Text Watermarking using Combined Image and Text for Authentication and Protection

Jaseena K.U.
Indira Gandhi National Open University
India

Anita John
Rajagiri School of Engineering and Technology, Kochi
India

ABSTRACT
Digital watermarking provides authentication and copyright protection for multimedia contents over the internet. In addition to image, audio, and video, now a day's text is the most important medium traveling over the internet. Hence it needs to be protected. Text watermarking techniques that have been developed in past protects the text from illegal copying, forgery, and prevents copyright violations. In this paper, we propose a new text watermarking technique that uses combined image and text watermark and encryption to protect the text document. The watermark is logically embedded in the text and the text is encrypted. Later the text is decrypted and then the watermark is extracted to prove authenticity. Experimental results demonstrate the effectiveness of proposed algorithm.

General Terms
Text Document Security.

Keywords
Watermarking; watermark embedding; watermark extraction; authentication; encryption.

1. INTRODUCTION
Digital watermarking is the means for providing authentication and copyright protection for digital contents over the internet. Now a day's text is the most important medium traveling 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[1][6]. Hence text needs to be fully protected.

Digital watermarking is a technique for inserting information into an image or text or audio, which can be later extracted for various purposes including identification and authentication [2].Text watermarking techniques have been developed in past to protect the text from illegal copying, forgery, redistribution and to prevent copyright violations. Text watermarking techniques proposed so far for English language text uses either an image watermark or a textual watermark [1]. In many ways the text can be attacked, but generally an attack is the random insertion, deletion or reordering of words and sentences to and from the text [1][4][10]. The existing text watermarking algorithms are not robust against random tampering attacks (insertion, deletion or reordering of words) [1]. Watermarks composed of both image and text make the text secure and has better robustness. So for enhancing robustness, it is better to use combined image and text watermark instead of using plain textual or image watermark to fully protect the text document [1].

The remainder of the paper is organized as follows. An introduction to the concept of digital watermarking is given in Section 2 and a description of proposed watermarking (embedding and extraction) process is given in Section 3. Section 4 demonstrates the algorithms used for embedding and extraction. In section 5 the experimental results are specified. Finally, Section 6 gives the conclusion.

2. DIGITAL WATERMARKING
A digital watermark can be visible (perceptible) or invisible (imperceptible) [4] [6]. In the case of visible watermarking, watermarks are embedded in such a way that they are visible when the content is viewed. Invisible watermarks cannot be seen with the naked eye but they can be recovered with an appropriate decoding algorithm.

Based on the type of document to be watermarked, watermarking can be classified as [5]:

- Image Watermarking
- Video Watermarking
- Audio Watermarking
- Text Watermarking

Watermarking can be again robust or fragile. In robust watermarking technique, the modification to the watermarked content will not affect the watermark. But fragile watermarking is a technique in which watermark gets destroyed when watermarked content is modified or tampered with.

Based on processing method used, watermarking can be classified as:

- Spatial-domain techniques
- Transform-domain techniques

Watermarking in spatial-domain is straightforward. Spatial-domain watermarking are also the first watermarking schemes that were investigated by researchers. Transform-domain watermarking techniques have the convenience of allowing us more direct understanding of the content of cover data. This ease in understanding the content is exploited in many ways.

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

a. Blind

b. Informed

In blind watermarking original document is not required during watermark detection process. But in informed, original document is required during watermark extraction process.

The different issues that arise in the study of digital watermarking techniques are the following [10]:

- Capacity: What is the minimum amount of data that can be embedded in a given signal? Which is the optimum way to embed and the later extract this information?
- Robustness: How do we embed and retrieve data securely so that it would survive attacks at removal?
- Transparency: How do we embed data such that it does not degrade the underlying content?
- Security: How do we know that the information embedded has not been tampered, forged or removed?

Cryptography only provides security by encryption and decryption. But encryption cannot help the seller to monitor how a legitimate customer handles the content after decryption [8]. Hence we can say that there is no protection after decryption [3][8]. Unlike cryptography, watermarks can protect content even after they are decoded. Cryptography cannot prevent illegal replication of the digital content. Cryptography is only about protecting the content of the messages [8]. Digital watermarking technique is a prospective method to solve the problems mentioned above [7]. Watermarks are inseparable from the cover in which they are embedded. So besides content protection, they provide many other applications like copyright protection, copy protection, ID card security etc.

