2 domains namely 1) Text Image Watermarking and 2) Natural Language Watermarking. The aims of both these watermarking systems are the embedding of information by modifying original data in a discrete manner, such that the modifications are imperceptible and the embedded information is robust against possible attacks. In image text watermarking this goal is achieved by exploiting the redundancy in images and the limitations of the human visual system. Similar approaches are used in other signal-based watermarking domains, such as video and audio. On the other hand, language has a discrete and syntactical nature that makes such techniques more difficult to apply. Specifically, language, and consequently its text representation have two important properties that differ from image representations.

Sentences have a combinatorial syntax and semantics. That is, structurally complex (molecular) representations are systematically constructed using structurally simple (atomic) constituents, and the semantic content of a sentence is a function of the semantic content of its atomic constituents together with its syntactic/formal structure of representations defined by this combinatorial syntax.

III. TEXT IMAGE WATERMARKING

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., The line shift (word shift) algorithm moves a line (word) upward or downward (left or right) depending on binary signal to be inserted. The detection algorithm requires the control lines or words to identify the direction of movement, and the algorithms are non-blind since the original document should be available. Huang et al developed a word shift algorithm that modifies the inter-word spaces that represent a sine wave. The signals are encoded in the phase, amplitude and frequency of the sine waves. The algorithm can operate both in non-blind modes. For signal insertion, spaces between characters should also be adjusted.

The feature and pixel-level algorithms were also found. The text documents can be watermarked by modifying the stroke features such as width or serif. The detection requires an accurate extraction of character strokes. Algorithms working on grayscale images were also developed. Their applications are limited to the scanned gray scale document images.

IV. NATURAL LANGUAGE WATERMARKING

Compared to image text watermarking, natural language watermarking is a relatively new area. In addition to content protection, robust NL watermarking algorithms will enable a wide range of application such as text auditing, mete-data binding, tamper-proofing and traitor tracing.

V. DEFECTS

Watermarking inserted by most of the image text based systems is not robust against attacks such as scanning the document and performing optical character recognition or re-formatting of the document file. For example, when a Microsoft word document, that was embedded with copyright material by way of modification of formatting such as spacing adjustment between words, font size modifications etc., is converted into plain text document by copying the text matter and pasting it into a new ASCII editor such as notepad, all the format will be lost and that the chance of retrieving of the copyright information will also be lost.

Images in general do not lend themselves to a syntactical decomposition similar to the one for language. The atomic/syntactical nature of language brings about unique challenges for NL watermarking. For example, Least Significant Bit (LSB) embedding used in image watermarking that modifies text locally, i.e., based on words, without making perceptually significant changes to sentence structure is a hard problem. This is due to the fact that even small local changes in a sentence can change its semantics and/or make it ungrammatical.

VI. PROPOSED TECHNIQUES

Watermarking techniques that modify perceptually significant portions of an image are more robust against attacks than techniques that modify only perceptually insignificant portions (such as LSB embedding). Similarly, language based text watermarking techniques that embed information in the underlying structure of a sentence will be more robust than those that modify the surface representation of the sentence.

The proposed algorithm consists of combining these two techniques to form a new one which is both robust and feasible. Fig. 1 shows the general stages of this technique.

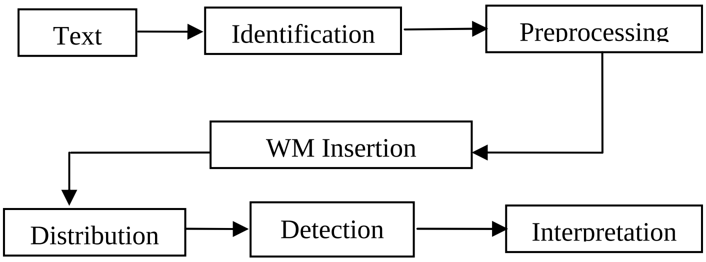

Fig. 1 Insertion & Detection Implementation Technique.

The first stage of the proposed watermarking technique is the identification of the matter whether it is a printed copy or a digital copy. Once the identification process is over, depending on a user needs or preferences, the user can decide to employ one or more functionality or a combination of them.

Preprocessing is the next stage. In this, if the matter is a scanned image text, then adjustments of the image qualities are done: else, parsing and classification etc., are done since language based watermarking depends mostly on substitution of synonyms and transformation based on syntax of the language.

The next stage is insertion of watermarking stage. In this stage if the text is a scanned one, then the following process is adopted.

VII. IMAGE TEXT WATERMARKING

English language has points in two letters, small “i” and small “j”, the positions of these points can be utilized for information security and watermarking. To be specific, information can be hidden in the point’s location of the pointed letters “i” and “j”. First, the hidden information is located at as binary. The first few bits are used to indicate the length of the hidden bits to be stored. Then, the cover medium text is scanned. Whenever a pointed letter is detected its point location may be affected by the hidden info bit. If hidden value is one the point is slightly shifted up; otherwise, the concerned cover-text character point location remains unchanged. The pointed letters are used to hold secret information bit “one” and the letters other than “i” and “j” are used to hold secret bit ‘zero’. Not all letters are holding secret bits since the secret bits since the secret information needs to fit in accordance to the cover-text letters.

