Mehran University Research Journal of Engineering and Technology
Vol. 40, No. 2, 323 - 334, April 2021
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: 10.22581/muet1982.2102.07

An Intelligent Three-Level Digital Watermarking Method for Document Protection

Umair Khadim[1,2], Muhammad Munwar Iqbal[2], Muhammad Awais Azam[3]

RECEIVED ON 23.07.2019, ACCEPTED ON 03.12.2019

ABSTRACT

Digital text is the most frequent interchange form of data that could hold sensitive information such as audit firms, banks, and educational institutes. This sensitive information needs to preserve its integrity and originality so that it could not only secure the data but also helps to identify ownership of text documents. This paper presents a novel and invisible digital watermarking approach for the secure exchange of text documents over the internet. Digital watermarking serves from the last decade for detection of forgery and tempering from digital text documents and maintained the copyright and authentication successfully. Many states of the art watermark techniques achieve high imperceptibility, robustness, and high hidden capacity; unfortunately failed to maintain the balance among these three conflicting parameters. As resolvent, we propose an intelligent Three-Level Digital Watermarking (3LDW) system for text documents copyright protection. 3LDW system can be applied to Microsoft Word objects, document open spaces, and text feature coding without affecting the content of the original document. Experimental results reveal that our proposed 3LDW system strongly resist against formatting attacks and efficiently preserves the imperceptibility. Additionally, embedding capacity analysis demonstrates a prominent improvement of the proposed system as compared to other similar approaches.

Keywords: Digital Watermarking, Copyright Protection, Ownership Verification, Imperceptibility, Robustness, Capacity.

1. INTRODUCTION

Information security has become a significant problem in the digital world in recent years because a vast amount of data is exchanged via a global network [1]. The exchanged data is a variation of text, image, audio, and video. Malicious attacks, threats, violations and illegal use of information are the major challenges for information security research [2]. The information security model consists of confidentiality, integrity, and availability as shown in Fig. 1. Confidentiality means that information will not be disclosed to unauthorized individuals. Integrity refers to the fact that the state of data must remain the same when it is transferred from one system to another.

Confidentiality
People
Information Security
Process
Technology
24
Integrity
Availability

Fig. 1: Information Security Model

[1] Department of Software Engineering, University of Kotli Azad Jammu and Kashmir, Pakistan.
Email: umair_khadim@uokajk.edu.pk (Corresponding Author)
[2] Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan.
Email: munwar.iq@uettaxila.edu.pk
[3] Department of Computer Engineering, University of Engineering and Technology, Taxila, Pakistan.
Email: awais.azam@uettaxila.edu.pk

The availability means that the data or information is accessible only for an authorized person. The communication makes our daily life easier in so many things with the growth of the internet. Nowadays, hackers use malware to access information and violate copyrights [3]. The booksellers and publishers spend millions of dollars for copyright authentication. There are various illegal actions that revile copyright protection regarding digital contents, like illegal copying, tampering, and forgery [4].

Digital watermarking protects the intellectual property of digital content and has played a critical role in the protection of copyright [5, 6]. In the past, steganography and cryptography were also used to solve the problem, but digital watermarking provides the best solution [7]. A general life cycle of digital watermarking is presented in Fig. 2. A watermarking system is divided into three parts: watermark embedding, attacks, and watermark detection as shown in Fig. 2.

In the embedding phase, secret data is embedded into the original signal that produces a watermarked signal, which is usually transmitted to other persons or stored. If someone modifies the document, then it is called an attack. The detection phase is also called verification that is used to authenticate the contents [8-10].

It is crucial to maintain data integrity while ensuring the confidentiality and availability of information [11]. Sensitive text documents are part of every company or organization like banks, audit firms, and educational institutes. A reliable method is required to authenticate text documents [12, 13]. In this research our main contribution is the following:

- The proposed three-level digital watermarking (3LDW) technique is imperceptible, secure and robust against formatting attacks. The proposed 3LDW technique can be applicable for text document ownership verification and copyright protection.
- The proposed technique is applicable to certain languages. As Microsoft word supports 91 languages, therefore the proposed technique can be applied in all 91 languages.
- The rest of the paper is structured as follows. Section II illustrates the related work. Section III is about the proposed methodology and presents the watermarking embedding and extraction process. Experimental results and discussion are presented in section IV. Section VI demonstrates the conclusion and feature work.

2. RELATED WORK

Digital text watermarking emerged in 1994 and grew with the passage of time. It is an active area of research and categorized into text, image, audio, and video [14]. The techniques used for text watermarking are classified into image-based, structural-based and hybrid techniques [15].

2.1 Image-Based Approach

In this approach, the contents of watermark information are treated as images or logos [16]. This approach is considered safe against formatting attacks, but it has limited applicability because it is not robust against re-typing attack [17]. Rizzo *et al.* [18] suggest a method based on a password that embeds the watermark in short text and preserves the appearance and content without converting text to the image. A graphical digital text watermarking method and algorithmic framework is proposed in Liu *et al.* [19], which consists of 8 levels that are a line, pixel, character, paragraph, page pixel, row, etymon, and three other aspects. Each level is divided into three typical features like similarity, structural and self-

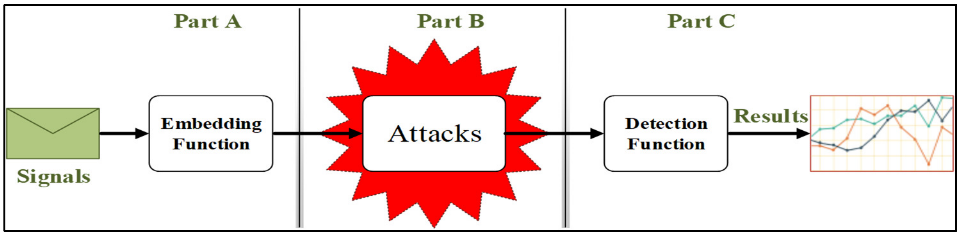


Fig. 2: General Digital Watermarking Life-Cycle Phases

characteristics features. Kundu *et al.* [20] propose a technique that measures print–and–scan transformation which can correlate the image before printing and then embed the watermark information.