The process of breaking the system is different for cryptosystems and watermarking systems. The breaking of a cryptographic system occurs when the attacker can read the secrete message. But there are two stages for breaking of a watermarking system [11].

1. The attacker can detect the presence of watermarking.

2. The attacker is able to read, modify or remove the hidden message.

3. PROPOSED ALGORITHM

A robust text watermarking algorithm using combined image and text watermark to fully protect the text document is proposed as in [1]. A watermark is a unique logo or signature of an individual or an organization who owns the copyright of a digital content [4].

The previous work of text watermarking used combined image and text as watermark [1]. But in [1], text document is not encrypted. In the proposed work, we encrypted the text document to increase security.

In the proposed algorithm, the watermark is logically embedded in the text and then the text is encrypted. Later the text is decrypted and the watermark is extracted. In the proposed algorithm the occurrences of double letters existing in text are utilized to embed the watermark as in [1]. The watermark embedding is done by the original copyright owner of text and a watermark key is generated. The watermark is later extracted to prove authenticity. Thus the watermarking process involves two stages

1. Watermark embedding, and

2. Watermark extraction.

3.1 Steps of Embedding Process

The algorithm which embeds the watermark in the text is called embedding algorithm. The inputs for embedding algorithm are combined image and text watermark and the text document. The embedding algorithm performs preprocessing of image and the text to convert the watermarks pure alphabetical in nature.

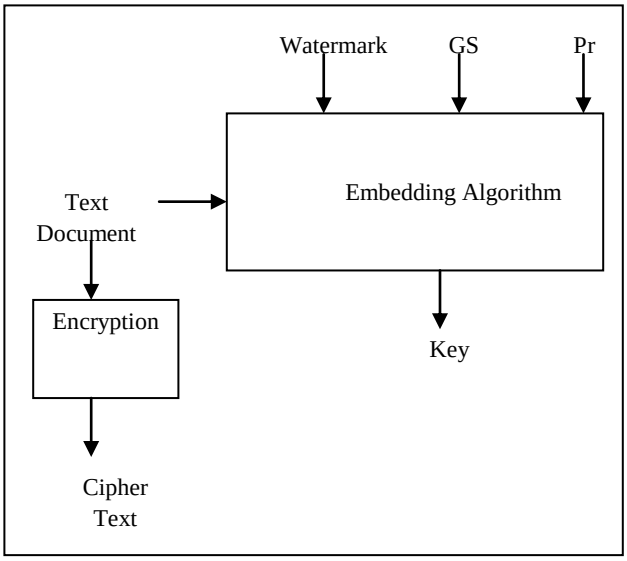

Figure 1. Embedding Process

1. Inputs are combined image and text watermark and text document.
2. Split the combined watermark into text and image watermarks.
3. Preprocess text watermark which includes, discarding white spaces, special characters, digits etc to make the watermark pure alphabetical.
4. Preprocess image which converts image to grayscale and scaling to standard size (100 x 100 pixels).
5. Convert image to plain text by normalization process.
6. The two textual watermarks (watermarks obtained after text preprocessing and image preprocessing) and partial key containing a partition size (Pr) and group size (GS) is given as input to the embedding algorithm(which is described in section 4).
7. The embedding algorithm generates the watermark key using the inherent properties of text.
8. Encrypt the text document using RSA encryption algorithm to increase security of text which is presented below.

3.2 Encryption and Decryption

In cryptography, encryption is the process of converting the plain text information into an unintelligible format called cipher

text so that the new cipher text cannot be easily read by another person. By encryption the confidentiality of the information can be maintained. Here any encryption algorithm can be used. If we are using more powerful algorithm, then it will be difficult for an attacker to decode the original information. Since we are concentrating mainly on watermarking, we have used a simple encryption called RSA algorithm for encryption.

