Performance Evaluation of Natural Language Text Watermarking using Encryption Techniques

Manmeet Kaur
Department Of Computer Science & Engineering
Rayat and Bahra Institute of Engineering & Bio Technology
Mohali, India

Kamna Mahajan
Department Of Computer Science & Engineering
Rayat and Bahra Institute of Engineering & Bio Technology
Mohali, India

ABSTRACT

Advances in communication technologies have made it easier to distribute and communicate information effectively. The increasing use of Internet has caused the information to be paperless and all the working to be electronic as compared to the conventional paper distribution. As the information is available on internet, it is liable to many kinds of threats like illegal copying, distribution, tampering, authentication etc. Till now, the electronic information can be secured by using different techniques like steganography, cryptography, watermarking. In text watermarking various techniques are implemented for English, Chinese, Turkish and Arabic language text using different methods. This paper includes an improved text watermarking technique for English language text documents. This paper proposes a technique which uses natural language components and UMARAM encryption technique. This study has focussed on grammatical rules like conjunctions, pronouns and modal verbs to generate encrypted watermark message. The performance of proposed technique is compared with AES encryption technique. It has been concluded that UMARAM algorithm is robust against content modifications and at the same time, is capable of detecting tampering attacks. The resulted technique is tested over various text documents to check the effectiveness of the algorithm.

Keywords

Text watermarking, Copyright Protection, Security, Encryption, Tampering, UMARAM, AES, robustness.

1. INTRODUCTION

The dawn of internet, e-commerce and other efficient communication technologies have resulted in many new ways for creation and delivery of digital contents. Besides, making the access to information easier within a very short span of time, the digital contents face the threats of copyright violations, digital counterfeiting, privacy, and plagiarism issues. Digital contents are composed of text, image, audio and video. The ease of dissemination and reproduction of digital contents has made it difficult to protect its copyrights[4]. Authentication and copyright protection of digital images, audio, and video has been given due thought by the researchers in past but the amount of work to protect the text is very inadequate.

Now a day's, text is the most important medium travelling 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. These text documents sometimes have various threats. The threats of electronic publishing like illegal copying of important information, redistribution of copyrighted text documents, tampering, forgery, plagiarism, theft and authentication must be seriously and specifically addressed. To overcome these threats various solutions such as authenticity, integrity, confidentiality, and copyright protection are urgently required. Digital watermarking is one of the solutions which can watermark the digital contents and can be used to claim ownership later. Digital watermarking methods are used to identify the copyright owner of origin (s) of content that may be an image, a simple text, sound, video or a combination of all[1].

The process of embedding and extracting a digital watermark to and from a digital text document which uniquely identifies the original copyright owner of that text is called **Digital Text Watermarking.** Text watermarking is done using the steps as shown in Figure 1.

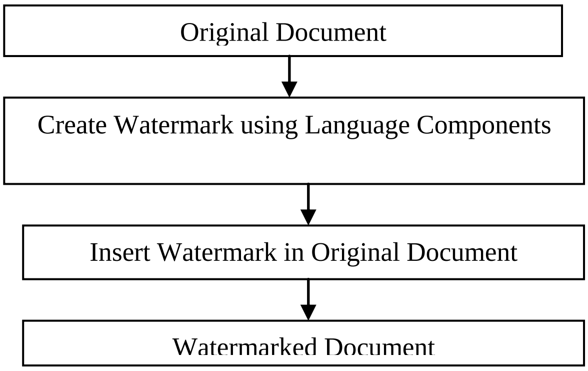

Original Document
Create Watermark using Language Components
Insert Watermark in Original Document
Watermarked Document

Figure1: Watermarking Process

This paper evaluates the performance of a novel text watermarking technique on various parameters.

2. NATURAL LANGUAGE WATERMARKING

Natural language watermarking means embedding a watermark into a text document using the natural components of language as the carrier in such a way that the modifications are imperceptible to the reader and the information is robust against the possible attacks. The structure of the sentences can be used in order to insert watermark in the text document.

The text watermarking algorithms developed in the past for plain text embedded a watermark in the original text document which resulted in degradation of text quality and meaning. . Previous work on the text watermarking can be classified into the following categories; an image based approach, a syntactic approach, a semantic approach and a structural approach.

