ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Text Watermarking using Sinusoidal Greyscale Variations of Font based on Alphabet Count

Jeebananda Panda[1] , Nishant Gupta[2] , Parag Saxena[3], Shubham Agrawal[4], Surabhi Jain[5], Asok Bhattacharyya[6]

Associate Professor, Department of ECE, Delhi Technological University, Delhi, India[1]

UG Student, Department of ECE, Delhi Technological University, Delhi, India[2]

UG Student, Department of ECE, Delhi Technological University, Delhi, India[3]

UG Student, Department of ECE, Delhi Technological University, Delhi, India[4]

UG Student, Department of ECE, Delhi Technological University, Delhi, India[5]

Professor, Department of ECE, Delhi Technological University, Delhi, India[6]

ABSTRACT: In this paper, an imperceptible, fragile text watermarking algorithm has been proposed. The font color of each alphabet in the text document is varied in the greyscale in accordance with a sine function. The amplitude for the same is generated using a hash function. The authenticity of the received document can be verified by comparing the actual color values of the letters in the document with the expected values generated by the sine function. This scheme is highly sensitive to various malicious text related tampering attacks hence preserving the integrity of the document. Unlike earlier methods, the proposed algorithm provides tamper detection while maintaining watermark invisibility. The attack analysis illustrates that the scheme is efficient and secure.

KEYWORDS: Greyscale; sine wave; text watermark; digital watermark; hash function; tamper detection

I. INTRODUCTION

The increased role of internet and networking techniques in modern communication has raised concerns over the security of digital information. Greater access to mobile devices like flash drives, memory cards, iPods etc. has enabled large volumes of text material to be transferred over these channels, exposing them to plagiarism, copyright violation, redistribution and other forms of malicious attacks. While extensive work has been done in the fields of image, video and audio watermarking; the research in the field of text watermarking is rather limited. The growth of e-commerce, e-business and digital libraries has augmented the need for efficient text watermarking techniques. Over the years the methods of encryption, steganography and watermarking have been used to solve these problems. Recently digital watermarking has emerged as a more advantageous method. This method is preferable over its counterparts as it maintains the comprehensibility of the documents while ensuring their authenticity and integrity. The proposed algorithm is based on the variation of the colour values of the font which follows a sine function. A hash function is implemented on a paragraph of the text to generate amplitude values for the sine function. The algorithm is sensitive to any form of tampering attack. The paper is organised into 6 sections. Section II examines the previous related work done in the area of text watermarking. Section III describes the proposed algorithm. Section IV illustrates the implementation of the same. Section V describes the experimental results and Section VI lists the conclusions.

II. RELATED WORK

A digital watermark may be described as an identification code that is permanently embedded in the document. The invisible watermarks are more secure at preserving the authenticity of the document. In the past many techniques have been proposed. These include text watermarking using text images, synonym based, pre-supposition based, syntactic

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

tree based, noun-verb based, word and sentence based, acronym based, typo error based methods etc. The text watermarking approaches are classified as follows.

A. *Image Based Approach:*
In image based approach towards text watermarking, the watermark is embedded in the text image. Brassil, et al. were the first to propose a few text watermarking methods utilizing text image [1]-[2]. Later, the performance of these methods were analysed by Maxemchuk, et al. [3]-[4]. Huang and Yan [5] proposed an algorithm based on an average inter-word distance in each line. Wiyada Yawai and Nualsawat Hiransakolwong showed how to use the intersection position of horizontal line, virtually run across text character skeleton line on a text image under the cross ratio applying, to be the marking point of zero watermarks [15].

B. *Syntactic Approach:*
The constituents of a sentence such as nouns, verbs, prepositions etc. determine the syntactic structure of the sentence which depends on language and its conventions. Applying syntactic transformations on text structure to embed watermark has also been one of the approaches towards text watermarking in the past. Mikhail J. Atallah, et al. first proposed the natural language watermarking scheme by using syntactic structure of text [6]-[7]. Hassan et al. performed morpho-syntactic alterations to the text to watermark it [8].

