Proc. of the Second Intl. Conf. on Advances in Computer and Information Technology -- ACIT 2013
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-981-07-6261-2

An Adaptive Zero-Watermarking Approach for Text Documents Protection

Omar Tayan, Yasser M. Alginahi and Muhammed N. Kabir
IT Research Center for the Holy Quran (NOOR)
College of Computer Science and Engineering
Taibah University
P.O Box 344
Al-Madinah Al-Munawarrah, Saudi Arabia

Abstract—**This paper presents a new approach to addresses the problems associated with text documents protection, such as copyright protection and content verification. This problem has not been investigated much for text media and most of the work available in literature is based on multimedia forms. Therefore, due to the large amount of text documents available and with the extensive use of the Internet to distribute such documents it becomes essential to protect and secure information delivered through the Internet. Few techniques have been proposed to attain this goal; however, they are not sufficient to protect sensitive documents such as the Holy Quran. This paper presents a new adaptive approach based on zero-watermarking for highly-sensitive documents in order to achieve content verification and authentication without physically modifying the cover-text in anyway. Finally, this work is anticipated to open new research directions in text and multimedia watermarking.**

Keywords- security; protection; text-documents; zero-watermarking.

I. Introduction

The information technology era has enabled the ease of creation, distribution and reproduction of digital documents over the Internet. Therefore, with such benefits comes other challenges and threats in order to ensure the protection and security of such documents. Such issues opened the eyes of researchers in the area of security and document analysis to propose ways to protect documents from such threats. The content on the Internet is increasing exponentially, and finding ways to protect sensitive documents during communication becomes crucial in order to ensure the copyright protection, content verification and prevent counterfeiting. In addition, other threats which could be encountered during the communication process. This paper is primarily concerned with the security aspect of integrity with regards to widely disseminated digital text resources. Digital-watermarking is a known technique used to provide the necessary security for such digital content. Significantly, digital watermarking techniques are used to embed identification data into the host cover-document data, in which the embedded data is a function of the host data/content bit-sequences [1 - 3]. In this paper, we focus on zero watermarking, whereby only a certification-authority is able to decode the digital content and examine the watermark for purposes of authenticity verification. Such an approach would clearly provide an essential tool for a future reference body/organization concerned with the dissemination of sensitive/critical text-resources. The watermarking life cycle goes through the encoding process, followed by the distribution or transmission phase where some modifications due to attacks or transmission media could take place, and finally, the decoding phase where the watermark is detected at the receiver end.

This paper proposes an adaptive zero-watermarking technique to be used for text documents where the modifications of text such as the position of letters and/or diacritics (used in semantic languages) are not allowed. Such approach in text-watermarking has not received much interest as it is evident from the overview on zero-watermarking techniques [4], since here most of the techniques found in the literature were for images, with only few being found for text watermarking.

This paper is organized as follows: section II provides the literature survey, section III explains the proposed adaptive zero-watermarking framework, section IV discusses the results and finally section V concludes the paper.

II. Literature Survey

A review of the literature shows the maturity of watermarking based-techniques in digital natural language documents especially digital text-content in Latin and Chinese languages [5-7], with only few techniques presented for other languages such as Arabic [6, 12-13]. Moreover, watermarking of text-documents has been classified as linguistic and non-linguistic [1]. Linguistic techniques manipulate the lexical, syntactic and semantic properties of a document while trying to preserve the meanings; however, in the non-linguistic approaches changes are made to the text by using different text-attributes to embed a message. Text-watermarking techniques have been based on shifting techniques: such as, line-shift coding, word-shift coding and feature/character-coding, and natural-language based watermarking techniques, such as synonym-substitutions or semantic-transformation techniques which are language-dependent [1]. On the other hand, the relevant literature had also classified text-watermarking techniques into either of i) image-based techniques, ii) syntactic-based manipulation, and, iii) semantic-based manipulation

ISBN: 978-981-07-6261-2

techniques which involves replacing the original text with alternative words in order to embed a hidden message while preserving the meanings as far as possible [9].