In this paper, an approach is presented in which the original text document is not altered to embed watermark, instead the characteristics of the text are used to generate a watermark. This watermark is fragile in nature and is used to authenticate the text documents. This paper proposes a technique which uses natural language (English) grammatical rules such as conjunctions, pronouns, nouns and modal verbs to generate a watermark. The user has to choose any two of the four rules. The count and actual occurrences of the chosen rules along with author Id are concatenated. The combined key is further encrypted by UMARAM encryption algorithm [23] to form the encrypted key. Watermark is registered with the Certifying Authority (CA) and is used in the extraction algorithm to authenticate text document.

The performance of proposed technique with UMARAM encryption is compared and evaluated with AES encryption technique using delay, robustness and tampering attack as performance measures.

In this approach watermark is generated using following steps:

1. Select a text document

First, we select a text document on which watermarking is to be done. In this approach we have used text documents of extension .txt.

2. Key Generation

Watermark is generated using the English language components like noun, pronoun, model verbs and conjunction. The user first selects combination of any two from the given components. Then count the total number of noun, pronoun, conjunction or model verbs present in the document. Author name is entered and thus author name is concatenated with the count calculated above. Key is made up of combination of count of language components concatenated with author name. This final combination is further encrypted using UMARAM encryption algorithm before adding it to the text document.

Watermark Key= Concatenation of Author name + Count of combination of any two (noun, pronoun, model verb or conjunction)

3. Embedding

The algorithm which embeds the watermark in the text is called embedding algorithm. The watermark embedding algorithm requires original text document as input and the key generated above.

4. Extraction Algorithm

The algorithm which extracts the watermark from the text is called extraction algorithm. The proposed extraction algorithm takes the plain text and keyword as input. The text may be attacked or un-attacked. The watermark is generated from the text by the extraction algorithm and is then, compared with the original watermark registered with the CA to check it for its authenticity.

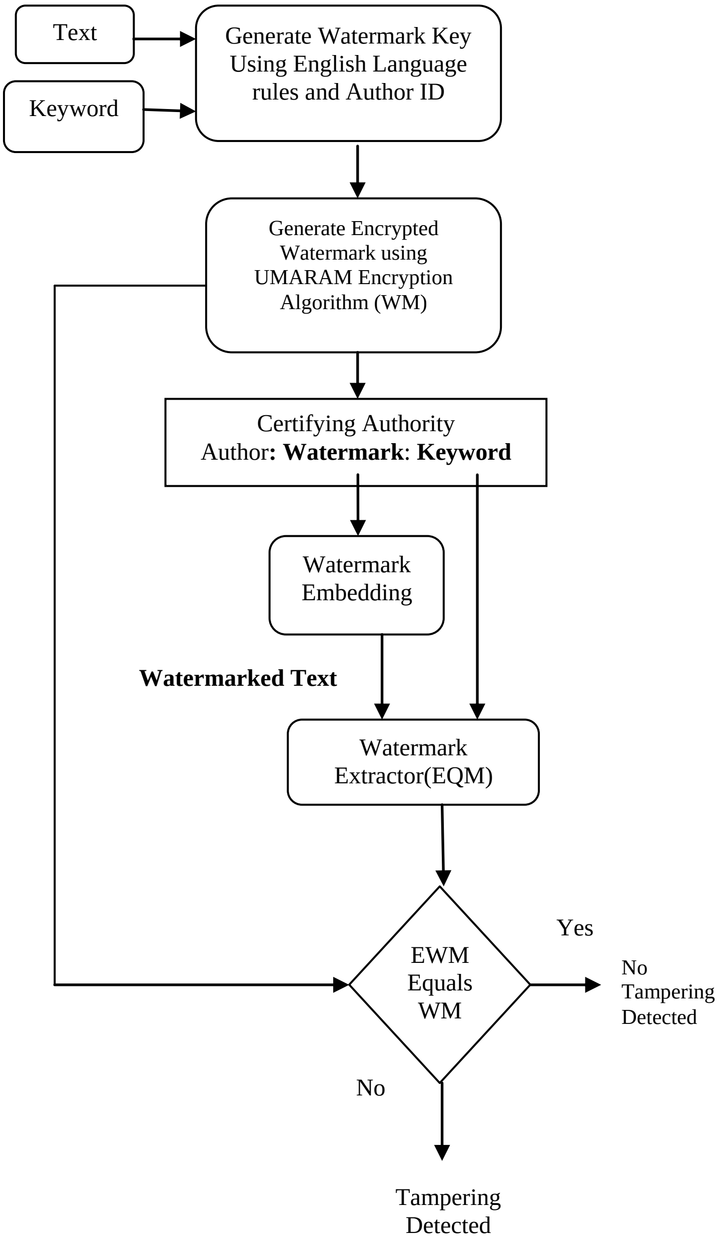

Text
Keyword
Generate Watermark Key Using English Language rules and Author ID
Generate Encrypted Watermark using UMARAM Encryption Algorithm (WM)
Certifying Authority
Author**: Watermark**: **Keyword**
Watermark Embedding
Watermarked Text
Watermark Extractor(EQM)
EWM Equals WM
Yes
No Tampering Detected
No
Tampering Detected

