An Invisible Zero Watermarking Algorithm using Combined Image and Text for Protecting Text Documents

Jaseena K.U.[1]
Indira Gandhi National Open University
Delhi, India

Anita John[2]
Rajagiri School of Engineering and Technology
Kochi, India

***Abstract*—Authentication and copyright protection for digital contents over the Internet can be achieved through digital watermarking. The major components of the Internet are textual contents. Hence protection of plain text documents requires more attention. In this paper, we propose an invisible watermarking algorithm based on the occurrence of non-vowel ASCII characters that uses combined image and text watermark for protection of the text document. The watermark is logically embedded in the text and a watermark key is generated. Later the watermark is extracted to prove the identity. The experiments show that the new method is efficient as well as effective for protecting text documents.**

Keywords- Digital watermarking; authentication; watermark embedding and extraction; copyright protection

I. INTRODUCTION

Digital watermarking provides authentication and copyright protection for digital contents over the Internet. Nowadays the major content of Internet is plain text besides image, audio and video. For example the contents of newspapers, research papers, e-books, messages, and articles are plain texts [1][7]. So plain text needs complete protection.

Digital watermarking is a technique for inserting information into an image or text or audio, which can be later extracted for variety of purposes which includes identification and authentication [9]. A watermark is a unique logo or signature of an individual or an organization who owns the copyright of a digital content [5]. The important characteristics of a good watermarking algorithm are imperceptibility, authenticity, integrity and security.

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 watermarking methods for protecting text documents developed so far use either an image watermark or a textual watermark [1]. Watermarks consisting of both image and text make the text more secure and has better robustness. So it is efficient to use watermarks composed of both image and text watermark instead of using plain textual or image watermark in order to achieve better robustness [1].

The paper is organized as follows. The introduction section is followed by the fundamental concept of digital watermarking in Section II. A description of proposed watermarking (embedding and extraction) algorithm is presented in Section III. In section IV, the experimental results are specified and finally section V concludes the paper.

II. DIGITAL WATERMARKING

There are two types of digital watermarking: visible (perceptible) and invisible (imperceptible) [5] [7]. 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 [6]. The classifications are as follows:

1. Image Watermarking
2. Video Watermarking
3. Audio Watermarking
4. Text Watermarking

On the basis of necessary data for extraction, watermarks can be divided into two categories:

1. Blind
2. Informed

Blind watermarking is a technique in which original document is not required during watermark detection process. Whereas in informed watermarking, original document is required during watermark detection process.

The important issues that arise in the study of digital watermarking techniques are capacity, robustness, transparency and security [11]. Cryptography only provides security by encryption and decryption. However, encryption cannot protect the content after decryption [4] [10]. Unlike 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 [10]. But watermarks not only protect the content but also provide many other applications like copyright protection, copy protection, ID card security etc [8].

Text watermarking is an emerging area of research. Text watermarking algorithms developed so far can be classified in to following categories [1].

1. Image based methods
2. Syntactic methods
3. Semantic methods
4. Structural methods.

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. These types of text watermarking schemes are robust zero watermarking [3]. Many text watermarking techniques utilizing existence of double letters (aa-zz) in the text have been proposed for protecting text documents [5, 12**].**

Text watermarking solutions are not robust against random tampering attacks such as insertion, deletion and re ordering attacks. In this paper, we propose a zero text watermarking algorithm which is resistant towards random tampering attacks. The performance of this algorithm is analyzed with the algorithm specified in [1].

III. PROPOSED ALGORITHM

In [1], 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 [1]. 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.

In [2], 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.

In the proposed work, we have combined the algorithms in [1] and [2]. As in [1], we have utilized combined image and text watermark instead of using text watermark as in [2]. As in [2], we have utilized the occurrence of non-vowel ASCII characters for embedding watermark into the text document and for generating key instead of using the occurrences of double letters existing in text to embed the watermark as in [1].The proposed algorithm is a zero watermarking algorithm since the text document is not modified while embedding watermark, but the characteristics of text are used to generate a watermark key.

