IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 15, 2019, accepted May 7, 2019, date of publication May 14, 2019, date of current version May 31, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2916674*

# Digital Watermarking Technique for Text Document Protection Using Data Mining Analysis

**UMAIR KHADAM[1], MUHAMMAD MUNWAR IQBAL[1], MUHAMMAD AWAIS AZAM[2], SHEHZAD KHALID[3], SEUNGMIN RHO[4], AND NAVEEN CHILAMKURTI[5]**

[1]Department of Computer Science, University of Engineering and Technology, Taxila 47050, Pakistan
[2]Department of Computer Engineering, University of Engineering and Technology, Taxila 47050, Pakistan
[3]Department of Computer Engineering, Bahria University, Islamabad 44220, Pakistan
[4]Department of Software, Sejong University, Seoul 100083, South Korea
[5]Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, VIC 3086, Australia

Corresponding author: Seungmin Rho (smrho@sejong.edu)

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2016R1D1A1A09919551.

**ABSTRACT** In the current era, information security is on its top priority for all organizations. The individuals, government officials, and military with the rapid development of Internet technologies like the Internet of Things (IoT), big data, and cloud computing facing data security problems. As the massive rate of data growth, it is a challenging task for the researchers, that how to manage the vast amount of data safely and effectively while designing smart cities. It has been quite easy to produce an illegal copy of digital contents. The verification of digital content is one of the major issues because digital contents are generated daily and shared via the internet. The limited techniques are available for document copyright protection. However, most of the existing techniques produce distortion during watermark insertion or lack of capacity. In the said perspective, a digital watermarking technique is proposed for document copyright protection and ownership verification with the help of data mining. The techniques of data mining are applied to find suitable properties from the document for embedding watermark. The proposed model provides copyright protection to text documents on local and cloud computing paradigm. For the evaluation of the proposed technique, 20 different text documents are used to perform many attacks such as formatting, insertion, and deletion attacks. The proposed technique attained a high-level of imperceptibility where peak signal to noise ratio (PSNR) values are between 64.67% and 71.03%, and similarity (SIM) percentage is between 99.92% and 99.99%. The proposed technique is robust and resists from formatting attacks and capacity of the proposed technique is also improved as compared to the previous techniques.

**INDEX TERMS** Information security, digital watermarking, copyright protection, data mining, cloud computing, Internet of Things.

## I. INTRODUCTION

Nowadays, one of the major types of information in the world is in digital format. Both positive and negative aspects of the digital format are progressing in the modern digital world. Positive aspects include progress in astronomy, medical science, and technologies. On the other hand, corresponding to the negative aspects, the misuse of these technologies raises many issues like copyright protection and data manipulation. Due to the advanced technologies such as high-speed computer networks, and Internet, etc. different ways have been used to illegal copy, redistribute and store the digital contents easily. It is necessary to secure digital contents and protects them against unauthorized copy [1]. Internet of Things (IoT) and cloud has received significant support from governments and research institutes around the world [2]. Data is shifted on cloud computing in the form of audio, video, image and text. The ownership verification and copyright protection of data is a challenging task. Data is the crucial element in smart cities which sustains the infrastructure of data and helps people to gain access to digital contents. The architecture of the smart city is presented in Fig. 1, where data is stored, processed and analyzed the central location. Digital watermarking provides a solution for digital contents copyright

The associate editor coordinating the review of this manuscript and approving it for publication was Chun-Wei Tsai.



2169-3536 © 2019 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.


**FIGURE 1.** The architecture of a smart city.

protection and ownership verification. A secret message is placed inside a digital content without compromising valuable data. This secret information is used later for ownership identification. Digital watermarking is categorized into text watermarking, image watermarking, audio watermarking and video watermarking. Most of the research has focused on image, audio, and video. Currently, text watermarking has received popularity due to large numbers of the text document are produced and shared [3].

The individuals, government officials, and military facing data security problems which also affect the smart cities. Digital publishers have rights but facing many threats, such as illegal use of copyrights, data manipulation, and redistribution of information [4]. Text documents are part of almost every organization or company such as audit firms, banks, or any large private or public corporation. These documents are in the form of financial statements, legal notes, birth certificates, soft degrees, classified reports and declarations [5].