Figure2: Overview of Watermark Generation and Extraction Process

3. ALGORITHMS

3.1 UMARAM Encryption Algorithm:

This new symmetric key encryption algorithm prevents the outside attacks and avoids fixed key exchange between the users and reduces the time taken for encryption and decryption. It operates at higher data rate in comparison with DES, triple DES, AES and RC6 algorithms[22].

The new algorithm adds some difficulties to the attackers to discover the key. The difficulties are:

- The longer key size, 512 bits, compared with DES, TDES, AES-256 and RC6.
- The key updating with each packet.

This algorithm uses a key size of 512 bits to encrypt a plaintext of 512 bits in 16 rounds. A series of transformations have been used depending on S-Box, different shift processes, XOR-gate, and AND gate. Each slide in the S-box is described by the following equation:

$S^{i}|_{X*Y} = S|_{X*Y*i}$

Where i =1, 2, ….., 16 and i is defined as the round number used in key generation, encryption and Decryption process[22].

- **Key generation in UMARAM:**

The key generation generates 16 keys during 16 rounds. One key of them is used in one round of the encryption and decryption process. Figure 3 explains the procedure of the key generation.

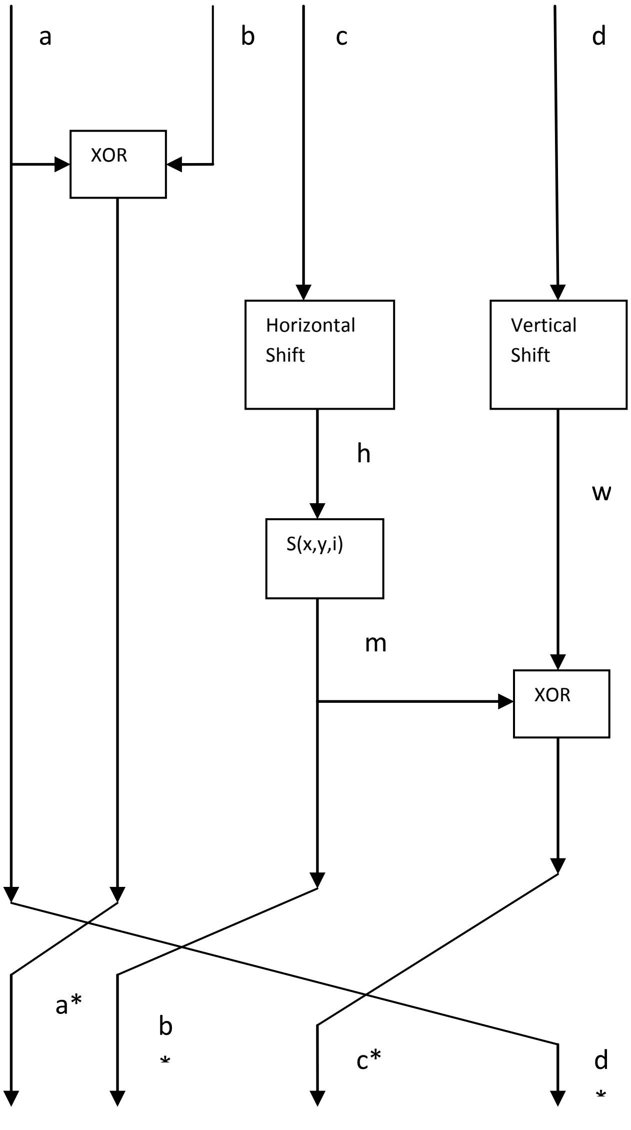


Figure 3: Key generation process in one round[22]

- **Encryption Process:**

The encryption process in this algorithm is used to encrypt the plain text of 512 bits by a key of size 512 bits in 16 rounds. Series of transformations are applied on the plaintext in each round as shown in figure 4 to obtain the cipher text.