In this proposed algorithm, the text is first partitioned based on partition size (Pr). This Pr is considered as a delimiter to form text partitions. Depending on the value of GS (Group Size), partitions are combined to form text groups. Then the occurrence of all non-vowel ASCII characters is calculated in each group and maximum occurring non-vowel ASCII character is identified in each group to create MONV (Maximum Occurring Non-Vowel) list. This MONV list and combined image and text watermark is used to generate the watermark key. Then the watermark key is registered with a certification authority (CA), a trusted third party for copyright protection. The watermarks and watermark key is kept with the CA along with time and date. Later this key is used in the extraction algorithm to identify the original owner. In general, the watermarking process involves two stages,

1. Watermark Embedding
2. Watermark Extraction

Watermark embedding is done by the original author and the extraction of watermark is done by CA for the original author.

A. Embedding Process

The algorithm which is used to embed the watermark in the text and to generate watermark key is called embedding algorithm. The embedding algorithm takes the combined image and text watermark as input and produces a watermark key as output. The embedding process is shown in figure 1. First the watermark is split into image and text watermarks. In figure 1, the preprocessing of text and pre processing of image watermarks is done to make the watermark pure alphabetical.

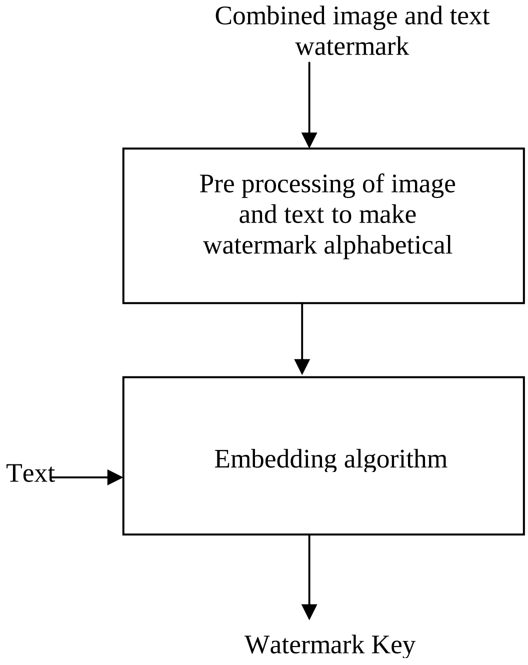
Combined image and text watermark
Pre processing of image and text to make watermark alphabetical
Text
Embedding algorithm
Watermark Key

Figure 1. Embedding Process

Preprocessing of text is the process of removing white spaces, special characters, digits etc to make the watermark pure alphabetical. During image pre processing, image is first converted in to grey scale and then scaling to 100x100 pixels. After image pre processing, image is converted in to plain text by normalization process. The two textual watermarks (watermarks obtained after text preprocessing and image preprocessing), partition size (Pr) and group size (GS) is given as input to the embedding algorithm.

1) Algorithm: Watermark Embedding

The algorithm used for embedding watermark as in [2] is presented below.

```
1. Input W, GS, Pr and T.
2. Split W into $W_{Img}$ and $W_{Txt}$
3. Preprocess $W_{Img}$ and $W_{Txt}$
4. Convert $W_{Img}$ to WT
5. Make partitions of T based on Pr
6. Make groups of text based on GS, where No. of groups = No. of partitions/GS
7. Count occurrence of non-vowel ASCII characters in each group and find Maximum Occurring Non-Vowel (MONV) in each group
8. Generate Watermark Key using steps from 9 to 12.
9. W = Merge (WT, $W_{Txt}$)
10. While (j<watermark_length) repeat steps 11 to 12
11. if ($W_j$ €MONVlist)
            Key (i) =0, Key (i+1) = Groupnumber(MONV)List
    else
            Key (i) =1,Key(i+1)=($W_j$ +k)MOD26, the Cipher text where k is in Z26 and Z26 represents 26 alphabets (a- z)
12.  Increment i by 1
13. Output Key
```