After embedding the watermark into the text document and generated the watermark key, the text is encrypted using RSA encryption algorithm as described below to produce the cipher text.

The reverse operation of encryption is called decryption which coverts the cipher text (the encrypted information) back to the plain text.

3.2.1 RSA Algorithm

The RSA algorithm is named after Ron Rivest, Adi Shamir and Len Adleman, who invented it in 1977. RSA Algorithm is based on integer factorization assumption. The RSA algorithm can be used for both public key encryption and digital signatures. It consists of the following procedures: key generation, encryption, decryption, signature and verification [9].

3.2.1.1 Key Generation

1. Choose two big primes: p and q;
2. Calculate n=p*q;
3. Randomly choose an integer e, satisfying $1<e \leq \varphi(n)$, gcd (e, φ (n)) = 1. Totient function φ (n) denotes the number of positive integers less than n and relatively prime to n. Here φ (n) = (p-1)*(q-1). The public key is (e, n);
4. Calculate d, satisfying $ed \bmod \varphi(n) = 1$, the private key is (d, n);

3.2.1.2 Encryption Procedure

1. Partition the message m to groups m_i, i=1,2,…, $|m_i| = |n|-1$; (|a| means the length of a in binary form)
2. Encrypt each group: $c_i = m_i^e \bmod n$.
3. Connect each c_i and get the cipher text c.

3.2.1.3 Decryption Procedure

1. Partition c to c_i, i=1,2,…, $|c_i| = |n|-1$;
2. Decrypt each c_i : $m_i = c_i^d \bmod n$
3. Connect each m_i and recover the plain text m.

The signature and verification procedures of RSA are similar to encryption and decryption procedures.

3.3 Steps of Extraction Process

The algorithm which extracts the watermark is called extracting algorithm. It takes the key and watermarked text as input and extracts the watermark (image and text) from the watermarked text.

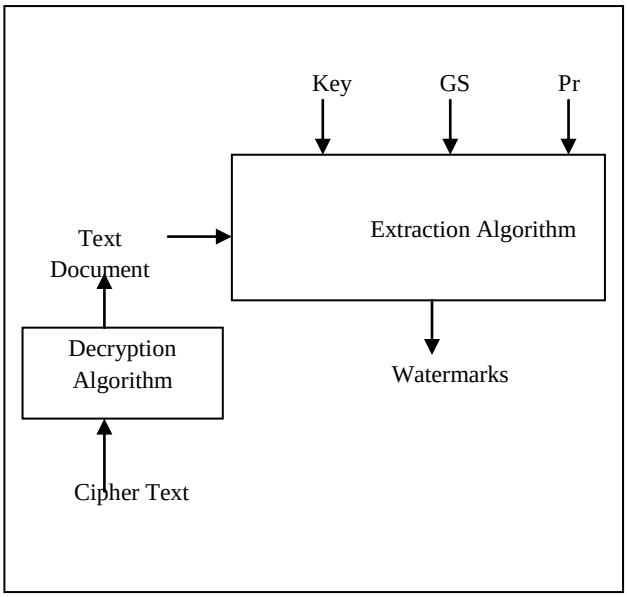

Figure2: Extraction Process

1. The watermark key and encrypted watermarked text are the inputs.
2. Decrypt the text to recover the watermarked text (using RSA Algorithm explained above).
3. Partition the text using Pr from watermark key.
4. Combine the partitions based on GS to make text groups as done in the embedding algorithm.
5. Occurrence of double letters in each group is analyzed and second maximum occurring letter (2MOL) in each group is identified.
6. Using the contents of watermark key, generate watermark from the text using extraction algorithm described in section 4.

4. ALGORITHMS

4.1 Watermark Embedding

The watermark (W) is initially separated in to image (WImg) and text (WTxt). WImg is then converted to alphabet and we obtain an alphabetical watermark (WT). The algorithm used for embedding watermark as in [1] is presented below:

1. Input W, GS, Pr and T.
2. Make partitions of T based on Pr.
3. Make groups of text based on GS, where Number of groups = No. of partitions/GS
4. Count occurrence of double letters in each group and find second largest occurring double letter in each group.
5. Populate 2nd Largest Occurring Letter (OL) list for each group.
6. Merge WT and WTxt to get W.