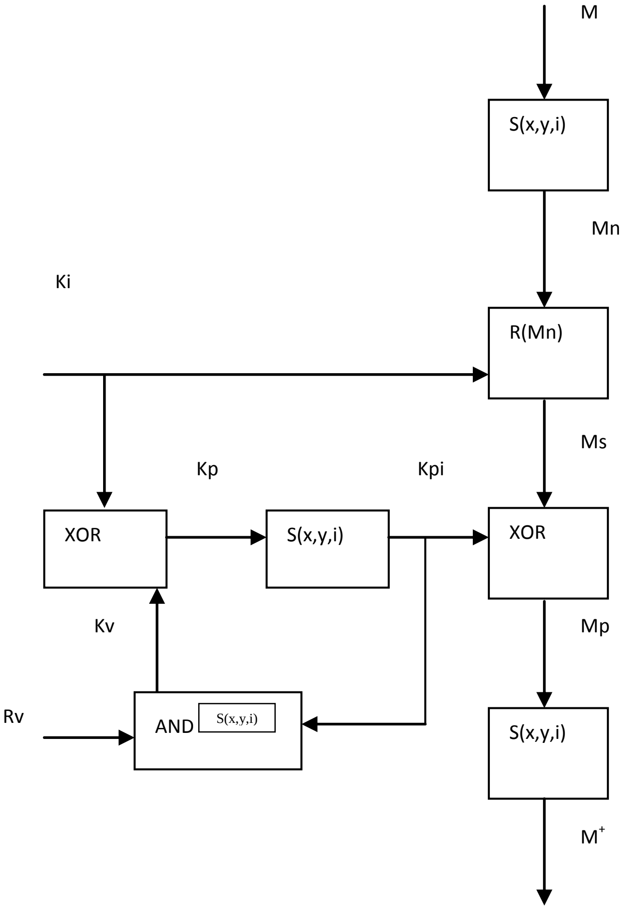


Figure 4: Encryption Process in each round[22]

- **Decryption Process:**

The decryption process is same as the encryption process but the direction of the encryption process is reversed as shown in figure 5.

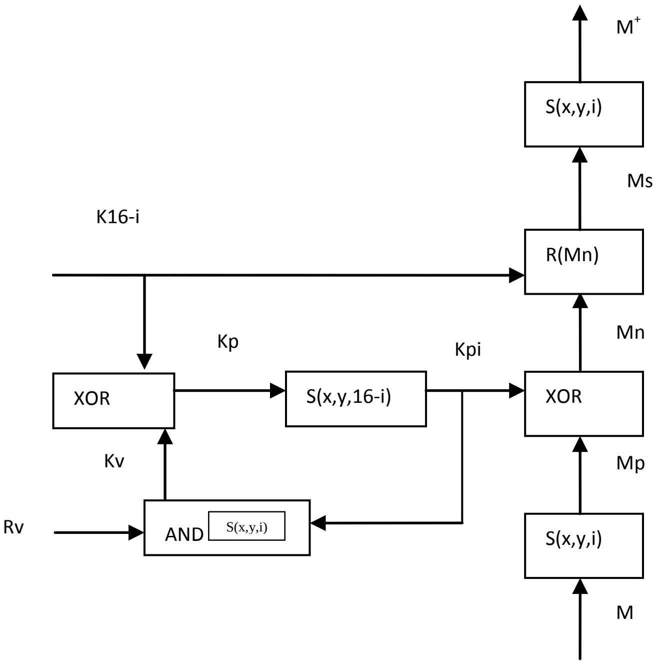


Figure 5: Decryption Process in each round [22]

4. AES ENCRYPTION ALGORITHM

AES is a block cipher with a block length of 128 bits. AES allows for three different key lengths: 128, 192, or 256 bits. The key length taken for our work is 128 bits. Encryption consists of 10 rounds of processing for 128-bit keys, 12 rounds for 192-bit keys, and 14 rounds for 256-bit keys. Except for the last round in each case, all other rounds are identical. The overall structure of AES encryption/ decryption is shown in Figure 6.

