Text Watermarking Using Combined Image & Text

Dr. K. Rameshbabu*,P. Prasannakumar**,Dr. K E Balachandrudu***

Professor& Dean(academics),jcem,shivajiuniversity,M.H,India,
M.Tech(eee) student,EVMCET,JNTUK,A.P,India,
professor &Dean(admin),jcem, shivajiuniversity,M.H,India,

Abstract

Authentication and copyright protection of digital contents over the internet is an important issue. Digital watermarking provides a complete authentication and copyright protection solution for this problem. Besides, image, audio, and video; text is the most dominant medium travelling over the internet and it requires complete protection. Text watermarking techniques have been developed in past to protect the text from illegal copying, forgery, redistribution and to prevent copyright violations. In this paper, we propose a novel text watermarking algorithm using combined image-plus-text watermark to fully protect the text document. The watermark is logically embedded in the text and is extracted later to prove ownership. Experimental results demonstrate the effectiveness of proposed algorithm under localized as well as dispersed tampering attacks on the text.

Keywords—authentication, copy right protection, tempering attacks, novel text, complete protection

I. Introduction

Increasing use of digital media and internet has made this world, a global village. Besides, the digital world is also encountering problems of copyright protection, authentication, illegal copying and redistribution of digital contents due to the ease of information sharing in a nominal time. Text is the most dominant medium existing in the digital world, besides image, audio, and video; hence requires complete protection. The major component of websites, newspapers, e-books, research papers, legal documents, letters, SMS messages, poetry, blogs etc is the plain text; therefore, it is necessary to protect text.

II.Motivation

Increasing use of digital media and internet has made this world, a global village. Besides, the digital world is also encountering problems of copyright protection, authentication, illegal copying and re-distribution of digital contents due to the ease of information sharing in a nominal time. Text is the most dominant medium existing in the digital world, besides image, audio, and video; hence requires complete protection. The major component of websites, newspapers, e-books, research papers, legal documents, letters, SMS messages, poetry, blogs etc is the plain text; therefore, it is necessary to protect text.

III. Proposed algorithm

A text watermarking algorithm based on the occurrence of double letters existing in text are used for protection of the text document is proposed. In this algorithm, the occurrence of all double letters is analyzed in each partition and maximum occurring double letters is identified to form MOL (Maximum Occurring Letter) list. The author key is generated using this MOL list 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. 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.

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. we have utilized combined image and text watermark instead of using text watermark ., we have utilized the occurrence of double letters 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 .The proposed algorithm is a Novel text watermarking algorithm using combined image plus text watermark 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 double letters is calculated in each group and second maximum occurring double letters is identified in each group to create 2MOL (Second Maximum Occurring letter) list. This 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. Water marking Extraction

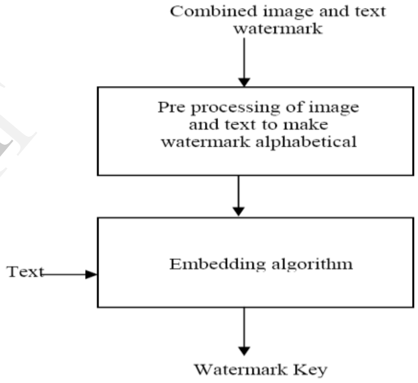

Combined image and text watermark
Pre processing of image and text to make watermark alphabetical
Text
Embedding algorithm
Watermark Key

Fig: 3.1embeding watermark

III.1 embedding process

The algorithm which is used to embed the watermark in the text and to generate water mark 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 reprocessing of text and preprocessing of image watermarks is done to make the watermark pure alphabetical.

Preprocessing of text is the process of removing white spaces, special characters, digits etc to make the watermark pure alphabetical. During image preprocessing, image is first converted in to grey scale and then scaling to 100x100 pixels. After image preprocessing, 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.