7. Generate watermark key using steps 8, 9, and 10.
8. While (j<watermark_length) repeat step 9 to 10.
9. if (wj € MOL list)

   Key (i) =0, key (i+1) = groupnumber (2 Maximum Occurring Letter)

   else

   Key (i) =1, Key (i+1) = (wj+k) MOD 26, where k is in Z26 and Z26 represents 26 alphabets (a- z)
10. Increment i.
11. Output Key.

W: watermark, WImg: image watermark, WTxt: text watermark, GS: Group size, Pr: Partition, T: text file, WT: text watermark

4.2 Watermark Extraction

The algorithm used for extracting watermark as in [1] is presented 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. Number of groups = Number of partitions/GS
5. Count occurrence of double letters in each group and find second largest occurring double letter in each group.
6. Populate 2Largest OL (Occurring Letter) list in each group.
7. Extract watermark from text using steps 8 to 11 with the help of Key.
8. L=length (Key), I=6
9. While (I<L) repeat 10 to 11
10. If (Key(I) equals 0) then

    W (I) =groupnumber(2Largest OL)

    else W(I)= Key(I+1) i.e. cipher letter.
11. Increment I by 1.
12. Split W into WImg and WTxt.
13. Output WImg and WTxt .

5. EXPERIMENTAL RESULTS

We used different values for Pr for experiments. Group size was kept 5 in all experiments. We used spatial domain technique for embedding watermark in to text document. The combined image and text watermark used in experiments is shown in figure 3.

Figure 3. Original Watermark (Combined Image and text)

The accuracy of extracted watermark for image, text and combined image and text watermarks under tampering attacks are given Table 1 and Figure 4. For comparing the accuracy of extracted watermarks, we have taken five values for Pr as 100, 120, 140, 160 and 180.

Table 1. Accuracy of Extracted watermark (Image, Text and Overall) under tampering attack

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

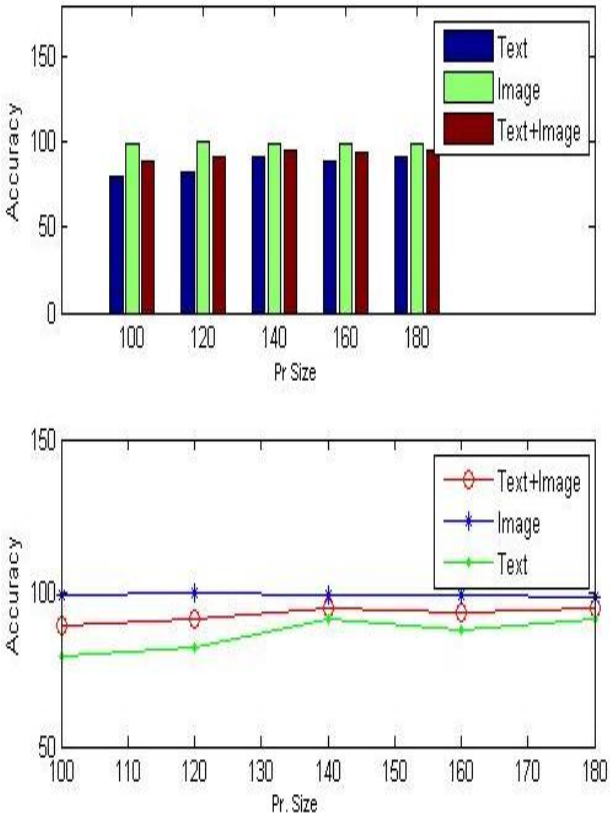

Figure 4. Accuracy of Extracted watermark under tampering attack

From Table 1 and Figure 4, it is clear that the accuracy of extracted watermark is always greater than 79% under tampering attacks. Textual watermark is more sensitive to tampering attacks (insertion, deletion and reordering) than image watermark. Hence the accuracy of text is lesser than image. However the combined accuracy is around 90%.

When no tampering is detected, 100% accuracy is obtained as shown in Table 2 and Figure 5.