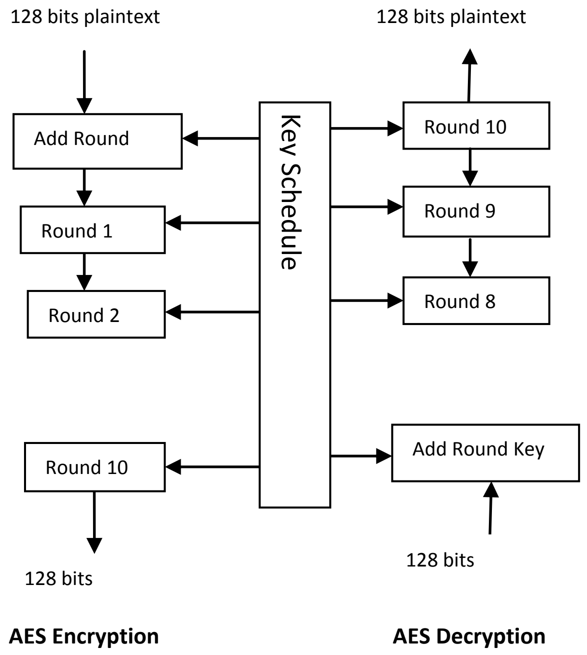
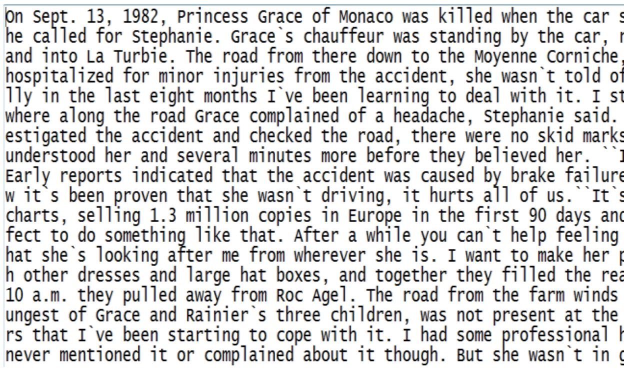


Figure 6: Structure of AES for 128 bits encryption key

5. RESULTS AND DISCUSSIONS

This section discusses the results obtained by using the above mentioned approach. MATLAB version 7.10.0 has been used. Text documents with various word counts (300,1000, 2100, 4200, 5100 and 7100) have been taken. We apply both the encryption techniques (UMARAM and AES) on the all text documents and compare their performances.

Figure 5 below shows the original plain text document having word count of 2100.



```
On Sept. 13, 1982, Princess Grace of Monaco was killed when the car s
he called for Stephanie. Grace`s chauffeur was standing by the car, r
and into La Turbie. The road from there down to the Moyenne Corniche,
hospitalized for minor injuries from the accident, she wasn`t told of
lly in the last eight months I`ve been learning to deal with it. I st
where along the road Grace complained of a headache, Stephanie said.
estigated the accident and checked the road, there were no skid marks
understood her and several minutes more before they believed her. ``I
Early reports indicated that the accident was caused by brake failure
w it`s been proven that she wasn`t driving, it hurts all of us.``It`s
charts, selling 1.3 million copies in Europe in the first 90 days and
fect to do something like that. After a while you can`t help feeling
hat she`s looking after me from wherever she is. I want to make her p
h other dresses and large hat boxes, and together they filled the rea
10 a.m. they pulled away from Roc Agel. The road from the farm winds
ungest of Grace and Rainier`s three children, was not present at the
rs that I`ve been starting to cope with it. I had some professional h
never mentioned it or complained about it though. But she wasn`t in g
```

Figure 7: Original Text Document

A graphical user interface (GUI) has been created in which we select the text document to be watermarked, two english language components and author ID along with encryption algorithm as shown in figure8.

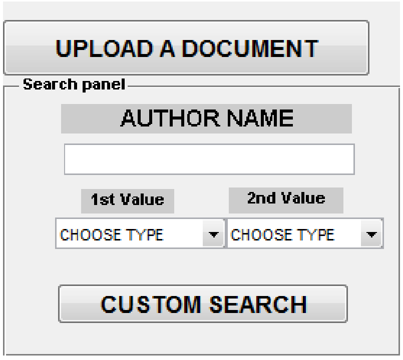
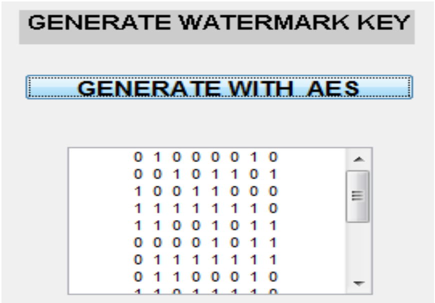


Figure 8: Components for watermark key generation

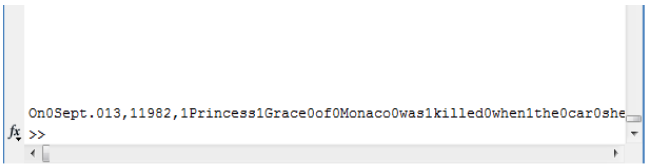
The figure below shows the encrypted key (128-bits) generated for watermarking using AES encryption algorithm.



```
0 1 0 0 0 0 1 0
0 0 1 0 1 1 0 1
1 0 0 1 1 0 0 0
1 1 1 1 1 1 1 0
1 1 0 0 1 0 1 1
0 0 0 0 1 0 1 1
0 1 1 1 1 1 1 1
0 1 1 0 0 0 1 0
```