However, most of the existing techniques produce distortion during watermark insertion, which directly impacts on imperceptibility. Moreover, most of the existing techniques are not robust or lack of capacity. Converting a digital file to another format has the risk of losing the embedded watermark. The challenge is to ensure the originality and copyright protection of text document, which required a proper watermark technique that is robust against formatting attacks, imperceptible, achieve high embedding capacity and secured. This issue can be addressed by a new framework which is proposed here to overcome the text watermarking current challenges.

Our main contribution in this research are following:

1. A novel digital text watermarking model is proposed with the help of data mining techniques. Suitable properties of MS Word document are selected using data mining for embedding large size of watermark information.
2. We proposed a secure and robust digital watermarking technique which provides copyright protection to text documents on local and cloud paradigm with the help of data mining.
3. The proposed technique is robust 99.9% against formatting attacks, imperceptible and secure with the objective to protect the text documents with high capacity.
4. The proposed technique supports format transformation and applicable to certain languages.
5. The proposed technique will be applicable for big data and the Internet of Things (IoT) which enhance the security of digital text documents in smart cities.

The rest of the paper is structured as follows. Section 2 illustrates the related work. Section 3 demonstrate the watermarking embedding and extraction process. Section 4 describe the methodology of the proposed work. Section 5 evaluates the results of the experiments, whereas Section 6 concludes the presented work and future direction.

## II. RELATED WORK

Text digital watermarking is a critical area of research and emerged in 1991. With the passage of time as internet grow and communication starts all over the world several numbers of text watermarking techniques are proposed. These techniques are image-based, linguistic-based which includes (semantic and syntactic), structural-based and hybrid approaches [6], [7].

### A. IMAGE-BASED APPROACHES

In image-based approach, the text is interpreted as an image for embedding the watermark into the cover file [8]. The watermark logo or image converted to a text string, and the watermarked data is produced. It can be interpreted as; a watermark logo used for copyright protection and ownership verification. In case of formatting attacks, this technique is considered safe, but it has limited applicability because simple Optical Character Recognition (OCR) will ruin hidden information [9].

Rizzo *et al.* [10] proposed a technique that uses a password for the embedded watermark in short text while the contents are strictly preserved. When text changed into an image, the content and appearance cannot change. Blind watermarking is used to shows invisibility and content preserving properties. Tayan *et al.* [11] introduced a hybrid technique based on zero watermarking. The watermark is converted into an image as a string and then embedded in the cover file. It has one drawback that large storage is required to store Certified Authority (CA) keys. Thongkor and Amornraksa [12] propose a spatial domain image watermarking method for scanned and printed documents. The white, and blue components of the color image are used to embed the watermark. The performance is investigated on the bases of different scanning resolutions, printable materials, and quality, which shows that the proposed technique is imperceptible.

### B. LINGUISTIC-BASED APPROACHES

The linguistic-based approach consists of semantic and syntactic techniques which emphasize the semantic that is used for embedding the watermark and does not change the meaning of the text. Using a synonym substitution technique semantic approach is developed, which specifics that words are exchanged with their synonyms for data hiding.

In this technique, grammatical alternations are used for watermark embedding without affecting the original meaning

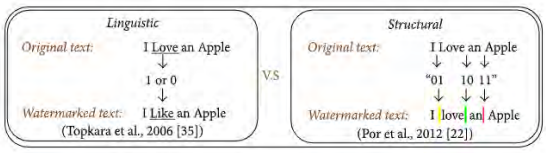



**FIGURE 2.** Comparison of linguistic and structural approaches [13].

of the text. The verb, adverb, noun, pronoun, adjective, preposition, acronyms, and conjunction are language parts which are used for the watermarking. The comparison of structural and linguistic-based approach is presented in Fig. 2. Liu *et al.* [14] proposed a method which is based on merging features of Chinese text sentences. Every word sentence entropy is calculated, and the weight of each sentence is also obtaining through entropy. The proposed method performs well in term of formatting attacks. Yingjie *et al.* [15] proposed a linguistic approach for text watermarking which based on characteristics of prose writings. Representative words are used to generate keyword, core verb set and proportional feature of adjectives. Verbs, adjective, noun, and adverb are used for embedding watermark. The proposed technique has a low embedding capacity.

### C. STRUCTURAL-BASED APPROACHES

In these techniques, essential bits are integrated into the structure or characteristic of the text. The spaces between lines and words are utilized for watermark embedding. This approach does not solve the problem of ownership authentication and for all types of text documents. If the spaces between words, lines, and paragraphs are removed then hidden data will ruined. The analysis of line shift coding is presented in Fig. 3, where the first group of three lines, the middle line is 1/300 inches shifted down. In Syntactic-based approach, this technique also retained all-natural textual features supported by the advances of Natural-language programming (NLP) techniques and resources. Taha *et al.* [17] suggest a technique for the Arabic language which utilizes the Kashida extension character and extra small whitespace for watermarking. The proposed technique is not robust against formatting attacks, because if the spaces between the words are removed, the hidden information will be lost.

