A Robust Method for Integrity Protection of Digital Data in Text Document Watermarking

Leena Goyal
Assistant Professor
Department of Computer Science
Vivekananda Institute of Technology, Jaipur

Manoj Raman
Assistant Professor
Department of Computer Science
Vivekananda Institute of Technology, Jaipur

Prateek Diwan
Assistant Professor
Department of Computer Science
Vivekananda Institute of Technology, Jaipur

Mukesh Kumar Vijay
Assistant Professor
Department of Computer Science
Vivekananda Institute of Technology, Jaipur

Abstract

Digital watermarking provides authentication, validation and copyright protection for multimedia contents over the internet. Text is the most widely used means of communication in addition to images, audio and video clips. So it may be required to be protected. Text watermarking techniques that have been developed in past protects the text from illegal copying, imitation, and prevents copyright violations. In this paper, we have proposed an algorithm that ensures the integrity and confidentiality of the document. In this technique watermark is created based on the contents of the document and embeds it without changing the contents of the document and also encrypts the text to provide confidentiality. To authenticate and prove the integrity of the document the watermark can be easily extracted and verified for tampering.

Keywords: Confidentiality, Integrity, encryption, shuffling, substitution, text watermarking, substitution, tampering

I. INTRODUCTION

In present scenario a good amount of information is being stored and distributed in digital form. Growth in internet services and availability of different materials like videos, images, e-books, e-journals and other contents increases the need of copyright protection of all digital data. Data stored in digital form is always being in risk in a number of ways like; digital data can be copied and republished by some other name. Due to advancement in computer technology, digital data can be easily edited and manipulated.

One way to protect digital information is encryption, but once the data has been decrypted it is easy to copy and use the same data. Digital watermarking can be a partial solution to this problem of copyrighting of data. Watermarking is a process of embedding information in any digital media like image, audio or video so that the information can be easily exchanged over the internet (1). Digital watermarking techniques provide different methods to control the copyright and to provide integrity to various forms of digital data(4). If in any case an opponent or copier modified the data or duplicates it, then the owner or authorized user can claim to the originality of the document by applying any required watermarking process.

II. PROPOSED ALGORITHM

Most of the researches for text document integrity emphasis on the techniques to hide data which is difficult to perceive. These techniques are mainly classified into two categories. First one focuses on the blank spaces between two consecutive words, characters and lines. Second one focuses on embedding an image into text document and then by generating watermarking key or sometimes both image and document can be converted into encrypted form to provide confidentiality also(5). Drawback of first technique is, if we have the watermarked text file then we can easily identify that where the space is being modified between the words by rewriting the ext. In the second method, if the information (image) is not being embedded at the edges of textual characters then it could be easily detected. Even if we are encrypting text and image to provide confidentiality, but if we are able to get image then we can easily get information about the text.

In our paper we have proposed a new method for text document integrity and confidentiality protection. In this method watermark is created based on the most frequent contents of the document and embeds it without changing the contents of document and also encrypt the watermarked text to provide confidentiality. To authenticate and prove integrity of document the watermark can be easily extracted and verified for checking tampering attack.

In the proposed method first we take the text document as input and select three most frequent words as keywords. Length of these 3 keywords specifies the shifting amount and average of keywords length specifies the shifting amount for final shuffling. Scanning of the document is started word by word with comparing each word with keywords. How to shuffle the words is decided on the matching keywords and according to the length of keywords substitution technique is applied. Finally we get encrypted form of text and then we convert the encrypted text document into corresponding image.