W: Combined watermark, W_{Img}: image watermark, W_{Txt}: text watermark, GS: Group size, Pr: Partition, T: text file, WT: text watermark (image to text), Key: Watermark Key

First, the watermark is split into image and text watermarks. The preprocessing of text and image watermarks is done to make the watermark pure alphabetical (W_{Txt} and WT). Based on Pr text partitions are made and based on GS, text partitions are combined to form text groups. Then, the occurrence of all non-vowel ASCII characters is calculated in each group and maximum occurring non-vowel ASCII character is identified in each group to create MONV (Maximum Occurring Non-Vowel) list. Using this MONV list and textual watermarks, the watermark key is generated using the algorithm described above. Then the watermark key is registered with a certification authority (CA), a trusted third party for copyright protection.

B. Extraction Process

The algorithm which extracts the watermarks using watermark key is called extraction algorithm. The algorithm takes the watermark key as input and extracts the watermarks (image and text) using the key from the text document. The extraction process is shown in figure 2.

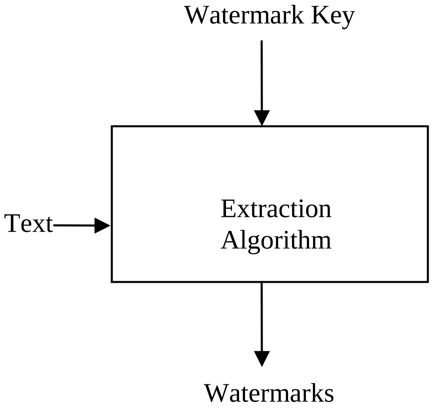

Figure 2. Extraction Process

1) Algorithm: Watermark Extraction

The algorithm used for extracting watermark as in [2] is given below.

```
1. Input Key and T.
2. Read Pr from Key and set counter=1.
3. Make partitions of T based on Pr
4. Make groups of text based on GS i.e. Numberofgroups =Number of partitions/GS
5. Count occurrence of non-vowel ASCII characters in each group and find Maximum Occurring Non-Vowel (MONV)
6. Populate MONV list in each group and extract watermark from text using steps 7 to 10.
7. L=length(Key), I=6
8. While (I<L) repeat 9 to 10
9. If (Key (I) equal 0)
          W (I)=group number(MONV)
     else
          W (I) = Key (I+1) i.e. cipher letter
10. I=I+1
11. Split W into $W_{Img}$ and $W_{Txt}$
12. Output $W_{Img}$ and $W_{Txt}$
```

Texts are partitioned and grouped based on Pr and GS value respectively. Occurrence of non-vowel ASCII characters is analyzed and maximum occurring Non-Vowel ASCII character (MONV) in each group is identified. Using the algorithm described above, watermarks are extracted.

IV. EXPERIMENTAL RESULTS

We have used different values for Pr for experiments. Group size was taken as 5 in all experiments. The combined image and text watermark used in experiments is shown in figure3.

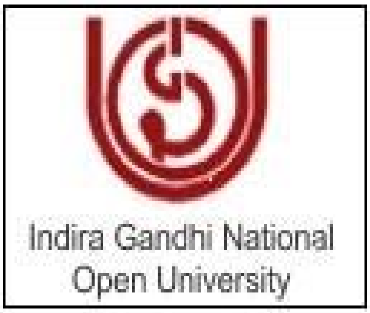

Figure 3. Original Watermark (Combined Image and text)

Table I shows the accuracy of extracted watermark for image, text and combined image and text watermarks under tampering attacks when algorithm based on non-vowel ASCII characters[2] are used. For comparing the accuracy of extracted watermarks, we have taken five values for Pr as 100,120,140,160 and 180.