Figure 9: Watermark key generated using AES

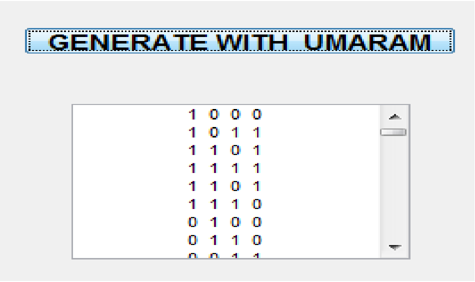
After generating the key, it is embedded in the selected text document. Figure10 below shows the visible key present in the text document at the backend and the embedded key is made invisible at the front end which is visible to the user as shown in figure 13 below.



```
On0Sept.013,11982,1Princess1Grace0of0Monaco0was1killed0when1the0car0she
>>
```

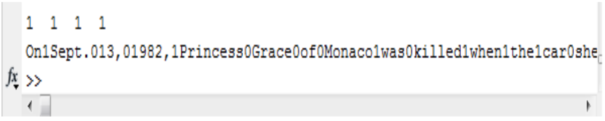
Figure 10: Key embedded at backend using AES

The figure below shows the encrypted key (512 bits) using UMARAM encryption algorithm and the generated key is embedded in the text document as shown in figure 12.



```
1 0 0 0
1 0 1 1
1 1 0 1
1 1 1 1
1 1 0 1
1 1 1 0
0 1 0 0
0 1 1 0
```

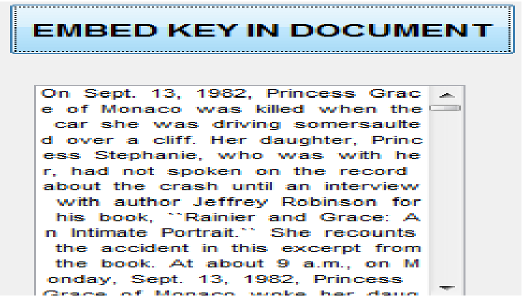
Figure 11: Watermark key generated using UMARAM



```
1  1  1  1
On1Sept.013,01982,1Princess0Grace0of0Monaco1was0killed1when1the1car0she
>>
```

Figure 12: Key embedded at backend using UMARAM

The front end document with watermark is shown in the figure below. This document appears same as the plain text document and the user will not be able to recognize that a watermark has been embedded in it.



On Sept. 13, 1982, Princess Grac
e of Monaco was killed when the
car she was driving somersaulte
d over a cliff. Her daughter, Princ
ess Stephanie, who was with he
r, had not spoken on the record
about the crash until an interview
with author Jeffrey Robinson for
his book, ``Rainier and Grace: A
n Intimate Portrait.`` She recounts
the accident in this excerpt from
the book. At about 9 a.m., on M
onday, Sept. 13, 1982, Princess
Grace of Monaco woke her daug

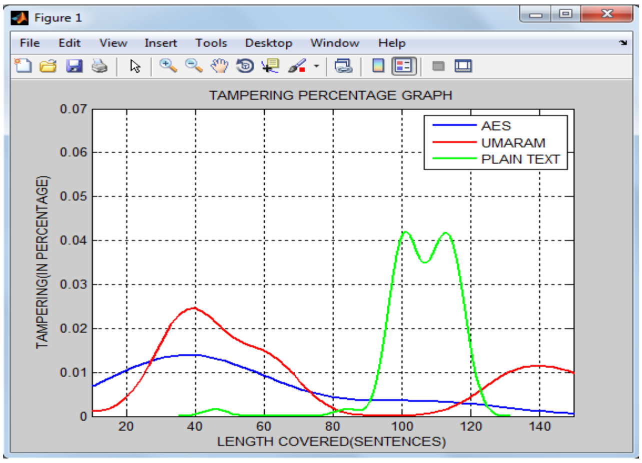
Figure 13: Key made invisible at front end generated using AES and UMARAM

6. COMPARISON OF UMARAM WITH AES

UMARAM and AES encryption techniques have been compared on three parameters- tampering, CPU time and robustness using text samples with different word count.