III.2 Watermark Embedding Algorithm

The algorithm used for embedding watermark is presented below.

```
1. Input W, GS, Pr and T.
2. Split W into WImg and WTxt
3. Preprocess WImg and WTxt
4. Convert WImg 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 double letters in each group
and find Second Maximum Occurring    (2MOL) in
each group
8. Generate Watermark Key using steps from 9 to 12.
9. W = Merge (WT, WTxt)
10. While (j<watermark_length) repeat steps 11 to 12
11. if (Wj €MONVlist)
      Key (i) =0, Key (i+1) =
Groupnumber(2MOL)List elseKey (i)
=1,Key(i+1)=(Wj +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: watermark, WImg: image watermark, WTxt: text watermark, GS: Group size, Pr: Preposition, T: text file, WT: text watermark, AK: Author keyThe Watermark (W) is first split into image (WImg) and text (WTxt). WImg is first converted to alphabet and we obtain an alphabetical watermark (WT). Then, depending on preposition (Pr) and group size (GS) Input by user (partial key), partitions and groups are formed. In the next step, the occurrence of each double letter is counted in each group and the 2nd largest occurring double letter in each group is identified (2MOL).The key generator generates the author key by using watermark (W) and 2MOL list as shown in the algorithm and generates the author key(AK). This author key is then registered with the A along with the watermark, original text, current date, and time.

III.3 Extraction process

The algorithm used to extract the watermark from the watermarked text is known extraction algorithm. It takes the author key and watermarked text as input and extracts the watermark (image-plus-text) from the text. The algorithm is kept with the Certifying Authority that uses it to resolve copyright issues, if any, at a later stage. The detailed extraction algorithm is as follows:

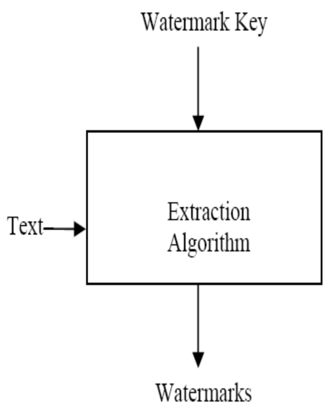


Figure 2. Extraction Process

III.4 Watermark Extraction Algorithm

The algorithm used for extracting watermark is given below.