A. Algorithm:

1) *Select three most frequent words as keywords (K1, K2, K3).*
2) *Calculate length of each keyword (L1, L2, L3)*
3) *Calculate average length of keywords (L= (L1+L2+L3)/3*
4) *Read each word (W) from document.*
5) *If W equals K1*
   Shuffle each character by L1
6) *If W equals K2*
   Shuffle each character by L2
7) *If W equals K3*
   Shuffle each character by L3
8) *Scan the encrypted document again and shuffle each character by L*
9) *Output is encrypted text document.*

B. Substitution technique:

1) *Substitution has been done on the basis of keyword length.*
2) *In the substitution technique we are taking first keyword as a start point and second keyword as an ending point.*
3) *In between these keywords we are adding each character with the length of keyword and then changing into character value.*
4) *When in between words are completed then we apply the same process of selecting start and end points.*
5) *Length of the selected keywords specifies the shifting amount for shuffling method1, method2 and method3 respectively.*
6) *Average length specifies the shifting amount for shuffling method4.*
7) *The whole process will continue up to the end of file.*
8) *Then after we get the encrypted file of text by substitution*

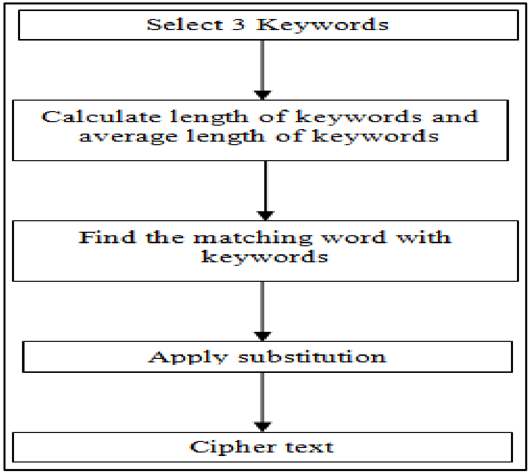

Select 3 Keywords
Calculate length of keywords and average length of keywords
Find the matching word with keywords
Apply substitution
Cipher text

Figure 1

C. Advantages of proposed method

- Encryption provides confidentiality
- Either if one has the image file he is not able to retrieve the text.
- We can send this image over the network without further encryption.
- If there is a minor modification in the text file whereas transmitting over the network we can easily identify it.

III. IMPLEMENTATION

A. Implementation at sender side

At sender side input.txt is taken as input file whose contents needs to be transmitted to the distant receiver and results in the three output files as ‘keyword.txt’, ‘sender_out.txt’ and ‘sender_out.jpg’ which are the keywords which are used to encrypt the file, the encrypted output file and the image of the encrypted output file respectively. The image output file ‘sender_out.jpg’ is used to compare against the modification of the text at the receiver side.

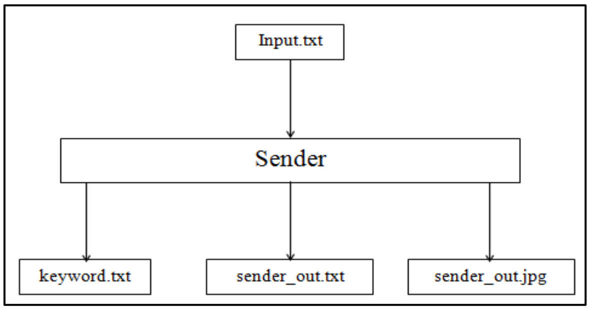


Figure 2

B. Implementation at receiver side

At receiver side input.txt and keyword.txt is taken as input file which is received at the receiver and the keywords used by the sender to encrypt the text file respectively and results in the two output files as receiver_out.txt and receiver_out.jpg which are the encrypted output file and the image of the encrypted output file respectively. The image output file receiver_out.jpg is used to compare against the modification of the text at the receiver side with sender_out.jpg.

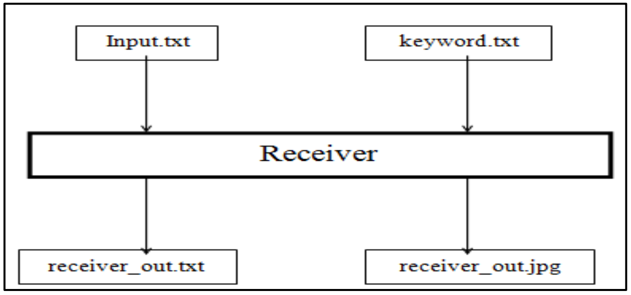


Figure 3

C. Comparison for integrity test

For comparison two input files sender_out.jpg and receiver_out.jpg are taken and executed at the receiver side. This script compares both the images and the comparison represents whether the two images are same or not. If both images are same then there is no modification in the text on the go, if comparison shows that the images are different then we can say that the text is being modified during the transmission from sender to receiver and some action needs to be taken to ensure the integrity of the text.