C. *Semantic Approach:*
Semantics of text like synonyms and antonyms are utilized to embed the watermark in text. Atallah et al. were the first to propose the semantic watermarking schemes in the year 2000[9]. Later, the synonym substitution method [10] was proposed. A noun-verb based technique for text watermarking was also proposed [11] 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 [12]. Algorithms were developed to watermark the Text using the linguistic semantic phenomena of prepositions [13]. The algorithm based on Text Meaning Representation (TMR) strings has also been proposed [14].

D. *Zero-Watermarking approaches:*
In zero watermarking approach, the host text document is not altered to embed the watermark; rather the characteristics of the text are utilized to generate a watermark. This watermark pattern is later matched using a pattern matching procedure with the pattern generated by tampered document to identify any tampering [16]. Jalil Z. et al. proposed a zero text watermarking algorithm based on occurrence frequency of non-vowel ASCII characters. The embedding algorithm makes use of frequency non-vowel ASCII characters and words to generate a specialized author key [17]. Zunera Jalil et al. developed an algorithm which utilizes a keyword from the text (selected based on author choice) to generate a watermark based on the length of preceding and next word length, to and from the keyword occurrences in text [19].

III. PROPOSED ALGORITHM

The aforementioned techniques have the drawback of attack specificity and often become unreliable when multiple attacks are performed. They also are not applicable to all types of text documents under random tampering attacks and are not specifically designed to solve tamper detection problem. The proposed algorithm aims to ensure authenticity and integrity over a wide variety of tamper attacks while also identification of location of tampering.

The algorithm uses variations in the font colour of the document. This watermarking technique exploits the fact that minor changes in colour are imperceptible to the human eye. The colour of every alphabet in the document varies sinusoidal within the greyscale. Xianmin Wei, earlier, proposed a sine wave based watermarking scheme which relied on word count, was limited to WORD documents only [18].

In the proposed algorithm, the count of each alphabet present in the raw text determines the parameters of the color variations on the entire text. It is, therefore, independent of any format of the document i.e. doc, docx, pdf etc. because the algorithm runs on raw text. The same algorithm may easily be extended to other languages with only slight

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

modifications involving the number of alphabets. The proposed technique of digital text watermarking can be adopted for IP protection of any text document.

A. *Watermark embedding process:*
The amplitudes of the sine waves are generated by applying the SHA-1 hash to any chosen paragraph of the document. This paragraph will be pre-decided between the sender and the receiver. The first 26 hexadecimal bits of the 40-bit hash so generated are used as amplitudes. The first 26 hexadecimal bits are used from the 40 bits generated as the amplitudes. Since these are hexadecimal values, the amplitude corresponding to sine wave for each alphabet hence vary between 1 and 16. The code of the proposed algorithm can be written in any high level language. The number of occurrences of each alphabet is recorded and this is used to calculate the sine function's argument. The argument is such that the sine wave completes one cycle over the total occurrences of one alphabet. This is true for all 26 alphabets, irrespective of case. The watermark is embedded such that the colour of the text varies in the greyscale from 85 to 100 on a scale of 0-100. This ensures that the changes in the intensity of the black colour remain imperceptible. The text is read alphabet by alphabet and its colour property is changed according to the sine wave of the corresponding alphabet. In the case of special characters (such as exclamation marks, commas and full stops), the output colour is the same as the preceding alphabet. The document with the embedded watermark is converted to a pdf.

B. *Tamper Detection:*
At the receiver end, the raw text is extracted from the received pdf document. The embedding algorithm is again run on the raw text to generate a new watermarked document. This generated document is compared against the received document. Any mismatch reported will indicate that the received document has been tampered with. The two documents can be compared by converting them to images and subtracting them using some software like MATLAB. If the document has not been tampered with, a resultant black image will be obtained.