2.2 Structural-Based Approach

In this approach, the structure of text is modified, like extra white spaces or line, and spaces are added for embedding watermark. The style of writing, letters and words location or the double letters are also incorporated for watermarking [21, 22]. The drawback of this approach has that when Optical Character Recognition (OCR) is applied, the spaces between the characters and words and the writing style are removed, which also ruin the watermark information. Aman *et al.* [23] proposes an open space format-based method that embeds the secret message in a Microsoft Word document. The white spaces are targeted to embed the watermark in a document. Liu *et al.* [24] proposes an algorithm based on Chinese text sentence features. The semantic code of the word calculates segmented text into sentences and sentence entropy. Sentence entropy, length, relevance, and weight function are used to find the weight of each sentence. The key is used for encryption and registered with a reliable third party. Zhu *et al.* [25] proposes an algorithm that connects the syllable of Chinese phonetic alphabet parts. The proposed algorithm has high resistance and strong robustness against tampering attacks.

2.3 Semantic-Based Approach

This approach uses the semantic of words for embedding the watermark and meaning of the text remains the same. In this approach, morphological alterations and words in the set are used for data hiding without disturbing original text meaning [26]. A syntactic and semantic approach is proposed by Mir *et al.* [27], where, the watermark information is first encrypted and then embeded into whitespaces using binary controlled characters. White spaces are used to embed the watermark throughout the text content. It is appropriate for web pages and offers security to protect watermarks.

2.4 Syntactic-Based Approach

In this approach syntactic tree is constructed first, then syntactic conversion is utilized for watermark embedding. In the syntactic structure, the text consists of sentences and words that can be nouns, verbs, adverbs, adjectives, prepositions, articles, *etc*. This technique is considered robust but cannot be applicable to all kinds of text like poetry, legal documents and transcripts [28]. A technique is proposed by Ren *et al.* [29] based on HiCod, HiOpt, HiPhs, and HiMax for text steganography in utilizing online short text. All proposed techniques are evaluated on the basis of security and performance about hiding ease and hiding rate.

2.5 Hybrid Approach

The hybrid approach is developed with the combination of different approaches of text watermarking to correct the weaknesses of each approach. This approach can be applied to extensive text documents and its robustness is also improved [30].

Saeed *et al.* [31] proposed a hybrid technique based on zero watermarking. The original structure of the document is not changed during embedding the watermark. Two steps are involved here, embedding watermark and extraction. A hybrid approach is proposed in [32], which is based on zero-watermarking. The integrity and originality of text documents is verified with the physical alteration. The proposed algorithm is robust against undetected content changes and is able to confirm proof of originality in temper detecting.

Every public or private organization or company transfers sensitive text documents, like legal documents, classified reports, declaration, and soft degrees. However, most of the existing schemes based on text watermarking are either imperceptible, robust or succession in obtaining high concealing capacity, but they do not achieve the balance between these conflicting parameters. In the said perspective, we propose an intelligent Three-Level Digital Watermarking (3LDW) system which provides copyright protection to text documents. Three-level embedding is applied, which includes Microsoft Word objects, document open spaces, and text feature coding without affecting the content of the original document.

3. METHODOLOGY

In this section, we describe the main characteristics of the proposed scheme. The proposed novel 3LDW digital watermarking technique is shown in Fig. 3, which utilizes the properties of a Microsoft Word document for watermarking without affecting the content and do not modify any word application. In the proposed technique, the watermark is embedded into text document. Three different properties, namely Microsoft Word objects, document open spaces, and text features are taken into consideration. The purpose of three-level embedding is to make the system more secure and efficient, so if any formatting attack disturbs the watermark then it recovers from other properties. Table 1 outlines the notations that are used throughout the paper.

Table 1: Notations

No.	Symbol	Description
1	M_S	Secret Message
2	K_{pu}	Public Key
3	M_b	Binary Conversion
4	M_d	Decimal Conversion
5	M_c	Cover Message
6	D_o	Original Document
7	D_w	Watermarked Document
8	W_{obj}	Microsoft Word Document Objects
9	M_e	Encrypted Message
10	W_n	Number of Groups

The proposed scheme is imperceptible, robust and incorporates a large amount of embedding capacity. The layout of the document is affected by manipulating these properties, and any standard word application command cannot amuse the watermark information. The secret message is encrypted with the help of a private key. Encryption is applied to preserve the watermark and make it difficult for the attacker. The watermarking is our focus in this research, so any encryption algorithm is applied for encrypting the secret message. Advanced Encryption Standard (AES) is applied here to encrypt the secret message.

After encryption, the secret message is converted into binaries then divided into n number of groups. The secret message called watermark W(n) if attacked, only 1/n of the watermark is demolished, where W_i is a group of watermark information as shown in equation (1).

$$W_n = \left\{\frac{W_i,\{i|i=1,2,\cdots,n\}}{W_n}\right. \quad (1)$$

3.1 Three Level Watermark Embedding

After generating the watermark information, the original Microsoft Word document is given as input to the system then three-level watermark embedding is applied. Microsoft Word document has a lot of properties that can be manipulated for watermarking and it will not affect the original document.