TABLE I. ACCURACY OF EXTRACTED WATERMARK (IMAGE, TEXT AND OVERALL) UNDER RANDOM TAMPERING ATTACK

Pr	Text %	Image %	text+image%
100	97.05	99.72	98.39
120	79.41	100	89.71
140	97.06	99.72	98.39
160	97.06	100	98.53
180	100	100	100
Average	94.12	99.89	97

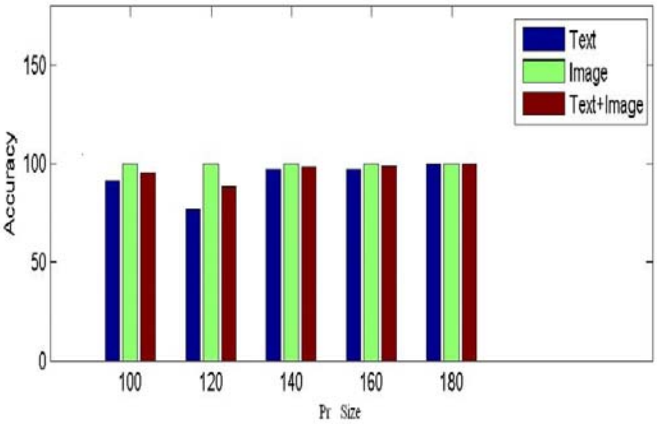

Figure 4. Graph Corresponding to Table I

From Table I, we can conclude that average accuracy of extracted watermark is 94% and that textual watermark is more sensitive to tampering attacks (insertion, deletion and reordering attacks) than image watermark. Hence the accuracy of text is lesser than image. But the combined accuracy (average) is 97%.

This is the case when the proposed algorithm is applied. In the proposed algorithm, we utilize the occurrence of non-vowel ASCII characters in the text to embed watermark. But if we are using the occurrence of double letters to embed watermark, the accuracy will be different. Table II demonstrates the accuracy of extracted watermarks when algorithm based on occurrence of double letters [1] is utilized.

TABLE II. ACCURACY OF EXTRACTED WATERMARK (IMAGE, TEXT AND OVERALL) WHEN CONCEPTS OF DOUBLE LETTERS ARE USED

Pr	Text %	Image %	text+image%
100	79.41	99.27	89.34
120	82.35	100.00	91.18
140	91.18	99.18	95.18
160	88.24	99.21	93.72
180	91.18	98.83	95.00
Average	86.47	99.30	92.88

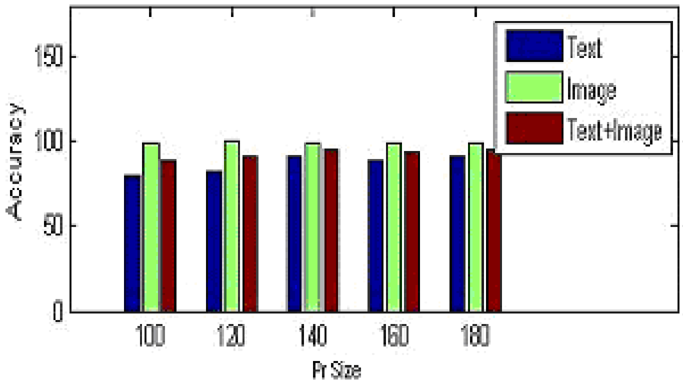

Figure 5. Graph Corresponding to Table II

It is clear from table II that the average accuracy of extracted watermark is always greater than 86%. In this case also the accuracy of text is lesser than image. But the combined accuracy (average) is 93%.

By comparing table I and table II, we can say that performance of proposed algorithm is better than the algorithm based on double letters. The comparisons of the two algorithms are demonstrated in table III and the graph corresponding to Table III is shown in Figure 6. From these, we can conclude that the proposed algorithm based on occurrence of non-vowel ASCII characters [2] is more resistant to random tampering attacks.