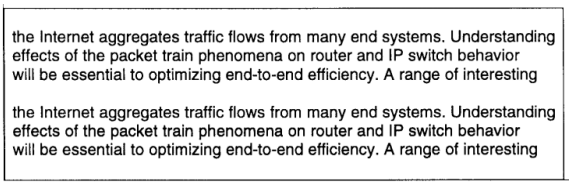



**FIGURE 3.** Example of line-shift coding [16].

Ba-Alwi *et al.* [18] introduced a new approach to zero text watermarking, which is based on the probabilistic model. The spacing between words and line are used for watermarking. As compared to other approaches, the proposed technique performs better in reordering attacks and robustness. Zhang *et al.* [19] proposed a novel method, where first the watermarking information is encrypted through the Caesar cipher with the user key then makes the groups of the message and packing into plain text. Through experiments, the proposed scheme is not imperceptibility. Liang and Iranmanesh [20] use a technique based on whitespaces between the words, which is not robust against formatting attacks and low embedding capacity. The main disadvantage of this is that large numbers of spaces are required to hide the secret message. Usmonov *et al.* [21] proposed a technique which aims to protect the data transferred between the logical, physical and IoT system virtual components, which combines the use of modern technologies of information security in the design and operation of IOT systems. Suciu *et al.* [22] presented an architecture for secure online health applications using IoT, Big Data and cloud convergence to enable remote monitoring. CloudView Exalead approach is used as a search platform which offers access to infrastructure-level information for both online and enterprise-based search applications.

Xiao *et al.* [23] suggest a structural approach based on Font-Code, where instead of changing text letters the glyphs of the fonts are used for embedding watermark. The proposed algorithm is robust and imperceptible but has Low capacity and only applicable for one font family. Large font size is required to detect the message and depending on the OCR library.

### D. HYBRID APPROACHES

A hybrid approach has been developed to combine different approaches to text watermarking. These techniques are considered robust and applicable to wide text documents [24]. Alotaibi and Elrefaei [25] introduced a method for Arabic text based on pseudo-space. The pseudo-space isolates connected letters are used for watermarking. The proposed method is imperceptible and robust against formatting and tampering attacks but cannot retain against retyping attack. Hamdan and Hamarsheh [26] proposed a new technique to hide information that conceals text messages in the text using Omega network structure. Zhang *et al.* [27] suggested the fragile watermark scheme to protect the integrity of the data in the IoT.

## III. WATERMARKING EMBEDDING AND EXTRACTION PROCESS

Digital Watermarking is often used to discourage illegal copying, and it is also used to stop the distribution of digital assets [42]. The architecture of text digital watermarking is shown in Fig. 4, the watermarking contains two phrases, embedding, and extraction of the watermark. The peace of secret information which is embedded into the original document is called watermark. The watermark embedding process includes three steps, first, watermark generation that includes the information about the owner, e.g., author name and other information like a publisher. Second, watermark securing, where the watermark is transformed into a binary

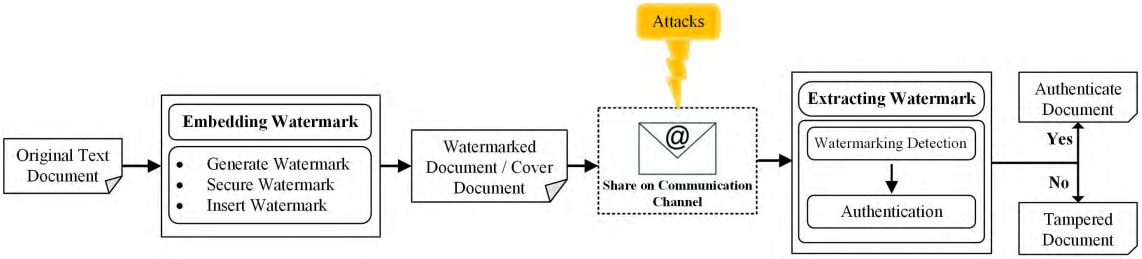



**FIGURE 4.** Architecture of digital text watermarking.

**TABLE 1.** The efficiency analysis of previous text watermarking techniques. Text watermarking existing techniques lack in capacity, not robust against the formatting attacks or not meet the requirements of imperceptible.