3.1.1 First Level Embedding (Word Object)

The Microsoft Word document comprises of a lot of word objects that authorize the users to interact and manipulate it. These objects are appropriate for two reasons: first, without affecting imperceptibly the vast capacity of watermark information is stored. Second, the watermark information is not affected by any mutual command of Microsoft Word. These objects are used in the documents to preserve the macro setting between macro sessions and stored as part of the document. The algorithm I describes the first level watermark embedding, where watermark information is embedded into text document's different objects, with the objective to attain the robustness and security. Any common instruction of Microsoft Word application cannot interrupt the watermark, because the watermark is stored in preserve micro setting.

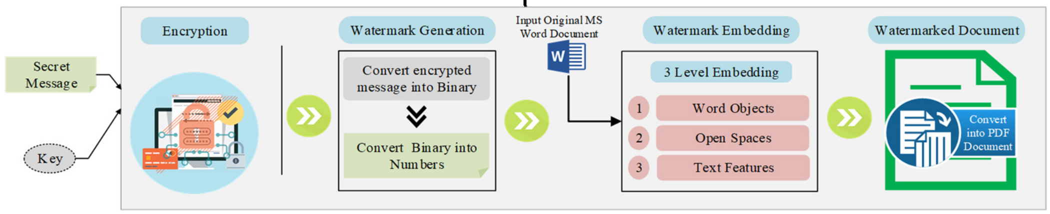


Fig. 3: Proposed 3LDW Model

```
The Algorithm1: First Level Watermark Embedding

    **Input:** $M_o$, $M_s$ and $K_{pu}$
    **Output:** $D_w \leftarrow (D_o, M_s, K_{pu})$
    **Start:**
    $M_e \rightarrow [M_s, K_{pu}]$ // Encrypt Secret Message
    $M_b$ = toBinary(Me)
        $W_n \rightarrow M_s$
        $W_n = \left\{\frac{W_i,\{i|i=1,2,\cdots,n\}}{W_n}\right.$
    **For each**
    Check $W_{obj}$//$W_{obj}$ = Microsoft Word Document Objects
    For (Range → R)
    $R \rightarrow W_n$
    $W_{i+1}$
    end for
    end for each
    Watermarked Document $D_w$
    end
```

3.1.2 Second Level Embedding (Open Spaces)

The open spaces are utilized for watermark embedding in second-level embedding. The open spaces are searched in the document and the watermark information is inserted one by one. The encrypted watermark information is converted into bits and embedded into selected open spaces. Algorithm II describes the complete procedure of second-level watermark embedding in document's open spaces.

```
Algorithm II: Second Level Watermark Embedding1:
**Input:** $D_o$ , $M_S$ and $K_{pu}$
**Output:**        $D_w \longleftarrow (D_o , M_S , K_{pu})$
**Start:**
        $M_e \rightarrow [M_s, K_{pu}]$        // Encrypt Secret Message
        $M_b$ = toBinary($M_e$)
        W(i) = $M_b$
        **For each**
          Find $D_{os}$// $D_{os}$ = Find Microsoft Word Document Open Spaces
          For ($D_{os} \rightarrow W_n$)
              $W_n$ = W(i)
              $W_{n+1}$
              i++
                        **End For**
        **End For each**
        Watermarked Document $D_w$
        **End**
```

3.1.3 Third Level Embedding (Text feature coding)

In third level embedding features of the text are used for watermarking. First, the secret message is encrypted and converted into numbers as shown in Fig. 4. The original text document is given as input then spaces, commas, full stop, semicolon, colon, question mark, inverted commas, special characters, and symbols are removed from the plain text. The Algorithm III provides the complete procedure for watermark embedding. For example, we have a decimal numbers array like [67027603…...], the first number is 6, and after increasing 1 it can locate the character at index 7. The increment of 1 is added in every number to handle the value of zero. The next number is taken from the array which became 8 after the increment, the current number is added in the last number (7) and locate character at 15th index. Then again increment 1 in next number which becomes 1 and donates the character at 16th index and does the changes in character format. After completion of all the numbers, Microsoft Word document is transformed into PDF and communicate.

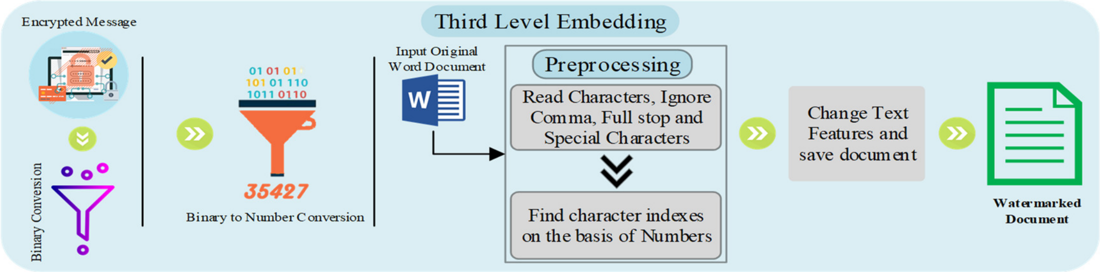

Third Level Embedding
Encrypted Message
Binary Conversion
01 01 01
101 01 110
1011 0110
35427
Binary to Number Conversion
Input Original Word Document
Preprocessing
Read Characters, Ignore Comma, Full stop and Special Characters
Find character indexes on the basis of Numbers
Change Text Features and save document
Watermarked Document