TABLE III. COMPARISONS OF ALGORITHMS

Pr	**Accuracy of Text+ image % based on algorithm[1]**	**Accuracy of text+ image% of proposed algorithm[2]**
100	89.34	98.39
120	91.18	89.71
140	95.18	98.39
160	93.72	98.53
180	95.00	100
Average	92.88	97

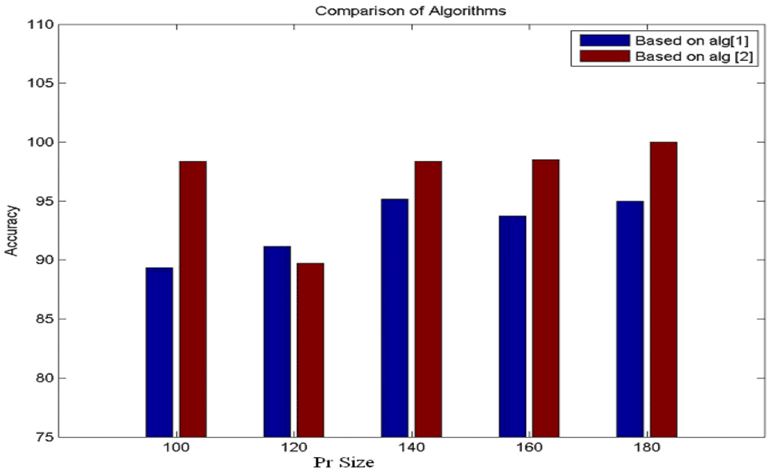

Figure 6. Comparisons of Algorithms (Graph Corresponding to Table III)

V. CONCLUSION AND FUTURE SCOPE

Watermarking is an emerging area of research for copyright protection and authentication of electronic documents and media. In this paper, a new watermarking technique based on the occurrence of non-vowel ASCII characters that uses combined image and text watermark for protection of the text document is proposed. Here, the occurrence of all non-vowel ASCII characters is analyzed and maximum occurring non-vowel ASCII character is identified to form MONV (Maximum Occurring Non-Vowel) list. The watermark key is generated using this MONV list and user given watermarks. Later this key is used by CA to extract watermarks to identify the original owner. We also compared this proposed algorithm with the algorithm specified in [1]. The experimental results show that the proposed algorithm has better performance.

According to our 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. Hence in future, some measures can be taken to reduce the length of the key.

REFERENCES

[1] Z.Jalil, A.M. Mirza ,“Text Watermarking Using Combined Image-plus- Text Watermark” ,IEEE, 2010.

[2] Z.Jalil, H.Aziz, S.B.Shahid, M.Arif, A.M.Mirza,” A Zero Text watermarking Algorithm based on Non-Vowel ASCII characters”,IEEE, 2010.

[3] Z.Jalil,., Farooq M., Zafar H., Sabir M., and Ashraf E.,” Improved Zero Text watermarking Algorithm against Meaning Preservation Attacks”, World Academy of Science, Engineering and Technology, 2010.

[4] X. Zhou,W. Zhao, Z. Wang, L. Pan,”Security Theory and Attack Anlysis for Text watermarking”, IEEE, 2009.

[5] Z. Jalil and A. M. Mirza, “An Invisible Text Watermarking Algorithm using Image Watermark”, International Conference on Systems, Computing Sciences, and Software Engineering (SCSS 2009), Innovations in Computing Sciences and Software Engineering, published by Springer, ISBN: 978-90-481-9111-6.

[6] M. Chandra, S. Pandey, R. Chaudhary, “Digital Watermarking Techniques for Protecting Digital Images”, IEEE, 2010.

[7] Z. Jalil, A.M. Mirza ,“A Review of Digital Watermarking Techniques for Text Documents” IEEE,,2009.

[8] X. Zhou, Z. Wang,W. Zhao,S. Wang,”Performance Anlysis and Evaluation of Text watermarking”, IEEE, 2009.

[9] Z. Xiao-hua1,M. Hong-yun,L. Fang, “A New Kind of Efficient Fragile Watermarking Technique“,Acta Electronica Sinica, 2004.

[10] F. Hartung, M. Kutter, “Multimedia Watermarking Techniques”, Proceedings of the IEEE, Vol. 87, No. 7, pp. 1085 – 1103, July 1999.

[11] 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.

[12] 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.