Zero-watermarking schemes are robust and offer high imperceptibility. Here, in contrast to physically inserting the watermark bits, zero-watermarking schemes generate bit-patterns during the encoding process by extracting essential characteristics from the host-data which are then used in the decoding process [9]. It is noted, however, that most of the existing zero-watermarking approaches are designed for audio or image media types with insufficient research found using such methods for electronic text documents.

In zero-watermarking of digital-texts, the publisher's key is generated by extracting properties from the text without any modifications on the cover-text; therefore, the watermark is not physically embedded in the text, but rather, the characteristics of the host document are used to generate a new key. Zero-watermarking techniques have not been widely exploited and their use could be applied to formal/religious and sensitive texts where no modifications to the text can be tolerated. Examples of zero-watermarking techniques are found in [9 – 11]. The following section presents the proposed adaptive zero text-watermarking technique which can be applied to any language.

III. PROPOSED ADAPTIVE ZERO-WATERMARKING FRAMEWORK

This paper introduces a new design approach for verification and authentication protection of electronic plain-text documents based on logical/zero-watermarking. The watermark-logo is a unique signature of an organization or publisher that owns rights to the digital content/online document. The proposed algorithm performs a logical, rather than physical, embedding of watermark-data in the cover-document. The motivation here is to provide a means for the secure dissemination of critical and sensitive documents in which any physical modification can deem the document as invalid. Application examples are numerous, and may include Holy Scriptures in order to prove the original publisher of the text. In the approach proposed in this paper, an adaptive watermarking framework is employed which supports the tuning of algorithmic-parameters at the different phases of the encoding and decoding process in order to optimize our application-dependent cost-function.

A. Description

The watermark encoding process is illustrated in Figure 1, and operates as follows. The original publisher embeds a data-sequence obtained from an image-logo (watermark image, W_L) in a copy of the cover-document, T_C, following processing and classification into its constituent word-sets. The embedding phase is based on a spread-spectrum technique that inserts one-watermark bit per set, with the set-size being a variable parameter. Next, a key-generation algorithm is applied to produce a unique key which is registered together with the logo at a certification-authority (CA) – a trusted third-party in the digital community.

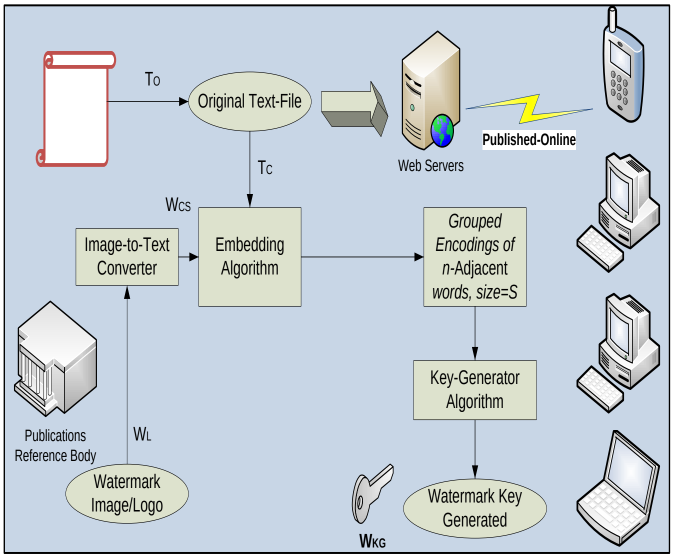

Figure 1: Watermark Encoding Process

Enquiries related to authentication, source-verification and tamper-detection tests of an online document are addressed using the decoding process shown in Figure 2, whereby the document under scrutiny, T_S, is passed for processing, which proceeds with the CA embedding the stored/embedded publisher logo. The text-classification, grouping and embedding used must be consistent with those in the encoder. A key-generator similar to the one in the encoder is used to extract a key based on the characteristics of the input document. The newly generated key is compared with the CA-registered key to confirm/reject document-ownership and authenticity (integrity). The document-owner is responsible for generating and registering the watermark-key (W_{KG}), logo and algorithm with the CA, whilst the CA would be responsible for the verification/decoding process upon client-requests. For the purposes of this paper, a blind and fragile watermark-extraction approach was used. The two-phase watermark encoding and decoding process can be comprehended in Algorithms 1 and 2.