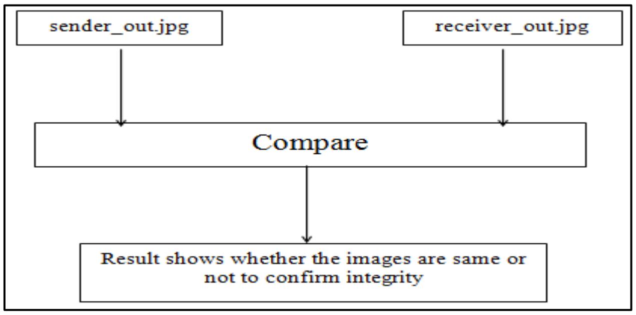
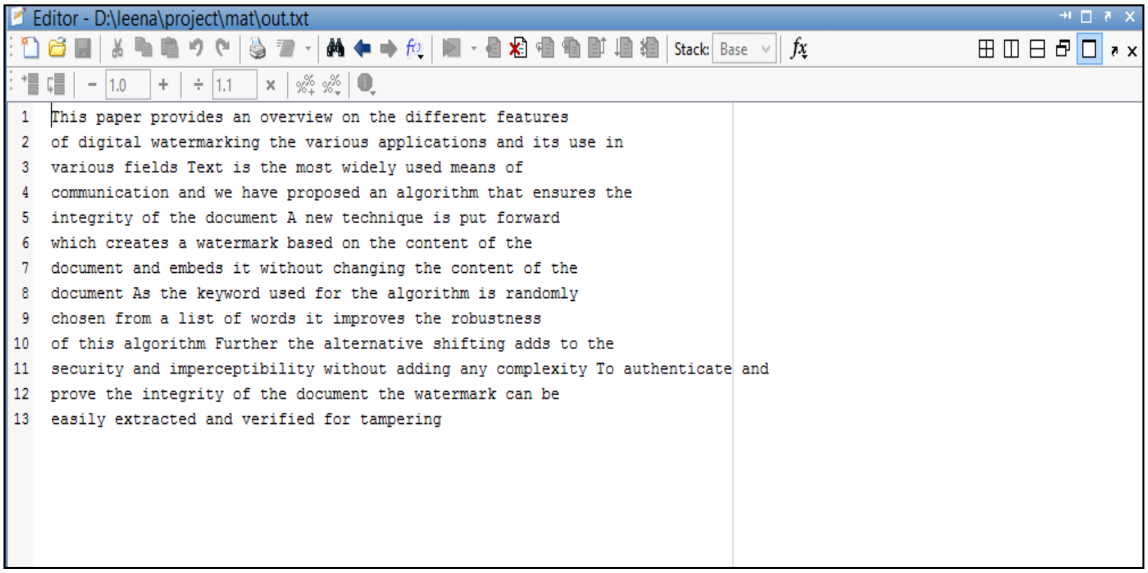

sender_out.jpg
receiver_out.jpg
Compare
Result shows whether the images are same or not to confirm integrity