IV. PROPOSED ALGORITHM

EMBEDDING ALGORITHM

1. Read the raw text.
2. Run the SHA-1 over a pre-decided paragraph of the document to generate a hash.
3. 26 hexadecimal values out of the generated 40-hexadecimal hash are used as amplitudes for the sine waves. Let these values be $\{A_1, A^2.....A_{26}\}$.
4. The number of occurrences of each alphabet is recorded which then used for forming the argument for the sine function. Let these values be $\{N_1, N_2, ...N_{26}\}$.
5. Watermark is then embedded by changing the colour according to the sine wave of the corresponding alphabet is calculated in the following manner.
6. Let the color value of the m^{th} occurrence of the n^{th} alphabet be denoted by S_{nm} Then S_{nm} is given by the expression

$$\mathbf{S_{mn} = A_n * \sin(2\pi * m/N_n)}$$

Suppose, in a document, the letter "B" occurs a total of 5 times. Let the amplitude corresponding to it, as calculated by the hash function be 7. Then the greyscale value of the 3^{rd} (m=3) occurrence of "B" in the watermarked document will be calculated as:

$$S_{3,5} = 7 * \sin(2\pi * 3/5) = -4.1$$

Hence, the grey-scale value will be 93-4 =89. This indicates the integer colour value that "B" would attain on a grey scale of range 0-100.

DETECTION ALGORITHM

1. Extract the raw text from the received document.
2. Run the embedding algorithm again on it.
3. Convert the new watermarked pdf and the received pdf document to any image format.
4. Subtract the above two images (I_{m1} and I_{m2}) in order to compare them.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

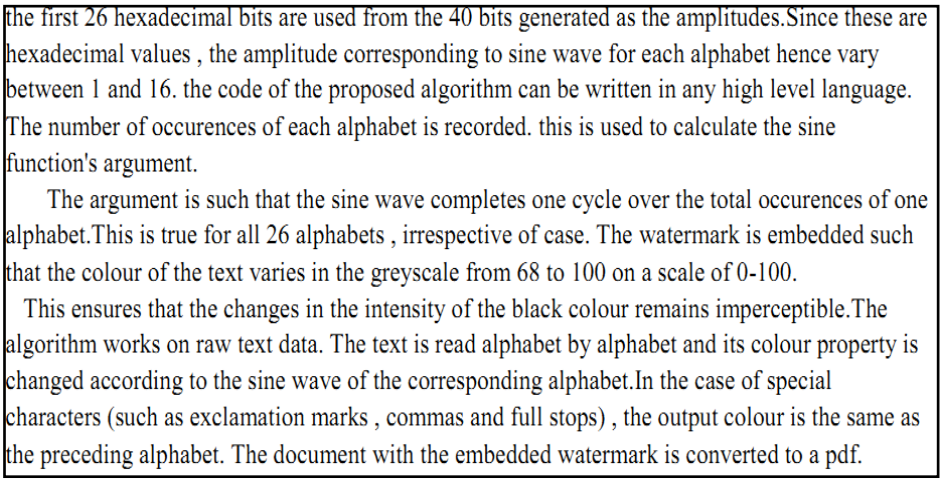
(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

5. The subtracted image is given by $\mathbf{I_{msub}= abs(I_{m1}-I_{m2})}$ and convert I_{msub} to greyscale format.
7. Plot the histogram.

V. EXPERIMENTAL RESULTS

The following figures demonstrate the experimental observations. **Figure 1** shows the raw text. The changes in the text document after embedding the algorithm are imperceptible, as seen in **Figure 2**. **Figure 3** shows the tampered document. The subsequent images show the detection process. **Figure 4** plots the histogram on comparing the un-tampered watermarked text obtained at the receiver's end against the text obtained by running the algorithm again on it. **Figure 5** illustrates the output on subtracting the images of the tampered watermarked text and the document obtained on running the algorithm again on the received tampered text. The histogram generated for **Figure 5** is shown in **Figure 6**. It can be seen from these results that if the document has been tampered with, then a resultant black image with grey patches is observed on subtraction. Otherwise, the resultant image will be completely black.


the first 26 hexadecimal bits are used from the 40 bits generated as the amplitudes.Since these are hexadecimal values , the amplitude corresponding to sine wave for each alphabet hence vary between 1 and 16. the code of the proposed algorithm can be written in any high level language. The number of occurences of each alphabet is recorded. this is used to calculate the sine function's argument.