Tampering attack: Tampering attack is any attempt to destroy, expose, alter, disable, steal or gain unauthorized access to or make unauthorized use of an asset. It is a form of attack where the attacker inserts and deletes, words and phrases from several places in the text [7].

Graph below shows the percentage of tampering in original text document and watermarked documents using AES and UMARAM encryption algorithm for a text document with 2100 word count. The graph clearly shows that the plain text document has the highest susceptibility to tampering with tampering percentage greater than 4%. Tampering ranges between 1.2 to 2.5% on using AES encryption technique and the tampering percentage on using UMARAM ranges between 0.2 to 1.4%. Hence, UMARAM is least susceptible to tampering.



Graph 1: Percentage of tampering

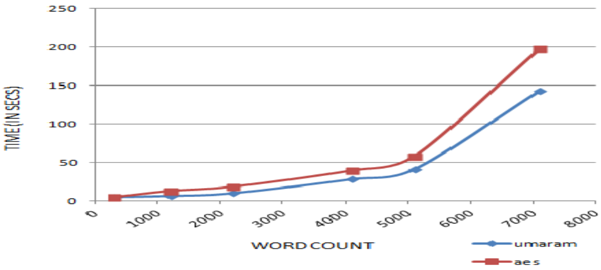
Tampering percentage of different sized documents has been compared in table 1 given below. UMARAM performs the best in all the cases.

Table1: Percentage of tampering

Document Size (In words)	AES	UMARAM
300	1.2%	0.1%
1200	1.7%	0.5%
2200	2.6%	1.3%
4100	3.8%	1.9%
5100	4.1%	2.1%
7200	4.8%	2.8%

CPU time: The CPU time is measured in seconds. It is the delay time taken for the encryption and the embedding process using UMARAM and AES encryption algorithm.

The Graph below shows the time taken in seconds for the encryption and embedding process using AES and UMARAM encryption algorithms for different sized text documents. The graph clearly shows that AES takes more time for execution as compared to UMARAM. The time taken for execution ranges between 4.5 to 197.7 seconds in case of AES and time delay ranges between 5.07 to 142.8 seconds in case of UMARAM. It clearly shows that UMARAM takes lesser time as compared to AES but the time taken for shorter documents is almost same.



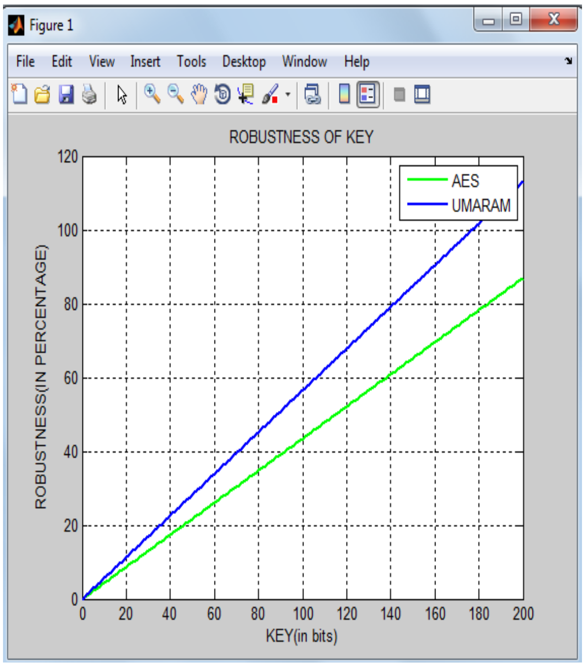
Graph 2: Time delay for UMARAM and AES algorithms

The time delay for different sized text documents have been compared in table 2 given below. UMARAM performs the best in all the cases.

Table2: Delay time comparison

Word Count	**UMARAM**	**AES**
300	5.07	4.552
1200	6.6768	12.6565
2200	10.1305	18.3365
4100	28.3922	39.4106
5100	41.0719	57.7395
7100	142.8023	197.7791

Robustness: Robustness of the watermark data means that the watermark data should not be destroyed if someone performs the common manipulations as well as malicious attacks. Graph 3 shows the robustness of the key made using UMARAM and AES encryption algorithm. The graph clearly shows that the key generated by UMARAM is more robust against attacks as compared to AES.



Graph 3: Robustness graph comparing keys of AES and UMARAM

7. CONCLUSION

The existing text watermarking algorithms are not robust against random tampering attacks. With the small volume of attack, it becomes impossible to identify the existence of attack and to prove authenticity of information. This study has developed a natural language component based watermarking algorithm, which utilizes the contents of text and suitable encryption technique to generate an encrypted watermark and this watermark is later extracted to prove the authenticity of text document. The performance of the algorithm has been evaluated for random tampering attack and robustness on variable size text documents. Results show that the new algorithm is more robust and is less susceptible to tampering.