Figure 4

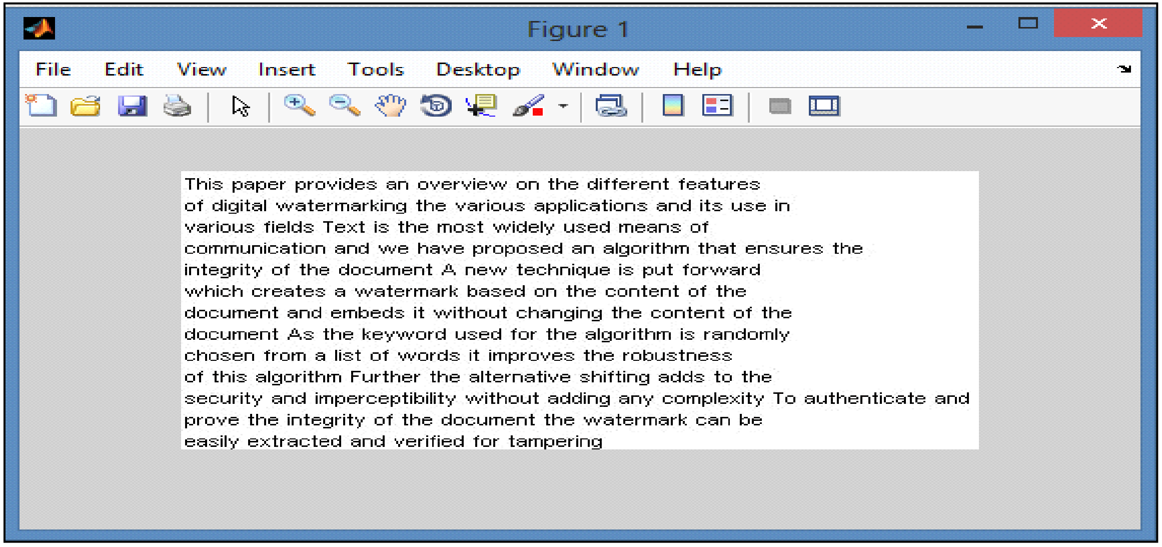
IV. EXPERIMENT AND RESULT

Mat lab simulations are performed on the text file. Figure 5(a) and Figure 5(b) Shows the text file and their equivalent image file. Now according to our work we applied some substitution algorithm and there after we get a encrypted form of that particular text file by Figure 6(a). Then again we convert the encrypted text file into image form and showing with Figure 6(b). This encrypted form of image would be used to send through network and if this image is altered or change by any intruder then we can easily check by our work. So this proposed work provides the feature of integrity in the text documents through image. Our algorithm depends upon the keywords which we have chosen from the text file. All the integrity would be given by these keywords only. The larger the frequency of keywords means the higher integrity algorithm provide.



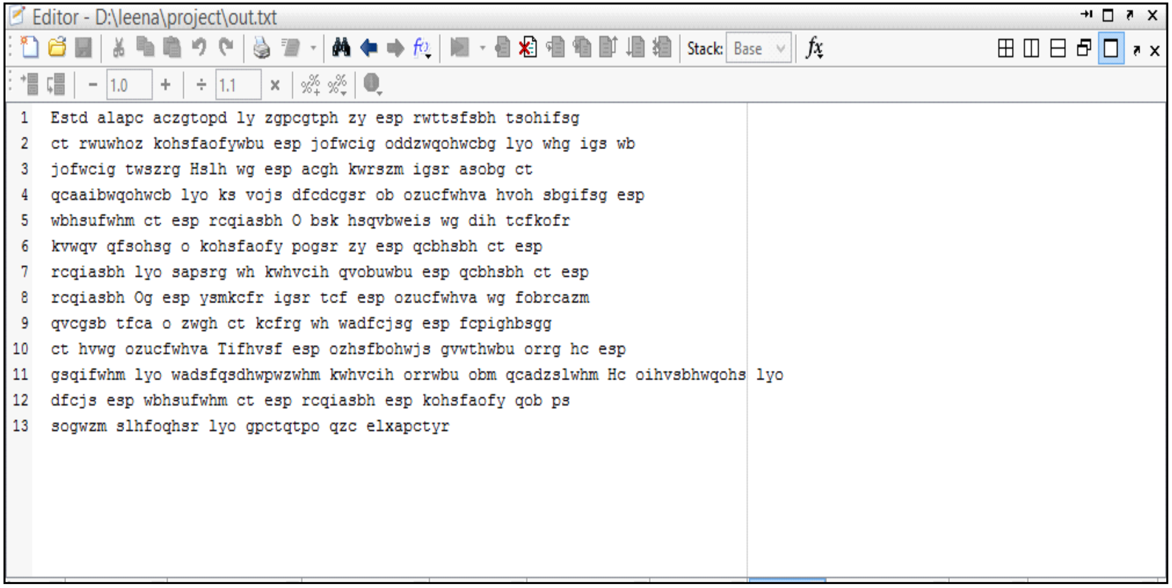
```
This paper provides an overview on the different features
of digital watermarking the various applications and its use in
various fields Text is the most widely used means of
communication and we have proposed an algorithm that ensures the
integrity of the document A new technique is put forward
which creates a watermark based on the content of the
document and embeds it without changing the content of the
document As the keyword used for the algorithm is randomly
chosen from a list of words it improves the robustness
of this algorithm Further the alternative shifting adds to the
security and imperceptibility without adding any complexity To authenticate and
prove the integrity of the document the watermark can be
easily extracted and verified for tampering
```