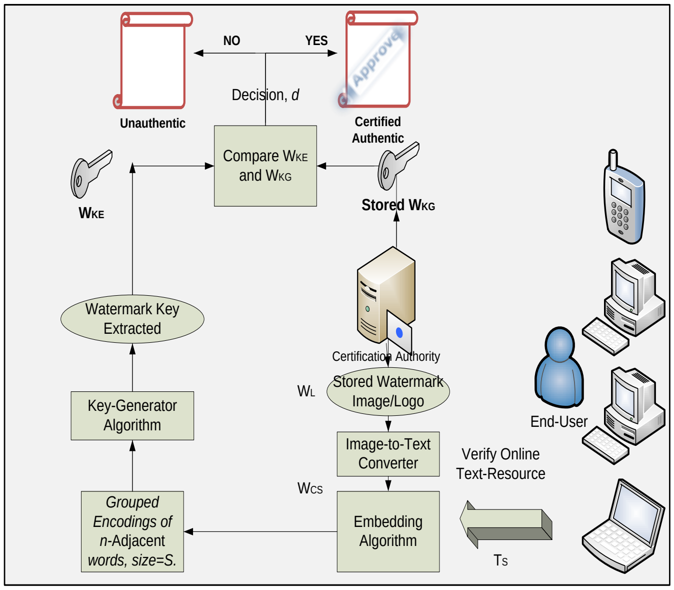

Figure 2: Watermark Decoding Process

B. *Watermark Encoding and Decoding*

Algorithm 1: *Encoding Algorithm*

```
Input: original cover-document $T_0$, logo-watermark $W_L$,
group-size S, key-generation policy $K_G$.
Output: key of watermarked text $W_{KG}$.
1. Convert $W_L$ to a bit-stream, $W_{CS}$.
2. Make a copy-document of the original cover-document, $T_C$.
3. Divide $T_C$ into $n$-word groups according to the Set-size $S$.
4. for $i$ = 1 to $n$
        get $i$-th word-group from $T_C$.
        add the Unicode values of all characters of all
        words in group $i$ to produce $sum_j$ . convert $sum_j$ to -
        binary values.
        get next embedding-bit $e$ from $W_{CS}$
        add $e$ to $sum_j$, and convert $sum_j$
   end for
5. Obtain the key-generated, $W_{KG}$, by combining the result for
all sets.
6. Output $W_{KG}$ and the character-stream, $W_{CS}$
```

Algorithm 2: *Decoding Algorithm*

```
Input: Document under scrutiny $T_S$, Watermark key: $W_{KG}$, and
original logo $W_L$.
Output: $W_{KE}$, decision-of-authenticity $d$.
1. Convert $W_L$ to a character-bit, $W_{CS}$.
Divide $T_S$ into $n$-word groups according to the set-size $S$.
4. for $i$ = 1 to $n$
        get $i$-th word-group from $T_S$.
        add the Unicode values of all characters of all
        words in group $i$ to produce $sum_j$ . convert $sum_j$ to
        binary values.
        get next embedding-bit $e$ from $W_{CS}$
        add $e$ to $sum_j$, and convert $sum_j$
   end for
5. Obtain the key-generated, $W_{KE}$, by combining the result for
all sets.
6. Compare $W_{KE}$ and $W_{KG}$ for similarity.
7. Output $d$ based on result of comparison.
```

Significantly, the watermark approach presented in this paper offers an adaptive scheme in the encoder-design, whereby key encoding-parameters such as the set-size, embedding and key-generation techniques are considered as variables set by the publisher based on the requirements of the target text-file to be encoded and published. The key-generation algorithm was only required to further encrypt the key generated from the encoder following the embedding phase.

IV. RESULTS AND DISCUSSION

The complete system, comprising of the encoder and decoder, was implemented in C++. Evaluation results had considered the computational-times required for the encoding and decoding processes using various sample electronic-texts. Numerical tests were executed for five sample text documents of different sizes. Tests were performed using a Linux operating system, and executed on a Pentium i3 processor with 1.7 GHz. The computational time required by the encoding and decoding algorithms is presented in Table I. The results show that the decoding process takes less time than the encoding time for each of the samples files observed. This reduced decoding time has an advantage, since the encoding process (and larger delay) for each text-document would only be encountered once, whereas the decoding process would typically be encountered multiple times for verification by many online clients.