The argument is such that the sine wave completes one cycle over the total occurences of one alphabet.This is true for all 26 alphabets , irrespective of case. The watermark is embedded such that the colour of the text varies in the greyscale from 68 to 100 on a scale of 0-100.

This ensures that the changes in the intensity of the black colour remains imperceptible.The algorithm works on raw text data. The text is read alphabet by alphabet and its colour property is changed according to the sine wave of the corresponding alphabet.In the case of special characters (such as exclamation marks , commas and full stops) , the output colour is the same as the preceding alphabet. The document with the embedded watermark is converted to a pdf.

Figure 1:Original Raw Text

The raw text shown in **Figure 1** when hashed with the SHA-1 algorithm (using a hash generator) generates the following 40-hexadecimal value:

"01c12fc24deaa65e4452335485854481603cd4bd"

The following table each alphabet along with its corresponding number of occurrences and the amplitudes calculated on the basis of the hash function. These values are then used as parameters {N_n and A_n} of the sine function to generate the instantaneous color value of the alphabets as explained in section IV. The total count and amplitudes for each alphabet are given in **Table 1**.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Alphabet	Count	Amplitude
A	81	1
B	18	2
C	49	13
D	30	2
E	133	3
F	18	16
G	15	13
H	59	3
I	49	5
J	0	14
K	5	15
L	37	11
M	24	11
N	45	7
O	56	6
P	25	15
Q	0	5
R	52	5
S	59	6
T	93	3
U	26	4
V	10	4
W	10	6
X	6	5
Y	7	9
Z	0	6

Table1: Alphabets with corresponding counts and calculated amplitudes

the first 26 hexadecimal bits are used from the 40 bits generated as the amplitudes.Since these are hexadecimal values , the amplitude corresponding to sine wave for each alphabet hence vary between 1 and 16. the code of the proposed algorithm can be written in any high level language. The number of occurences of each alphabet is recorded. this is used to calculate the sine function's argument. The argument is such that the sine wave completes one cycle over the total occurences of one alphabet.This is true for all 26 alphabets , irrespective of case. The watermark is embedded such that the colour of the text varies in the greyscale from 68 to 100 on a scale of 0-100. This ensures that the changes in the intensity of the black colour remains imperceptible.The algorithm works on raw text data. The text is read alphabet by alphabet and its colour property is changed according to the sine wave of the corresponding alphabet.In the case of special characters (such as exclamation marks , commas and full stops) , the output colour is the same as the preceding alphabet. The document with the embedded watermark is converted to a pdf.

Figure 2: Original Document after embedding Watermark

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Figure 3 shows the tampered document. The subsequent images show the detection process.