Fig. 4. Third Level Digital Watermark Embedding

Algorithm III: Third Level Watermark Embedding

```
Input:   $D_o$ , $M_S$ and $K_{pu}$
Output:  $D_w \longleftarrow (D_o , M_S , K_{pu})$
    Start:
    $M_e \rightarrow [M_s, K_{pu}]$ // Encrypt Secret Message with Key
    $M_b$= toBinary($M_e$)    // $B_n$ = Convert $E_n$ into Binary
    $N_d$ = ($M_b$)    // Dec[i] = Convert $B_n$ into Decimal Number Array
    Input (T) // T = Input Original Document
    sum=1; count=0; D[j]
    while (i to Dec[i]) do
        sum=sum+ D[j]
        for (i to totalchar(T)) do
          if (i! = Space + Comma + Full stop +….) then
              count = count +1
              if (count == sum) then
        Replace the character with text features
              end if
          end if
    end for
     end while
    Watermarked Document $D_w$
    Convert to PDF and share
End
```

3.2 Three Level Watermark Extraction

The watermark extraction or verification is the reverse process of embedding as shown in Fig. 5. The watermarked document is given as input then three-level extraction is applied. After extraction, regenerate the watermark information is regenerated, numbers are converted into binary and then binary to characters. The same key is applied to decrypt the message using AES algorithm. After decryption, secret message is compared with an original watermark that authenticates the originality of the document. If original watermarked information is matched with watermark, then it is called an original document otherwise it is tempered or changed document.

4. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we analyse the experimental results of our proposed scheme. The performance of the proposed scheme is measured in terms of capacity, imperceptibility, and robustness. Imperceptibility test is performed in the first sub-section, whereas the second sub-section analyses the robustness and the third sub-section analyses the rate of embedding capacity.

4.1 Imperceptibility Test

The imperceptibility has a primary and fundamental requirement of the watermarking. The watermark cannot be detected and seen by human eyes. Only through special processing or authorized agency can detect the watermark. The statistical analysis is performed, which is more powerful than visual analysis for imperceptibility test. The similarity score of two strings is computed using Jaro Winkler Distance [33] using equation (2). The threshold of 0 and 1 is standardized, where 0 is equivalent to no similarity and 1 represents the exact match.

$$dj = \begin{cases} 0, & \text{if } m=0 \\ \frac{1}{3}\left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{|m|}\right), & \text{otherwise} \end{cases} \qquad (2)$$

The imperceptibility measure between the two strings is illustrated in Fig. 6, where 20 different message stings remain similar in the proposed technique, but in the existence techniques [34, 35] it varies from 0.83 to 0.97. In the proposed technique, average similarity is 0.99, which demonstrates that the proposed technique is 0.99% imperceptible as compared to the previous techniques.

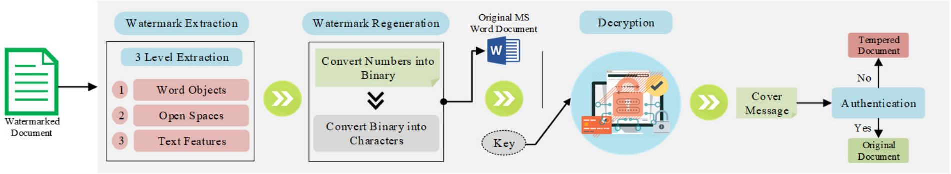


Fig. 5: Three Levels Digital Watermarking Extraction or Verification

The proposed technique performs excellently in terms of imperceptibility because the watermark is stored in special properties and it cannot affect the whole document. After embedding the watermark information, the original and the watermarked document look like the same as shown in Fig. 7. The overall layout of the document is also not affected.

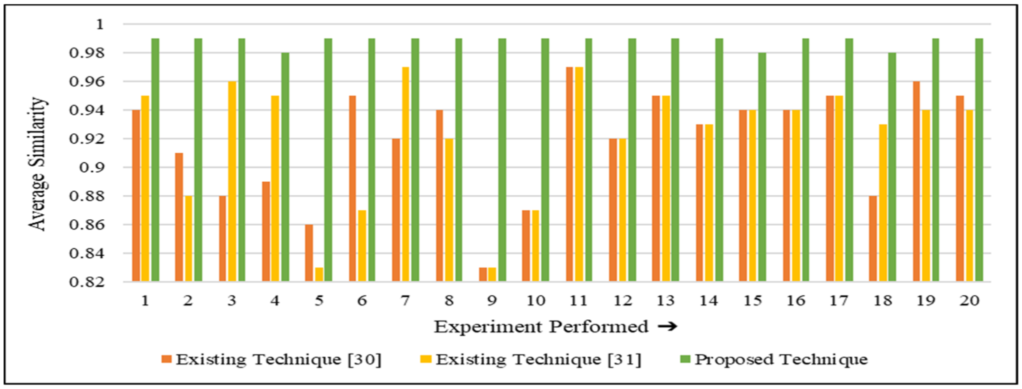
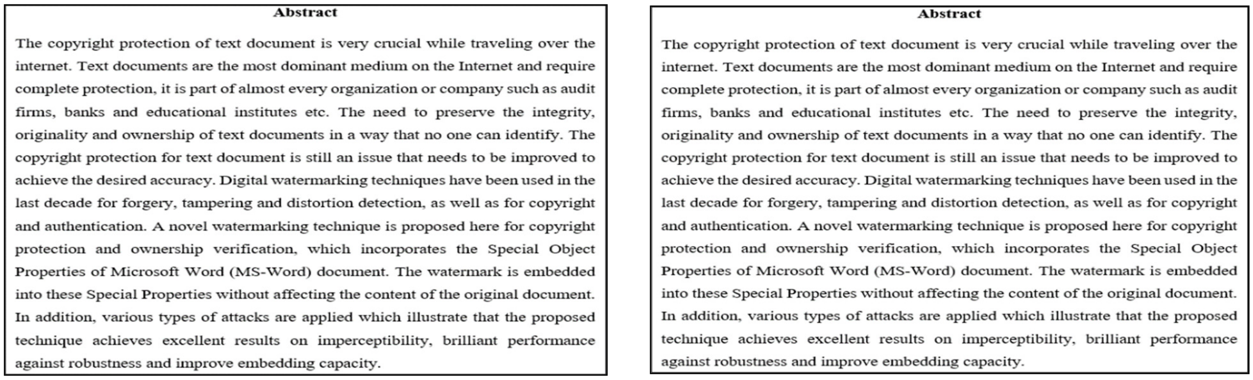