VIII. LANGUAGE BASED TEXT WATERMARKING PROCESS

In this process a list of words are created. In every list with count elements, we can hide count-1 bits just by sorting the items. The first step is to force the list items into a specific order, which can be alphabetically, or a customized sorting. Then the sorted list is re-sorted, depending on the bits of the secret message. If the default sorting list is known, the list can be sorted again and the watermarked-list can be compared with the sorted list. The differences tell everything about the message-bits that produced the word order.

IX. DISCUSSION

Language text watermarking processing aims at designing algorithms that will analyze, understand and generate natural language automatically. Success of an information hiding system depends on medium which can only be achieved with large data sets. A statistically representative sample of natural language text is referred to as a corpus. Since most of language text water marking research is based on statistical analysis and machine readable form are essential.

Synonym substitution is the most widely used linguistic transformation for information hiding systems since it is the simplest transformation. Synonym substitution has to take the sense of the word into consideration. In order to preserve the meaning of the sentence the word should be substituted with a synonym in the same sense. For example, the word “bank” has at least three different senses as a financial institution, a river edge, or something to sit on. An electronic dictionary like Wordnet that classifies all words and phrases into synonym sets can be used to search for words that are synonym sets can be used to search for words that are synonyms for a given word. However, determining the correct sense of a given word, referred to as the word sense disambiguation task in language based watermarking, may present hard problems since it is hard to even derive a general definition for word sense.

A second type of transformation is the class of syntactic transformations, such as passivization and clefting, which change the syntactic structure of a sentence with little effect on its meaning. In addition to these, there is another group of syntactic transformations that are solely based on the categorization of the main verb of the sentence. Verbs can be classified according to shared meaning and behavior, and different classes of verbs allow different transformations to be performed in the sentence.

X. Benefits of the Proposed System

Some of the benefits of using this technique are: Confirmation of property, Follow up of unauthorized copies, Validation of identification and verification of integrity, labelling, usage control and protection of contents.

XI. Conclusion

A new method has been proposed in this paper that combines the best of both image based text watermarking techniques and language based watermarking techniques for an efficient copyright protection mechanism. The benefits accruing from this include robustness, imperceptibility, easy implementation etc.

References

[1] F.Hartung and M.Kutter, “Multimedia Watermarking techniques” IEEE, Vol.87,No.7, pp.1079-1107, July 1999.

[2] Young-Won Kim and Il-Seok Oh. “A survey on text watermarking techniques”, Proceedings of Honam-Jeju Korea Information Science Society, Vol.14, No.1, pp.34-3, August 2002 (in Korean).

[3] J.T.Brassil. S.Low and N.F.Maxemchuk, “Copyright protection for the electronic distribution of text documents” Proceedings of IEEE, Vol.87, No.7, pp.1181-1196, July 1999.

[4] D.Huang and H.Yan. “Interword distance changes represented by sine waves for watermarking text images”, IEEE Transaction. Circuits and systems for video technolo, Vol.11, No.12, pp.1237-1245, Dec 2001.

[5] T.Amano and D.Misaki, “A feature caliberation method for watermarking of document images”, Proceedings of ICDAR, pp.91-94, 1999.

[6] A.Bhattachariya and H.Ancin, “Data embedding in text for a copier system”, Proceedins of ICIP, Vol.2, pp.245-249, 1999.

[7] N.Ide and J.Vronis, “Word sense disambiguation: The current state of the srt” Computational Linguistics, Vol.24, no.1, 1998.

Suganya Ranganathan is pursuing her Ph.D at Karpagam University. She holds her MCA degree from Anna University, Chennai. Also completed her B.Sc Computer Science degree from Bharathiar University, Coimbatore. Being her research area is in software testing she also has interest on security oriented networks, Digital Processing and Software Engineering. She was working as a software engineer in UST Global for 2 years. And recently entered into academics. Organized workshops, FDPs, attended International conferences and published an International journal paper on Networks.

Ahammed Johnsha Ali is pursuing his Ph.D. Completed MCA at Anna University, Coimbatore and also has completed B.Sc Computer Science in Bharathiar University, Coimbatore. He has two and a half years of experience in software industry as a programmer and currently working as a Assistant Professor in Nehru College of Management. He has mush interest towards research in networks and also has interest in testing and multimedia.

Kathirvel.K has completed his M.Phil in the area of networks and completed MCA in Bharathiar University, Coimbatore. Participated and also has presented papers in Conferences. He is specialized in Networks and also in software engineering. He has organized workshops and seminars. Presently he is working as a Assistant Professor in Karpagam University.

Mohan Kumar.K has completed his M.Phil in the area of software engineering. Has completed MCA. Worked as a ISO Assistant , Software engineer and DMS incharge with Lotus TVS. Also worked as a Quality analyst and Test engineer. Presently working as Assistant Professor at Karpagam University, Coimbatore. Also has interest in Artificial Intelligence.