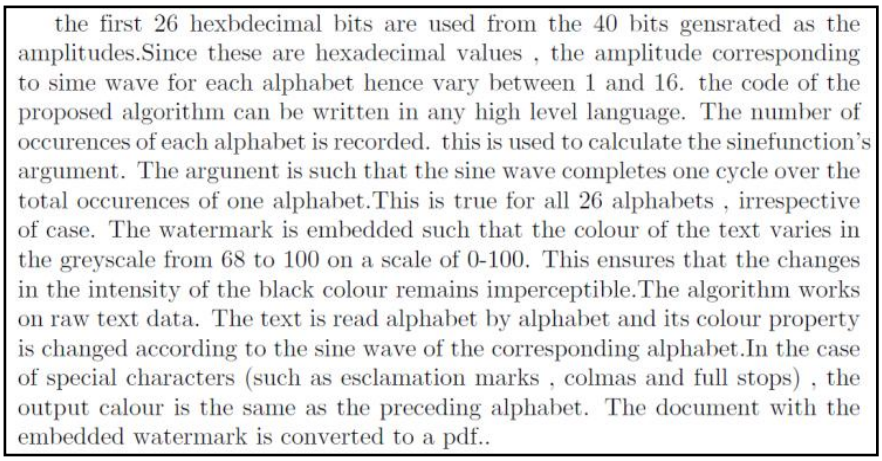

the first 26 hexbdecimal bits are used from the 40 bits gensrated as the amplitudes.Since these are hexadecimal values , the amplitude corresponding to sime wave for each alphabet hence vary between 1 and 16. the code of the proposed algorithm can be written in any high level language. The number of occurences of each alphabet is recorded. this is used to calculate the sinefunction's argument. The argunent is such that the sine wave completes one cycle over the total occurences of one alphabet.This is true for all 26 alphabets , irrespective of case. The watermark is embedded such that the colour of the text varies in the greyscale from 68 to 100 on a scale of 0-100. This ensures that the changes in the intensity of the black colour remains imperceptible.The algorithm works on raw text data. The text is read alphabet by alphabet and its colour property is changed according to the sine wave of the corresponding alphabet.In the case of special characters (such as esclamation marks , colmas and full stops) , the output calour is the same as the preceding alphabet. The document with the embedded watermark is converted to a pdf..

Figure 3: Tampered Document after running the Extraction Algorithm

When the document is untampered, the received pdf and the pdf generated after extraction are identical. Hence on subtraction, the resultant image is completely black. This causes the histogram to be concentrated around 0 as shown in **Figure 4**.

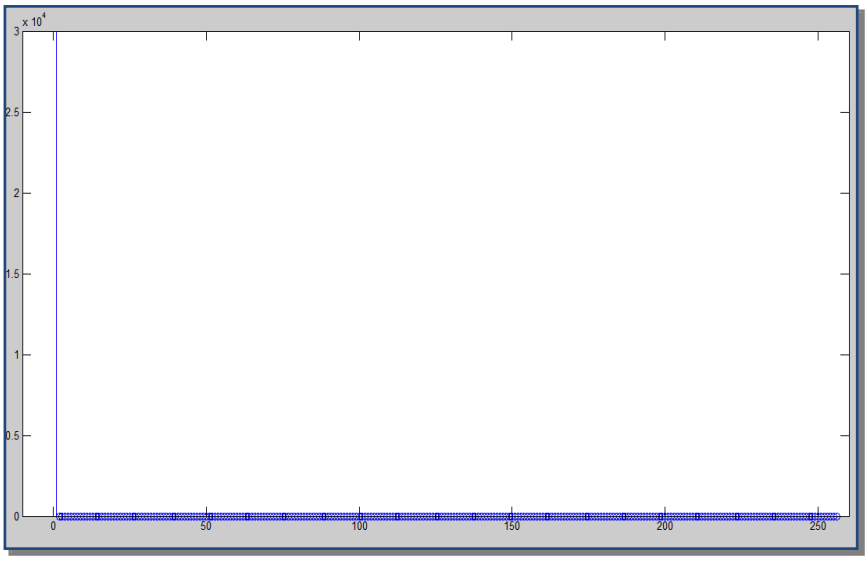


Figure 4: Histogram on comparing untampered watermarked document with the text obtained on running the algorithm again on it

Tampering of the document causes discrepancies in the parameters of the sine wave. Thus the resultant image has grey patches, shown in **Figure 5**.