Fig. 6: Imperceptibility Analysis with Existing Techniques



Abstract

The copyright protection of text document is very crucial while traveling over the internet. Text documents are the most dominant medium on the Internet and require complete protection, it is part of almost every organization or company such as audit firms, banks and educational institutes etc. The need to preserve the integrity, originality and ownership of text documents in a way that no one can identify. The copyright protection for text document is still an issue that needs to be improved to achieve the desired accuracy. Digital watermarking techniques have been used in the last decade for forgery, tampering and distortion detection, as well as for copyright and authentication. A novel watermarking technique is proposed here for copyright protection and ownership verification, which incorporates the Special Object Properties of Microsoft Word (MS-Word) document. The watermark is embedded into these Special Properties without affecting the content of the original document. In addition, various types of attacks are applied which illustrate that the proposed technique achieves excellent results on imperceptibility, brilliant performance against robustness and improve embedding capacity.

Abstract

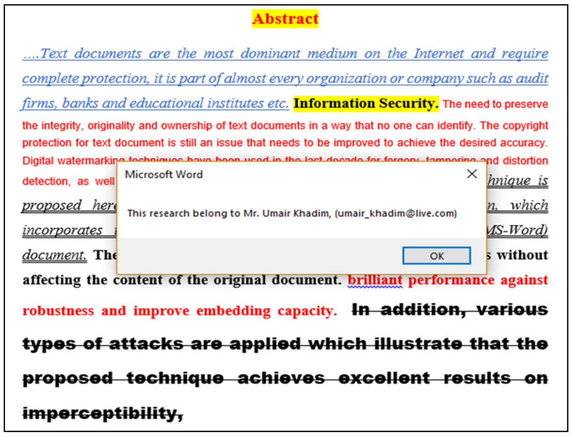
The copyright protection of text document is very crucial while traveling over the internet. Text documents are the most dominant medium on the Internet and require complete protection, it is part of almost every organization or company such as audit firms, banks and educational institutes etc. The need to preserve the integrity, originality and ownership of text documents in a way that no one can identify. The copyright protection for text document is still an issue that needs to be improved to achieve the desired accuracy. Digital watermarking techniques have been used in the last decade for forgery, tampering and distortion detection, as well as for copyright and authentication. A novel watermarking technique is proposed here for copyright protection and ownership verification, which incorporates the Special Object Properties of Microsoft Word (MS-Word) document. The watermark is embedded into these Special Properties without affecting the content of the original document. In addition, various types of attacks are applied which illustrate that the proposed technique achieves excellent results on imperceptibility, brilliant performance against robustness and improve embedding capacity.

Fig. 7: Before Watermark Embedding on Right Side, After Watermark Embedding on Left Side

4.2 Robustness Test

Robustness means that after applying different formatting attacks, f the embedded data remains safe. To measure the robustness of the proposed technique different formatting attacks are applied that can be seen in Fig. 8. Insertion, deletion, re-ordering and formatting attacks like font color, font size, italic, underline and text highlighting are applied on the watermarked document.

As shown in Fig. 8, different formatting attacks are applied where the font size of the first word "Abstract" is increased 2 points (16), red font color and highlighter is also applied. The first line of watermark document is "deleted" and replaced with dots (….). The formatting of the second line is also changed and



Abstract

....Text documents are the most dominant medium on the Internet and require complete protection, it is part of almost every organization or company such as audit firms, banks and educational institutes etc. **Information Security.** The need to preserve the integrity, originality and ownership of text documents in a way that no one can identify. The copyright protection for text document is still an issue that needs to be improved to achieve the desired accuracy. Digital watermarking techniques have been used in the last decade for forgery, tampering and distortion detection, as well

hnique is

proposed here

n, which

incorporates

MS-Word)

document. **The**

s without affecting the content of the original document. **brilliant performance against robustness and improve embedding capacity.** **~~In addition, various types of attacks are applied which illustrate that the proposed technique achieves excellent results on imperceptibility,~~**

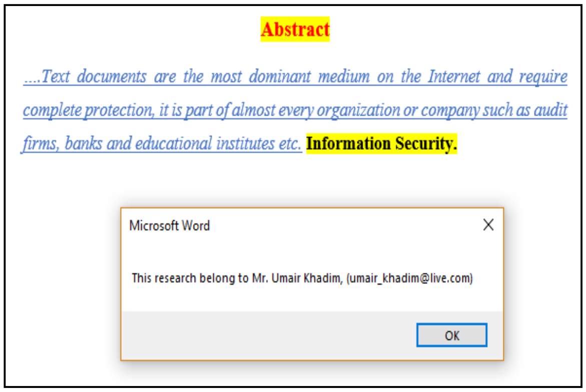
Microsoft Word

This research belong to Mr. Umair Khadim, (umair_khadim@live.com)

Fig. 8: 100% Watermark Information is extracted after formatting attacks

two words “information security” are inserted. The third, fourth- and fifth lines formatting is changed as; font-family: Arial, font-size:10 points, font colour: red and strike last three words. The sixth line is converted to italic and double underline is applied, and only bold the contents of line seven. In the last line, after “result on Imperceptibility” is cut and paste at the beginning of line eight. The font family and the font size of the eighth line are changed, and the strike is applied. After applying formatting attacks, the watermark information is still extracted from word objects.