Table 2. Accuracy of Extracted watermark (image, text and overall) when no tampering is detected

Pr	Text %	Image %	text + image %
100	100	100	100
120	100	100	100
140	100	100	100
160	100	100	100
180	100	100	100

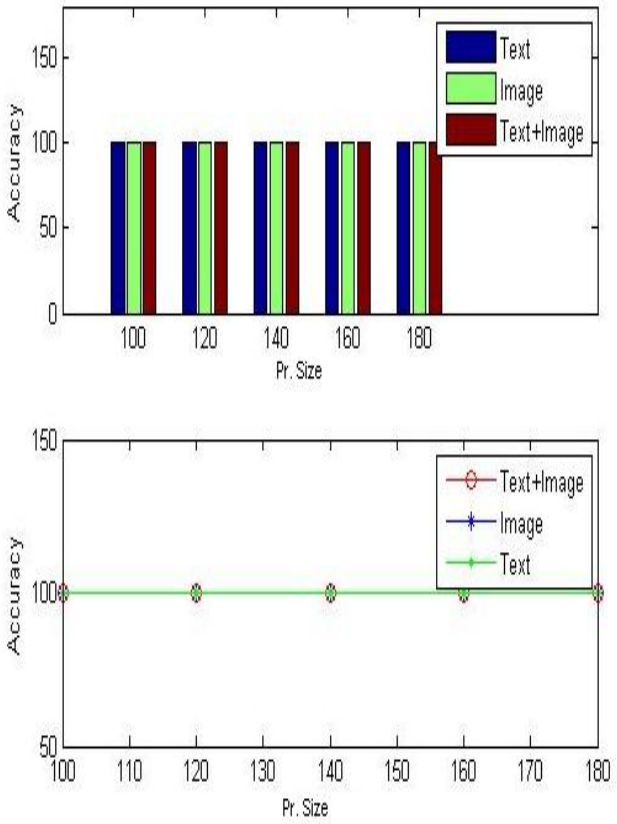

Figure5. Accuracy of Extracted watermark when no tampering is detected

The proposed technique of watermarking uses encryption to increase the security of text. For that we used RSA cryptographic algorithm. Proposed watermarking technique has following advantages:

1. Because our technique uses encryption, it increases the security of text document. If watermarking key is hacked by an attacker still the attacker will not be able to identify the text because it is encrypted.
2. It is a blind watermarking technique. So, original document is not required at the time of watermark recovery.

6. CONCLUSION

Watermarking is an emerging research area for copyright protection and authentication of multimedia content. In this paper, a new watermarking technique is specified that uses combined image and text watermark and encryption. First we embedded a watermark in to text using the algorithm described previously. After embedding watermark into text, a watermark key is generated. Then we encrypted the text with RSA. This provides an additional level of security for text documents. Later the cipher text is decrypted and watermark is extracted. Then extracted watermark is compared with original watermark to prove authenticity.

Most of the research related to watermarking is going on in the field of image watermarking. The research works related to text watermarking is less. The security of text can be enhanced by using another powerful encryption technique. The future work includes the development of benchmarking tool for text watermarking.

7. REFERENCES

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

[2] Z. Xiao-hua1,M. Hong-yun,L. Fang, "A New Kind of Efficient Fragile Watermarking Technique",Acta ElectronicaSinica, 2004.

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

[4] Z. Jalil, A. M. Mirza ,"An Invisible Text Watermarking Algorithm Using Image Watermark" ,Innovations In Computing Science and Software Engineering, 2010.

[5] M. Chandra, S. Pandey, R. Chaudhary, "Digital Watermarking Techniques for Protecting Digital Images", IEEE, 2010.

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

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

[8] Frank Hartung, Martin Kutter, "Multimedia Watermarking Techniques", Proceedings of the IEEE, Vol. 87, No. 7, pp. 1085 – 1103, July 1999.

[9] Jiezhao Peng1, Qi Wu, "Research and Implementation of RSA Algorithm in Java", IEEE, 2008.

[10] 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.

[11] R. Chandramouli, N.Memon, "Digital Watermarking".