8. REFERENCES

[1] Makarand L. Mali, Nitin N. Patil and J.B.Patil , "Implementation of Text Watermarking Technique Using Natural Language Watermarks", International Conference on Communication Systems and Network Technologies,2013.

[2] Ingemar J. Cox and Matt L. Miller , "The First 50 Years of Electronic Watermarking", EURASIP Journal on Apllied Signal Processing, Issue 2, pp. 126-132, 2002.

[3] Xinmin Zhou, Weidong Zhao, Sichun Wang and Rui Peng, "A Semi-FragileWatermarking Scheme for Content Authentication of Chinese Text Documents", IEEE,2009.

[4] Zunera Jalil, Anwar M. Mirza and Tahir Iqbal, "A Zero-Watermarking Algorithm for Text Documents based on Structural Components" ,IEEE, 2010.

[5] Suganya Ranganathan , Ahamed Johnsha Ali, Kathirvel.K & Mohan Kumar.M,"Combined Text Watermarking", International Journal of Computer Science and Information Technologies, Vol. 1 (5), 2010.

[6] Zunera Jalil and Anwar M. Mirza "A Review of Digital Watermarking Techniques for Text Documents", International Conference on Information and Multimedia Technology,2009.

[7] Zunera Jalil, M. Arfan Jaffar and Anwar M. Mirza,"A Novel Text Watermarking Algorithm Using Image Watermark",International Journal of Innovative Computing, Information and Control ,Volume 7, Number3,March2011.

[8] Patil Bharati Devidas and Patil Nitin Namdeo, "Text Watermarking algorithm using structural approach", IEEE, 2012.

[9] Mi-Young Kim, Osmar R. Zaiane, and Randy Goebel,"Natural Language Watermarking Based on Syntactic Displacement and Morphological Division".

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

[11] Yingli Zhang and Huaiqing Qin, "A Novel Robust Text Watermarking For Word Document", IEEE, 2010.

[12] Zhangjie Fu, Xingming Sun, Jiangang Shu and Lu Zhou, " Plain Text Zero Knowledge Watermarking Detection Based on Asymmetric Encryption", Advanced Science and Technology Letters ,Vol.48, 2014.

[13] Min Du and Quanyou Zhao, "Text Watermarking Algorithm based on Human Visual Redundancy", Advanced in Information Sciences and Service Sciences, Volume 3, Number 5, June 2011.

[14] Zhangjie Fu, Xingming Sun, Jiangang Shu, Lu Zhou and Jin Wang, "Verifiable Text Watermarking Detection to Improve Security" , International Journal of Security and Its Applications ,Vol.8, No.5,2014.

[15] Leena Goyal, Manoj raman, Prateek Divan and Mukaesh Kumar Vijay, " A Robust Method for Integrity Protection Of Digital Data in Text Document Watermarking", International Journal for Innovative Research in Science & Technology, Volume 1, Issue 6, November 2014.

[16] Chee Hon Lew and Chaw Seng Woo, "Design and Implementation of Text based Watermarking combined with Pseudo-Random Number Generator(PRNG) for Cryptography Application", Latest Trends in Applied Computational Science.

[17] Omar Tayan, Yasser M. Alginahi and Muhammed N. Kabir, "An Adaptive Zero-Watermarking Approach for Authentication and Protection of Sensitive Text Documents".

[18] Sukhpreet Kaur and Geetanjali Babbar, "A Zero-Watermarking algorithm on multiple occurrences of letters for text tampering detection", International Journal on Computer Science and Engineering, Vol. 5, No. 05 ,May 2013.

[19] Sonia Bajaj and Manshi Shukla, "Performance Evaluation of an approach for Secret data transfer using

interpolation and LSB substitution with Watermarking", International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014.

[20] Priyanka Verma, Rakhshan Anjum Shaikh and Ketki Deshmukh, " A Novel Approach to Angle based Invisible Text Watermarking with EBCDIC Coding", International Journal of Computer Applications, Volume61,No.20, January 2013.

[21] Manmeet Kaur and Kamna Mahajan, " An Existential review on Text Watermarking Techniques", International Journal of Computer Applications, Volume120, No.18, June 2015.

[22] G. Ramesh and R.Umarani, "UMARAM: A novel Fast encryption algorithm for data security in Loacal Area Network", ICCCCT,2010.