After applying the 90% deletion attack, the 100% watermark information is extracted as shown in Fig. 9. We apply three-level embedding, so if the data is deleted in formatting attacks it can be recovered from word objects or text feature coding.



Abstract

....Text documents are the most dominant medium on the Internet and require complete protection, it is part of almost every organization or company such as audit firms, banks and educational institutes etc. **Information Security.**

This research belong to Mr. Umair Khadim, (umair_khadim@live.com)

Fig. 9: Robustness result against 90% deletion Attack

The circos graph in Fig. 10 illustrates that after applying formatting attacks the Watermark Detection Rate (WDR) is 99.9%. Twenty experiments are performed with different insertion, deletion and re-ordering rate that is between 10 % to 90 %. The “W” in the circos graph defines the watermark document size, and “WDR” represents the Watermark Detection Rate which is 99.9%. As three-level digital watermarking (3LDW) is applied on Word document objects, document open spaces and text features without altering the physical contents of the original document are added, which can increase the watermark detection rate. The proposed technique is robust against formatting attacks because if document open spaces or text features are removed then we can extract the watermark from the document objects. Through the experimental results, it has been demonstrated that the proposed technique performs excellently against the formatting attacks.

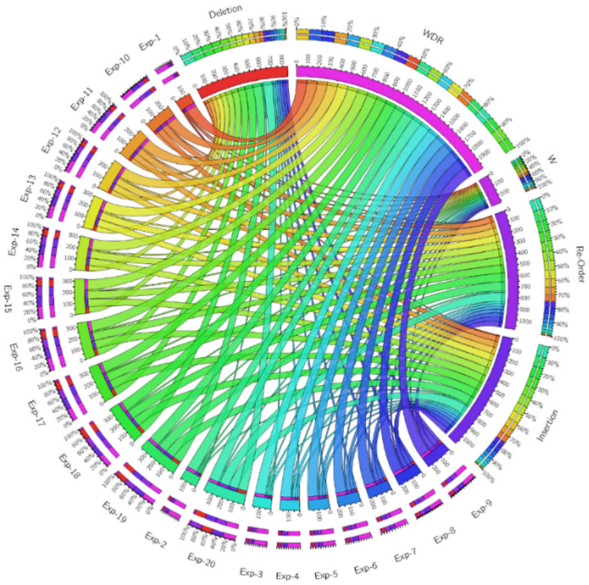


Fig. 10: Insertion, Deletion and Re-Ordering Attacks

4.3 Capacity Analysis

In text watermarking, hidden capacity analysis is the major parameter to measure the strength of the proposed algorithm [36]. The capacity indicates the maximum number of bits with the name of a watermark that can be embedded in the host document. A novel system is required that maximizes hidden capacity without affecting other conflicting parameters, such as robustness and imperceptibility.

A technique is considered decent if it has high embedding capacity and does not affect the visibility of the watermark. The capacity of the proposed system can be measured by equation (3).

$$\text{Capacity} = \frac{\text{Total No.of bits (Secret Data)}}{\text{Total No.of curves file data (Kb)}} \times 100 \qquad (3)$$

The proposed technique can incorporate 917 characters. The capacity analysis of the proposed technique is presented in Table 2 where ten different document sizes with watermarks of different lengths are examined. The original and the watermark document sizes are compared and the change in the size of the document is also measured. After embedding the 917 characters, the size of the water document is changed by 0.08% which is acceptable.

As mentioned above, the word objects of documents are chosen for two reasons, first for incorporating large embedding capacity and second, they do not affect the contents of the original documents.

Table 3 presents the capacity comparison, where the proposed technique has a higher embedded capacity as compared with [35, 37, 38, 39]. The size of the original and watermark document is compared in Fig. 11, which shows that after embedding the watermark information, there is no significant change in the watermark document. When the secret message is 84 bytes then watermarked document size is 13,432, and after increasing the amount of watermark information by 917 bytes, the watermarked document size is 13,508.

The proposed scheme performs brilliantly against robustness, imperceptible and also incorporates a massive amount of capacity. The proposed technique is robust and more secure because special properties are considered for watermarking. It can be applied for document authentication and copyright protection, which protects the document against unauthorized access and illegal use.

5. CONCLUSION

In this paper, a novel 3LDW system has been proposed to achieve robustness, imperceptibility and high embedding capacity. The experimental results demonstrate that the performance of the proposed technique is improved. A three-level watermark embedding is applied, which includes word objects, open spaces and text feature coding for concealing the watermark information. The proposed scheme is robust as we embed the watermark in three-levels. If any formatting attack or Optical Character Recognition is applied, the watermark information is retrieved from other properties. Furthermore, watermark information is still retrieved from word objects after applying 90% deletion attack. In the

Table 2: Watermarking Capacity Analysis of Proposed Technique

Experiment No	Secret Message in bytes	Words	Characters	Original Document Size in bytes	Watermarked Document Size in bytes	Size Change in %
1	84	12	84	13,407	13,432	0.11
2	105	19	105	13,432	13,442	0.10
3	170	25	170	13,442	13,455	0.13
4	213	29	213	13,455	13,463	0.08
5	257	38	257	13,463	13,472	0.09
6	325	55	325	13,472	13,478	0.06
7	515	73	515	13,478	13,485	0.07
8	713	133	713	13,485	13,490	0.05
9	879	171	879	13,490	13,500	0.10
10	917	145	917	13,500	13,508	0.08