TABLE I: Computational time of the proposed method

File Name	No. of Chars	Computational time Encoding(ms)	Computational time Decoding(ms)
Text1	28915	30	20
Text2	47974	40	30
Text3	54839	60	40
Text4	116794	80	70
Text5	166166	130	100

V. Conclusion

The increase use of the Internet suggests the need to secure and protect multimedia data of all types, especially digital text documents. From surveying the available text watermarking techniques, it is evident that most of the available techniques are not robust to all types of security attacks. This paper has proposed an adaptive algorithm for text-based zero-watermarking. The algorithm can be used to protect all digital textual-content from forgery and illegal content manipulation. Essentially, the algorithm works by embedding the watermark logo of the original-publisher in an identical-duplicate of the cover-document to generate a characteristic key, which is then compared with the characteristic key of another sample document to prove authenticity and ownership.

The zero-watermarking approach presented in this paper completely removes the vulnerability of watermarking attacks found in other/physical watermark-embedding approaches since no data embeddings are inserted into the host cover-document. The new design framework for text-based zero-watermarking can help in the protection of text documents from attacks which maybe encountered during the transmission process. It is anticipated that this work will open new research directions aimed at developing and advancing the state-of-the-art in text and multimedia based zero watermarking.

Acknowledgment

The authors would like to thank and acknowledge the IT Research Center for the Holy Qur'an (NOOR) at Taibah University for their financial support during the academic year 2012/2013 under research grant reference number NRC1-126.

References

[1] Adesina, A.O.; Nyongesa, H.O.; Agbele, K.K., "Digital watermarking: A state-of-the-art review", IST-Africa, 2010.

[2] Podilchuk, C.I.; Delp, E.J., "Digital watermarking: algorithms and applications", Signal Processing Magazine, IEEE, 2001, Volume: 18 , Issue 4.

[3] Swanson, M.D.; Kobayashi, M.; Tewfik, A.H., "Multimedia data-embedding and watermarking technologies", Proceedings of the IEEE, 1998, volume: 86 , Issue 6.

[4] YaYa Wang, Ke Luo, and Wei Gao; "Overview on Zero-watermarking Techniques," 2012 International Conference on Mechanical Engineering and Automation Advances in Biomedical Engineering, (ICMEA-2012), Vol.10, pp. 259 – 264.

[5] Zhichao Yu; Xiaojun Liu, "A New Digital Watermarking Scheme Based on Text" , International Conference on Multimedia Information Networking and Security, 2009. MINES '09.

[6] Shirali-Shahreza, M.; Shirali-Shahreza, S., "Persian/Arabic Unicode Text Steganography", Fourth International Conference on Information Assurance and Security, 2008.

[7] Davarzani, R.; Yaghmaie, K., "Farsi Text Watermarking Based on Character Coding", International Conference on Signal Processing Systems, 2009.

[8] Xinmin Zhou; Zhicheng Wang; Weidong Zhao; Sichun Wang; Jianping Yu, "Performance Analysis and Evaluation of Text Watermarking", International Symposium on Computer Network and Multimedia Technology, 2009.

[9] Jalil, Z.; Mirza, A.M.; Iqbal, T., "A zero-watermarking algorithm for text documents based on structural components", International Conference on Information and Emerging Technologies (ICIET), 2010.

[10] Zunera Jalil, Anwar 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.

[11] Jalil Z., Farooq M., Zafar H., Sabir M., and Ashraf E., "Improved Zero Text Watermarking Algorithm against Meaning Preserving Attacks,' World Academy of Science, Engineering and Technology, Vol. 46, 2010, pp. 592- 596.

[12] Al-Haidari, F.; Gutub, A.; Al-Kahsah, K.; Hamodi, J., "Improving security and capacity for Arabic text steganography using 'Kashida' extensions", IEEE/ACS International Conference on Computer Systems and Applications, 2009.

[13] Aabed, M.A.; Awaideh, S.M.; Elshafei, A.-R.M.; Gutub, A.A., "Arabic Diacritics based Steganography", IEEE International Conference on Signal Processing and Communications, 2007.