```
1. Input AK and T.
2. Read Pr from AK 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
6. Populate 2MOL (2nd Maximum Occurring Letter)list in each group.
7. L=length(AK), I=6
8. While(I<L)repeat 9 to 10
9. If(AK(I)equals 0) ;    W(I)=groupnumber(2MOL)
    Else    W(I)= AK(I+1) i.e. cipher letter
10. I=I+1
11. Split W in WImg and W
12. Output WImg and WTxt
```

In the extraction algorithm, text is partitioned using preposition(Pr) from author key (AK). Then partitions are combined to make text groups as done previously in the embedding algorithm. Afterwards, occurrence of double letters in each group is analyzed and second maximum occurring letter (2MOL) in each group is identified. The contents of author key (AK) are then used to obtain watermark from the text. The reverse process of figure 1 is performed in the extraction process, where extraction algorithm takes place of embedding algorithm.

IV.Results

Watermark Embedding

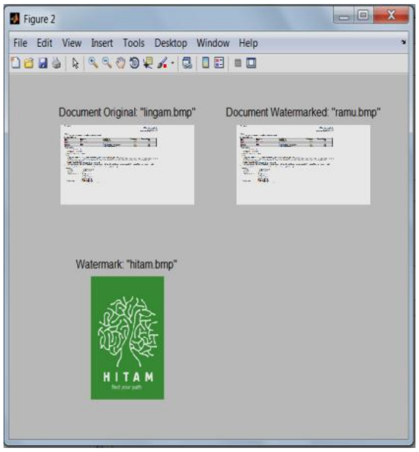
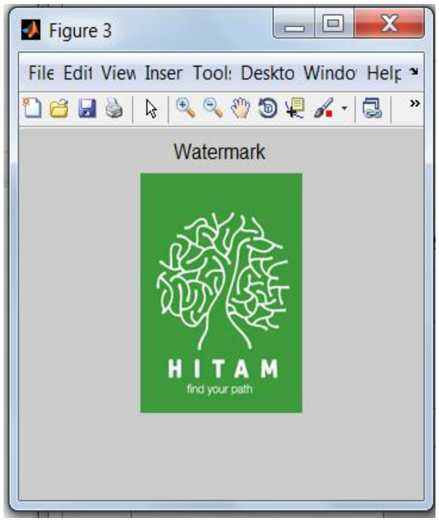


Figure 2
File Edit View Insert Tools Desktop Window Help
Document Original: "lingam.bmp"
Document Watermarked: "ramu.bmp"
Watermark: "hitam.bmp"
HITAM
find your path

4.1Fig: Watermark Embedded Process Result

Left most images isa original image and it is a input. Right most image is a watermarked image and below image indicates watermark.

Watermark Extraction: In Watermark extraction process watermark is extracted from original image to prove,ownership.



4.1.2 Fig: Extracted watermark

IV.1 Experimental results

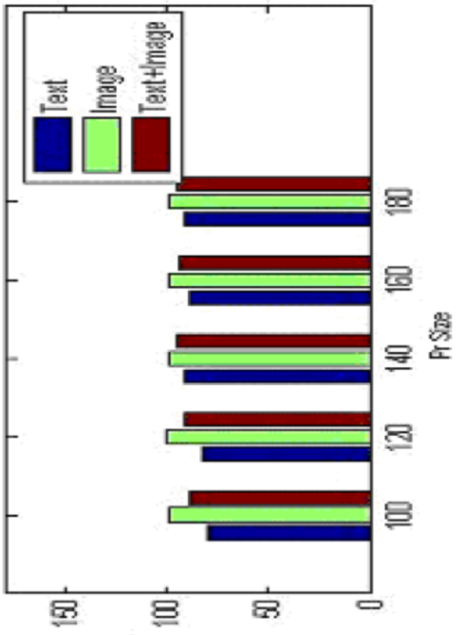
We have used different values for Pr for experiments. Group size was kept same in all experiments. The combined image and text watermark used in experiments is shown in figure.

The accuracy of extracted watermark under tampering attacks is shown in table However, the overall accuracy of extracted watermark (image plus text) is 92%.and Image watermark is 99% and it is more resilient towards dispersed tampering attacks since the accuracy is a 86 % for all text samples.

Table shows the accuracy of extracted watermark for image, text and combined image and text watermarks under tampering attacks.

4.1Table: Accuracy of extracted watermark (image, text and overall)

S.no	Pr	Text%	Image %	(Text+Image)%
1	100	79.41	99.27	89.34
2	120	82.35	100.0	91.18
3	140	91.18	99.18	95.18
4	160	88.24	99.21	93.72
5	180	91.18	98.83	95.00
6	Avg	86.47	99.30	92.88



<………. Accuracy-------

4.1.3 Figure: Graphs Corresponding to Table 4.1

V. Comparisions:

Text watermarking is an important area of research however; the previous work on digital text watermarking is quite inadequate. The previous work on digital text watermarking can be classified in the following categories; an image based approach, a syntactic approach, a semantic approach and the structural approach. In image based approach towards text watermarking, watermark is embedded in text image. Brassil, were the first to propose a few text watermarking methods utilizing text image. Later Maxemchuk,analyzed the performance of these methods.

Huang and Yan proposed an algorithm based on an average inter-word distance in each line. In syntactic approach towards text watermarking, the syntactic structure of text has been used to embed watermark. Mikhail J. Atallah, et al. first proposed the natural language watermarking scheme by using syntactic structure of text. Hassan et al. performed morpho-syntactic alterations to the text to watermark it. An overview of available syntactic tools for text watermarking was provided in. In semantic approach, semantics of text are utilized to embed the watermark in text. Atallah. Were the first to propose the semantic watermarking schemes in the year 2000. Later, the synonym substitution method was proposed. A noun-verb based technique for text watermarking was also proposed which exploit nouns and verbs in a sentence parsed with a grammar parser using semantic networks. Later Mercan, et al. proposed an algorithm of the text watermarking by using typos, acronyms and abbreviation to embed the watermark. Algorithms were developed towatermark the text using the linguistic semantic phenomena of presuppositions. The algorithm based on text meaning representation (TMR) strings has also been proposed. The structural approach is the most recent approach used for copyright protection of text documents.

A text watermarking algorithm for copyright protection of text using occurrences of double letters in text to embed the watermark has recently been proposed. Another algorithm which use preposition besides double letters to watermark text is also proposed recently. Text watermarking algorithms using binary text image are not robust against re-typing attack. The text watermarking methods based on semantics are language dependent. The synonym based techniques are not resilient to the random synonym substitution attacks. The structural algorithms are not applicable to all types of text documents and the algorithms are restricted to only alphabetical watermark or only image watermark. To increase robustness, it is better to use combined image-plus-text watermark instead of using plain textual or image watermark. Hence, we propose a text watermarking algorithm which uses combined image-plus-text watermark.