Table 3: Comparison of proposed technique with existing techniques in terms of capacity

Sr. No	Authors	Words	Characters	Size in Bits
1	Cheng et al. [35]	6	32	256
2	Alotaibi et al. [37]	21	130	1040
3	Taha et al. [38]	28	186	1490
4	Liang et al. [39]	32	197	1576
5	Proposed Technique	145	917	7336

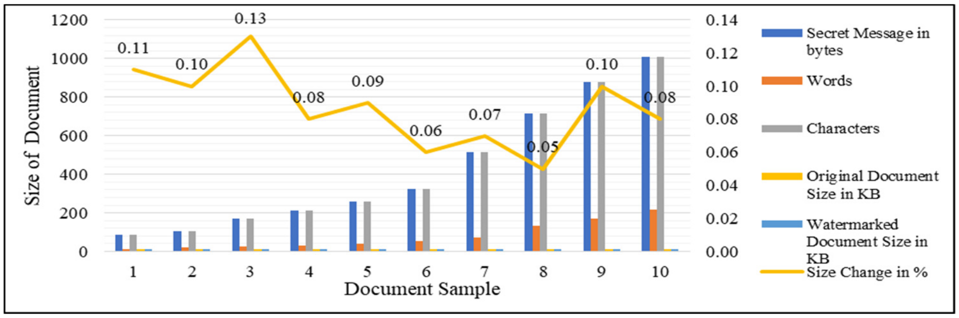


Fig. 11: Capacity Analysis Original Document Size and Watermarked Document Size

future research, the area of textual watermarks will be further investigated. We will analyse the other possible attacks to enhance the robustness in text documents. Moreover, the Portable Document Format (PDF) documents will also be investigated.

REFERENCES

1. Iqbal M.M., Khadam U., Han K.J., Han J., Jabbar S., "A Robust Digital Watermarking Algorithm for Text Document Copyright Protection based on Feature Coding", *Proceedings of the 15th IEEE International Communications and Mobile Computing Conference (IWCMC),* pp. 1940-1945, Tangier, Morocco, 2019.
2. Al-Maweri N.A.S., Adnan W.A.W., Ramli A.R., Samsudin K., Abdul Rahman S.M.S.A., "Robust Digital Text Watermarking Algorithm based on Unicode Extended Characters", *Indian Journal of Science and Technology,* Vol. 9, No. 48, pp. 1-14, 2016.
3. Zeeshan M., Ullah S., Husain R.G., Nasir N., "A Review Study on Unique Way of Information Hiding: Steganography", *International Journal on Data Science and Technology,* Vol. 3, No. 5, 2017.
4. Rizzo S.G., Bertini F., Montesi D., "Text Authorship Verification through Watermarking", *Proceedings of the European Intelligencel and Security Informatics Conference (EISIC),* pp. 168-171, Uppsala, 2016.
5. Singh P., Chadha R., "A survey of digital watermarking techniques, applications and attacks", *International Journal of Engineering and Innovative Technology,* Vol. 2, No. 9, pp. 165-175, 2013.
6. Habib M.A., Ahmad M., Jabbar S., Ahmed S.H., Rodrigues J.J.P.C., "Speeding Up the Internet of Things: LEAIoT: A Lightweight Encryption Algorithm Toward Low-Latency Communication for the Internet of Things", *IEEE Consumer Electronics Magazine,* Vol. 7, No. 6, pp. 31-37. 2018.
7. Khadam U., Iqbal M.M., Azam M.A., Khalid S., Rho S., Chilamkurti N., "Digital Watermarking Technique for Text Document Protection Using Data Mining Analysis", *IEEE Access,* Vol. 7, pp. 64955-64965, 2019.
8. Xiang S., He J., "Database authentication watermarking scheme in encrypted domain", *IET Information Security,* Vol. 12, No. 1, pp. 42-51, 2017.
9. Khadim U., Khan A., Ahmad B., Khan A., "Information hiding in text to improve performance for word document", *International Journal of Technology and Research,* Vol.3, No.3, 2015.
10. Dabbagh M., Rayes A., *"Internet of things security and privacy in Internet of Things From Hype to Reality",* pp. 211-238, Springer., 2019.
11. Kamaruddin N.S., Kamsin A., Por L.Y., Rahman H., "A Review of Text Watermarking: Theory, Methods and Applications", *IEEE Access,* Vol.6, pp. 8011-8028, 2018.
12. Ahvanooey M.T., Li Q., Shim H.J., Huang Y., "A Comparative Analysis of Information Hiding Techniques for Copyright Protection of Text Documents", *Security and Communication Networks,* Vol. 2018, 2018.
13. Habib M.A., Ahmad M., Jabbar S., Khalid S., Chaudhry J., Saleem K., Rodrigues J.J.P.C., Khalil M.S., "Security and privacy based access control model for internet of connected vehicles", *Future Generation Computer Systems,* Vol. 97, pp. 687-696, 2019.
14. Brassil J.T., Low S., Maxemchuk N.F., O'Gorman L., "Electronic marking and identification techniques to discourage document copying", *IEEE Journal on Selected Areas in Communications,* Vol. 13, No.8, pp. 1495-1504, 1995.
15. Brassil J., Low S., Maxemchuk N.F., O'Gorman L., "Hiding information in document images", *Proceedings of the Conference on Information Sciences and Systems (CISS-95),* Vol. 29, pp. 482-489, 1995.
16. Hao Y., Lin Chuang Q.F., Rong D., "A survey of digital watermarking", *Journal of Computer Research and Development,* Vol. 42, No.7, pp. 1093-1099, 2005.
17. Kaur M., Mahajan K., "An existential review on text watermarking techniques", *International Journal of Computer Applications,* Vol. 120,

2015.Rizzo S.G., Bertini F., Montesi D. "Content-preserving text watermarking through unicode homoglyph substitution", *Proceedings of the 20th International Database Engineering and Applications Symposium,* ACM, 2016.