.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

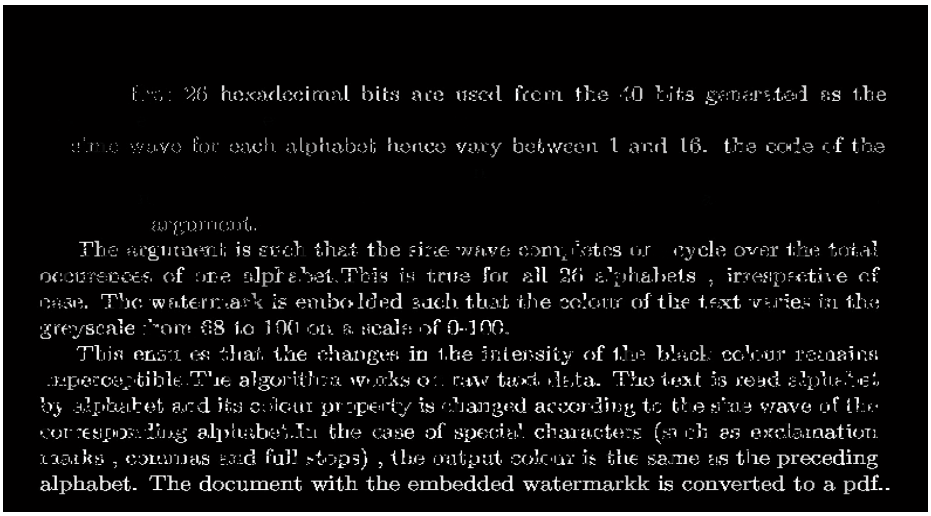


Figure 5: Subtraction of images of tampered text with that obtained in Figure 3

On plotting the histogram of this subtracted image, it is observed that it is no longer concentrated along zero but rather distributed along the whole greyscale as in **Figure 6**

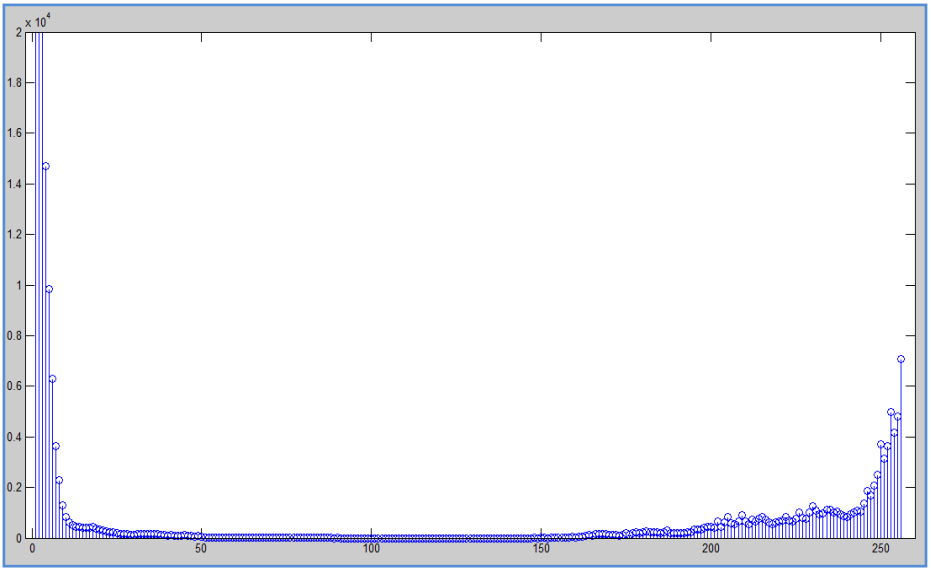


Figure 6: Histogram of Figure 5

VI. CONCLUSION AND FUTURE WORK

The algorithm implemented here is a novel watermarking scheme that is imperceptible and preserves the authenticity and integrity of the document. Using alphabet count as a variation parameter allows for an unprecedented amount of sensitivity of even small changes which is the strength of the proposed technique. The earlier works on watermarking for text authentication are not reliable in the case of random tampering attacks, especially when the amount of tampering is low. The algorithm hence is versatile, highly imperceptible and fragile. Future works may include extending the algorithm to watermark coloured documents and to detect changes in formatting, like indentation, of the document. With small changes, the algorithm may also be extended to other languages too.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