VI. Conclusions

Text watermarking methods for English language text proposed so far; use either an image watermark or a textual watermark. The existing text watermarking algorithms are not robust against random tampering attacks. Watermarks composed of both image and text, make the text secure and has better robustness. We have developed a text watermarking algorithm, which uses combined image-plus-text watermark to watermark the text document. Watermark can later be separately identified to prove the ownership. We evaluated the performance of the algorithm for localized and dispersed random tampering attack in 20 texts. The results show that the algorithm using text plus image watermarks are more robust, secure and efficient against random tampering attacks.

VII. Futurescope

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.

VIII. References

[1] A. Khan, A. M. MiSrza and A. Majid, "Optimizing Perceptual Shaping of a Digital Watermark Using Genetic Programming", Iranian Journal of Electrical and Computer Engineering, vol. 3, pp. 144-150, 2004.

[2] J. T. Brassil, S. Low, N. F. Maxemchuk, and L. O'Gorman, "Electronic Marking and Identification Techniques to Discourage Document Copying", IEEE Journal on Selected Areas in Communications, vol. 13,no. 8, pp. 1495-1504, October 1995.

[3] J. T. Brassil, S. Low, and N. F. Maxemchuk, "Copyright Protection for the Electronic Distribution of Text Documents", Proceedings of the IEEE, vol. 87, no. 7, pp.1181-1196, July 1999.

[4] N. F. Maxemchuk, S. H. Low, "Performance Comparison of Two Text Marking Methods", IEEE Journal of Selected Areas in Communications (JSAC*)*,vol. 16 no. 4 1998. pp. 561-572, May 1998.

[5] N. F. Maxemchuk, "Electronic Document Distribution," AT&T Technical Journal, September 1994, pp. 73-80. 6.[6] N. F. Maxemchuk and S. Low, "Marking Text Documents", Proceedings of the IEEE International Conference on Image Processing, Washington,DC, , pp. 13-16, Oct. 26-29, 1997.

[7] D. Huang and H. Yan, "Interword distance changes represented by sine waves for watermarking text images", IEEE Trans. Circuits and Systems for Video Technology, Vol.11,No.12, pp.1237-1245, Dec 2001.

[8] M. J. Atallah, C. McDonough, S. Nirenburg, and V. Raskin, "Natural Language Processing for Information Assurance and Security: An Overview and Implementations", Proceedings 9th ACM/SIGSAC New Security Paradigms Workshop, Cork, Ireland, pp. 51–65, September,2000.

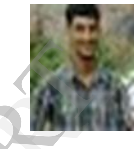
[9] M. J. Atallah, et al., "Natural language watermarking: Design,analysis,and a proof-of-concept implementation", Proceedings of the Fourth Information Hiding Workshop, vol. LNCS 2137, Pittsburgh, PA, 25-27 April 2001.

[10] Hassan M. Meral et al., "Natural language watermarking via morphosyntactic alterations", Computer Speech and Language, 23, 107-125, 2009.

[11] Hasan M. Meral, et al, "Syntactic tools for text watermarking", 19th SPIE Electronic Imaging Conf. 6505: Security, Steganography, and Watermarking of Multimedia Contents, San Jose,
12].www.rakshakfoundation.org/projects/c)
www.americanbarfoundation.org/d)
www.deepdyve.com/search-e)
www.pen.org/blogf)www.deepdyve.com/search-
[13]..www.rakshakfoundation.org/projects[14].h)www.deepdyve.comi)
www.deepdyve.com/searchj)doi.ieeecomputersociety.org/10.1109/ETCS.2010.494k)www.deepdyve.com

IX. Authors brief profiles:

Dr.K.Rameshbabu∗ professor&Dean(academics) ,E&TCEDept ,sspmJCEM, karad M.H. he is holding B.E (ece),M.Tech, Ph.D having 17+years of experience in electronics and Telecommunication Engineering area .he is member in ISTE,IEEE & Java Certified programmer(2,0)PGDST holder. he has lot of experience in academics and industrial related real time projects. He is paper setter for many autonomous universities and visiting professor for image processing, electron devices & communications etc.u can be reached at:



Mr .prasannakumar.ponnam** holding B.TECH (eee) with I class. fromCIT, JNTUK Guntur. Now he is parsueing M.Tech from EEE department ,EVMCET,NRT,Guntur.he is very much interested in research work in the image processing,power Electronics,Machines etc. he is very good human being.Hecanreach at:

DrK.E Balachandrudu*** professor& Dean(admin) ,CSE Dept ,sspmJCEM, karad, M.H. he is holding B.E (cse),M.E, Ph.D having 18+years of experience in compuer Science & Engineering area .he is member in ISTE,IEEE & sun micro system Certified programmer,. he has lot of experience in academics and industrial related real time projects. He is paper setter for many autonomous universities and visiting professor for image processing, cloud & mobile computingetc.u can be reached