18. Liu X., Zhang J., Wang H., Gong X., Cheng Y., A Novel Text Watermarking Algorithm Based on Graphic Watermarking Framework. in Broadband and Wireless Computing, *Proceedings of the Ninth IEEE International Conference on Communication and Applications (BWCCA),* pp. 84-88, Guangdong, 2014.
19. Kundu M.K., Maiti A.K., "An inexpens*ive* digital watermarking scheme for printed document", *Proceedings of the IET International Conference on Visual Information Engineering, pp. 298-303, Bangalore,* India, 2006.
20. Jeevan K., Krishnakumar S., Image hiding technique using a pseudo hexagonal structure approach, *International Journal of Computers and Applications,* Vol. 41, No. 5, pp. 359-366, 2019.
21. Bezerianos A., Dragicevic P., Balakrishnan R., "Mnemonic rendering: an image-based approach for exposing hidden changes in dynamic displays", *Proceedings of the 19th annual ACM Symposium on User interface software and Technology,* Montreux, Switzerland, October 2006.
22. Aman M., Khan A., Ahmad B., Kouser S., "A hybrid text steganography approach utilizing Unicode space characters and zero-width character", *International Journal on Information Technologies and Security,* Vol.9, No.1, pp.. 85-100, 2017.
23. Liu Y., Y. Zhu, Xin G., "A zero-watermarking algorithm based on merging features of sentences for Chinese text", *Journal of the Chinese Institute of Engineers,* Vol. 38, No.3, p. 391-398, 2015.
24. Zhu P., Xiang G., Song W., Li A., Zhang Y., Tao R., "A text zero-watermarking algorithm based on Chinese phonetic alphabets", *Wuhan University Journal of Natural Sciences,* Vol. 21, No.4, pp. 277-282, 2016.
25. Topkara U., M. Topkara, Atallah M.J.. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions, *Proceedings of the 8th workshop on Multimedia and Security,* ACM, pp. 164-174, Geneva, Switzerland, 2006.
26. Mir N., Copyright for web content using invisible text watermarking. *Computers in Human Behavior,* Vol. 30, pp. 648-653, 2014.
27. Jalil Z., Mirza A.M., Iqbal T., A zero-watermarking algorithm for text documents based on structural components", *Proceedings of the International Conference on Information and Emerging Technologies (ICIET),* pp. 1-5, Karachi, 2010.
28. Ren W., Liu Y., Zhao J., "Provably secure information hiding via short text in social networking tools", *Tsinghua Science and Technology,* Vol. 17, No. 3, pp. 225-231, 2012.
29. Nematollahi, M.A., C. Vorakulpipat, and H.G. Rosales, "*Text Watermarking, in Digital Watermarking*", pp. 121-129. Springer. 2017.
30. Saeed F., Dixit A., "Hybrid HSW Based Zero Watermarking for Tampering Detection of Text Contents", *Proceedings of the International conference on Computer Networks, Big data and IoT,* Medurai, India, 2018.
31. Tayan O., Kabir M.N., Alginahi Y.M., "A hybrid digital-signature and zero-watermarking approach for authentication and protection of sensitive electronic documents", *The Scientific World Journal,* Vol. 2014, pp. 1-14, 2014.
32. Ahmadoh, E.M., Gutub, Utilization of two diacritics for Arabic text A.A., "Steganography to enhance performance", *Lecture Notes on Information Theory,* Vol. 3, No.1, pp. 42-47, 2015.
33. Wang Z.-H., Chang C-C, Kieu D., Li M-C, "Emoticon-based text steganography in chat", *Proceedings of the Asia-Pacific Conference on Computational Intelligence and Industrial Applications,* pp. 457-460, 2009.
34. Cheng W., Feng H., Yang C., "A robust text digital watermarking algorithm based on fragments regrouping strategy", *Proceedings of the International IEEE Conference on Information Theory and Information Security (ICITIS),* pp. 600-603, Beijing, 2010.
35. Naqvi N., Abbasi A.T., Hussain R., Khan M.A., Ahmad B., "Multilayer partially homomorphic encryption text steganography (MLPHE-TS): a zero steganography approach", *Wireless Personal*

Communications, Vol. 103, No.2, pp. 1563-1585.2018.

36. Alotaibi R.A., Elrefaei L.A., "Improved capacity Arabic text watermarking methods based on open word space", *Journal of King Saud University-Computer and Information Sciences,* Vol. 30, No.2, pp. 236-248, 2017.
37. Taha A., Hammad A.S., Selim, "A high capacity algorithm for information M.M. hiding in Arabic text", *Journal of King Saud University-Computer and Information Sciences,* Vol. 32, No.6, pp. 658-665, July 2020.
38. Liang O.W., Iranmanesh V., "Information hiding using whitespace technique in Microsoft word", *Proceedings of the 2nd IEEE International Conference on Virtual Systems and Multimedia (VSMM)*, Kuala Lumpur, pp. 1-5, 2016.