Figure 5(a)



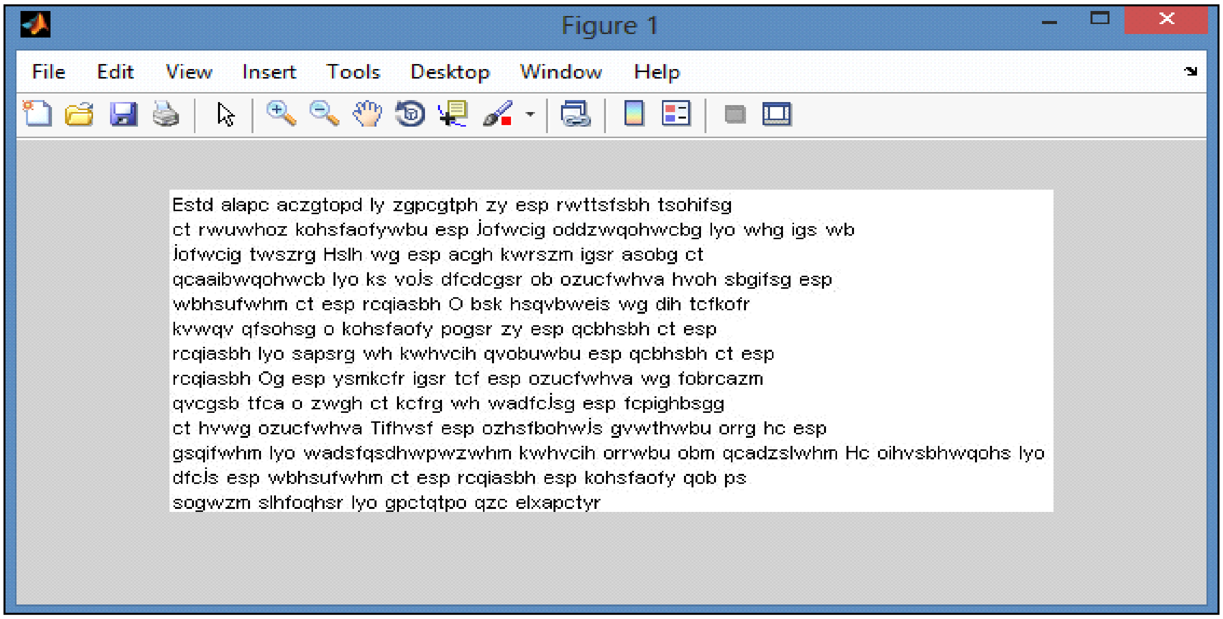
```
This paper provides an overview on the different features
of digital watermarking the various applications and its use in
various fields Text is the most widely used means of
communication and we have proposed an algorithm that ensures the
integrity of the document A new technique is put forward
which creates a watermark based on the content of the
document and embeds it without changing the content of the
document As the keyword used for the algorithm is randomly
chosen from a list of words it improves the robustness
of this algorithm Further the alternative shifting adds to the
security and imperceptibility without adding any complexity To authenticate and
prove the integrity of the document the watermark can be
easily extracted and verified for tampering
```