REFERENCES

[1] J. Brassil, S. Low, N.Maxemchuk, and L. O'Gorman, "Hiding Information in Document Images", Proceedings of the 29th Annual Conference on Information Sciences and Systems, Johns HopkinsUniversity, pp 482-489, March 1995.
[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.
[3] 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, pp 561-572, May 1998.
[4] N. F. Maxemchuk, "Electronic Document Distribution," AT&T Technical Journal, pp. 73-80. 6, September 1994.
[5] D. Huang and H. Yan, "Inter word 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, December, 2001.
[6] 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/SIG SAC New Security Paradigms Workshop, September, Cork, Ireland, pp. 51–65, 2000.
[7] M. J. Atallah, V. Raskin, M. C. Crogan, C. F. Hempelmann, F. Kerschbaum, D. Mohamed, and S.Naik, "Natural language watermarking: Design, analysis, and a proof-of-concept implementation", Proceedings of the Fourth Information
[8] PA. Hasan M. Meral, et al., "Syntactic tools for text watermarking", Hiding Workshop, vol. LNCS 2137, 9th SPIE Electronic Imaging Conf. 6505, Pittsburgh, 25-27th April 20011: Security, Steganography, and Watermarking of Multimedia Contents, San Jose , Jan. 2007.
[9] M. 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.
[10] U. Topkara, M. Topkara, M. J. Atallah, "The Hiding Virtues of Ambiguity: Quantifiably Resilient Watermarking of Natural Language Text through Synonym Substitutions".In Proceedings of ACM Multimedia and Security Conference, Geneva, 2006.
[11] Xingming Sun, Alex Jessey Asiimwe, "Noun-Verb Based Technique of Text Watermarking Using Recursive Decent Semantic Net Parsers", Lecture Notes in Computer Science (LNCS) 3612: 958-961, Springer Press, August 2005.
[12] Topkara, M., Topraka, U., Atallah, M.J., "Information hiding through errors: a confusing approach" In: Delp III, E.J., Wong, P.W. (Eds.), Security, Steganography, and watermarking of Multimedia Contents IX. Proceedings of SPIE-IS&T Electronic Imaging SPIE6505. pp. 65050V-1–65050V-12, 2007.
[13] B. Macq and O. Vybornova, "A method of text watermarking using presuppositions," in Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents, January 2007.
[14] Peng Lu et al., "An optimized natural language watermarking algorithm based on TMR", on proceedings of 9th International Conference for Young Computer Scientists, 2009.
[15] Wiyada Yawai and Nualsawat Hiransakolwong, "Increase the Hiding-bit Capacity and Strength for Text Watermarking with the Line Intersection on Text Image" IEEE ICCM, vol. 1, pp. 427-433 , 2012.
[16] ZuneraJalil, Anwar M. Mirza, and Hajira Jabeen, "Word Length Based Zero-watermarking Algorithm for Tamper Detection in Text Documents," 2nd International Conference on Computer Engineering and Technology, vol. 6, pp. 378-382, 2010.
[17] 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, pp. 592-596 , 2010
[18] Xianmin Wei "Sine-wave-based Text Watermark for WORD Document", IEEE International Conference on Computer and Information Application, vol. 10, pp. 99-102,2010.
[19] Zunera Jalil, Anwar M. 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, pp. 212-217, February 2010.

BIOGRAPHY

Jeebananda Panda
Designation: Associate Professor
Qualification: ME
Specilization: Applied Electronics

Nishant Gupta
Designation:Student
Qualification: B.Tech
Department: Electronics and Communication Engineering

Parag Saxena
Designation:Student
Qualification: B.Tech
Department: Electronics and Communication Engineering

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Shubham Agrawal
Designation:Student
Qualification: B.Tech
Department: Electronics and Communication Engineering

Surabhi Jain
Designation:Student
Qualification: B.Tech
Department: Electronics and Communication Engineering

Asok Bhattacharyya
Designation:Professor
Qualification: PhD
Department: Electronics and Communication Engineering