Figure 5(b)



```
Estd alapc aczgtopd ly zgpcgtph zy esp rwttsfsbh tsohifsg
ct rwuwhoz kohsfaofywbu esp jofwcig oddzwqohwcbg lyo whg igs wb
jofwcig twszrg Hslh wg esp acgh kwrszm igsr asobg ct
qcaaibwqohwcb lyo ks vojs dfcdcgsr ob ozucfwhva hvoh sbgifsg esp
wbhsufwhm ct esp rcqiasbh O bsk hsqvbweis wg dih tcfkofr
kvwqv qfsohsg o kohsfaofy pogsr zy esp qcbhsbh ct esp
rcqiasbh lyo sapsrg wh kwhvcih qvobuwbu esp qcbhsbh ct esp
rcqiasbh Og esp ysmkcfr igsr tcf esp ozucfwhva wg fobrcazm
qvcgsb tfca o zwgh ct kcfrg wh wadfcjsg esp fcpighbsgg
ct hvwg ozucfwhva Tifhvsf esp ozhsfbohwjs gvwthwbu orrg hc esp
gsqifwhm lyo wadsfqsdhwpwzwhm kwhvcih orrwbu obm qcadzslwhm Hc oihvsbhwqohs lyo
dfcjs esp wbhsufwhm ct esp rcqiasbh esp kohsfaofy qob ps
sogwzm slhfoqhsr lyo gpctqtpo qzc elxapctyr
```

Figure 6(a)



```
Estd alapc aczgtopd ly zgpcgtph zy esp rwttsfsbh tsohifsg
ct rwuwhoz kohsfaofywbu esp İofwcig oddzwqohwcbg lyo whg igs wb
İofwcig twszrg Hslh wg esp acgh kwrszm igsr asobg ct
qcaaibwqohwcb lyo ks voİs dfcdcgsr ob ozucfwhva hvoh sbgifsg esp
wbhsufwhm ct esp rcqiasbh O bsk hsqvbweis wg dih tcfkofr
kvwqv qfsohsg o kohsfaofy pogsr zy esp qcbhsbh ct esp
rcqiasbh lyo sapsrg wh kwhvcih qvobuwbu esp qcbhsbh ct esp
rcqiasbh Og esp ysmkcfr igsr tcf esp ozucfwhva wg fobrcazm
qvcgsb tfca o zwgh ct kcfrg wh wadfcİsg esp fcpighbsgg
ct hvwg ozucfwhva Tifhvsf esp ozhsfbohwİs gvwthwbu orrg hc esp
gsqifwhm lyo wadsfqsdhwpwzwhm kwhvcih orrwbu obm qcadzslwhm Hc oihvsbhwqohs lyo
dfcİs esp wbhsufwhm ct esp rcqiasbh esp kohsfaofy qob ps
sogwzm slhfoqhsr lyo gpctqtpo qzc elxapctyr
```

Figure 6(b)

V. CONCLUSION AND FUTURE SCOPE

In this paper a new method is proposed for text document integrity and confidentiality protection. In this method watermark is created based on the most frequent contents of the document and also encrypts the watermarked text to provide confidentiality. After that the comparison of sender side image and receiver side image is carried out. If there are any minor modifications in the text then it can be identified easily by comparing the images and some action can be taken to ensure the integrity of text. As advancement of our algorithm in future we can use AES (Advanced Encryption Standard) Algorithm for encryption of text document, which can provide a lot more benefits like-high security, and high efficiency.

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] 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, October 1995, pp. 1495-1504.

[7] N. F. Maxemchuk, "Electronic Document Distribution", AT&T Technical Journal, September 1994, pp. 73-80.

[8] N. F. Maxemchuk and S. Low, "Marking Text Documents," Proceedings of the IEEE International Conference on Image Processing, Washington, DC, Oct. 26-29, 1997, pg. 13-16.

[9] S. H. Low, N. F. Maxemchuk, and A. M. Lapone, "Document Identification for Copyright Protection Using Centroid Detection", IEEE Transactions on Communications, Mar. 1998, vol. 46, no.3, pg 372-381.

[10] 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, pg.1237-1245, Dec 2001.

[11] Young-Won Kim , Kyung-Ae Moon, and Il-Seok Oh, "A Text Watermarking Algorithm based on Word Classification and Inter-word Space Statistics" , Proceedings of the Seventh International Conference on Document Analysis and Recognition, IEEE, 2003

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

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

[14] Jaseena K.U., Anita John, "Text Watermarking using Combined Image and Text for Authentication and Protection", International Journal of Computer Applications ,Volume 20 No.4, April 2011