| # | Authors & Years | Medium | Capacity | Security | Imperceptibility | Robustness | Drawbacks |
|---|---|---|---|---|---|---|---|
| 1 | [26] - 2017 | Text | Low | High | Medium | High | The main drawback is the length (Capacity). |
| 2 | [28] - 2010 | Text | High | Medium | Low | Low | Not robust against formatting attacks. |
| 3 | [29] - 2003 | Text | High | NA | High | Low | Not robust against formatting attacks. |
| 4 | [30] - 2004 | Text | Low | High | High | Medium | Not robust against formatting attacks. |
| 5 | [31] - 2013 | Text | Low | Medium | High | Medium | Capacity and robustness issue. |
| 6 | [32] - 2010 | Text | Medium | Medium | High | High | Low capacity. |
| 7 | [33] - 2010 | Image plus Text | Low | High | Low | High | Low imperceptibility and capacity. |
| 8 | [34] - 2013 | Text | High | Low | High | Low | Not robust. |
| 9 | [35] - 2010 | Text | High | Medium | Low | Medium | Not robust nor imperceptible. |
| 10 | [36] - 2014 | Text | Medium | High | High | Medium | Not robustness against attacks. |
| 11 | [37] - 2011 | Text | Low | Medium | High | High | Low capacity. |
| 12 | [19] - 2010 | Text | High | High | Low | Medium | Not robust nor imperceptible. |
| 13 | [14] - 2015 | Text | Low | High | High | High | Low capacity. |
| 14 | [38] - 2014 | Text | High | Medium | Low | Low | Not robust nor imperceptible. |
| 15 | [20]- 2016 | Text | Low | Medium | High | Low | Capacity and robustness issue. |
| 16 | [25] - 2017 | Text | High | Medium | High | Medium | Not vulnerable to retyping attacks. |
| 17 | [15] - 2017 | Text | Low | High | Medium | High | Low capacity. |
| 18 | [39] - 2018 | Text | Low | High | Medium | High | Low capacity. |
| 19 | [40] - 2018 | Text | High | Medium | Low | High | Low imperceptibility. |
| 20 | [33] - 2010 | Image plus Text | Low | High | Low | High | Low imperceptibility and capacity. |
| 21 | [17] - 2018 | Text | High | Medium | Medium | Low | Not Robust against formatting attacks. |
| 22 | [23] - 2018 | Text | Low | High | Medium | High | Low capacity and only applicable to one font. |
| 23 | [41] - 2018 | Text | High | High | Low | Medium | Low imperceptibility. |

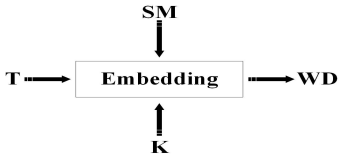



**FIGURE 5.** Watermark embedding process.

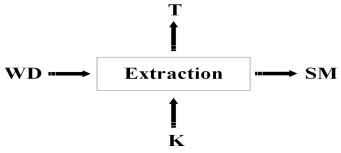



**FIGURE 6.** Watermark extraction process.

string or groups. The last one is inserting a watermark, where the watermark is inserted without affecting the whole document. The embedding process of watermarking is presented in Fig. 5, where ''SM'' denotes the secret message, ''T'' represents the original document, ''WD'' is a watermarked document and ''K'' denotes Key [43]. The watermarked document is shared via communication channels such as e-mail, website, and social media. The reverse process of watermark embedding is called extraction or verifying. Fig. 6 shows the extraction process of watermarking, where watermark document and key are provided as input and the secret message is extracted.

## IV. MATERIAL AND METHODS

In this section, we present our proposed model for text document copyright protection and ownership verification. A new framework as shown in Fig. 7, proposed to overcome the text watermarking current challenges. Data mining techniques are used to find the suitable properties of Microsoft Word (MS-Word) document.

Data mining aims to extract useful information from data. MS-Word document consists of a set of objects, where each object has many methods and attributes that permit users to interact with it and manipulate. MS-Word document has many properties in which we can store watermark information, and it cannot affect the whole document. Data mining main challenge is to find suitable properties which can incorporate a large size of the secret message. In the proposed method, we embed the watermark into different properties of MS-Word document. The proposed technique is robust, imperceptible, and incorporate large embedding capacity.

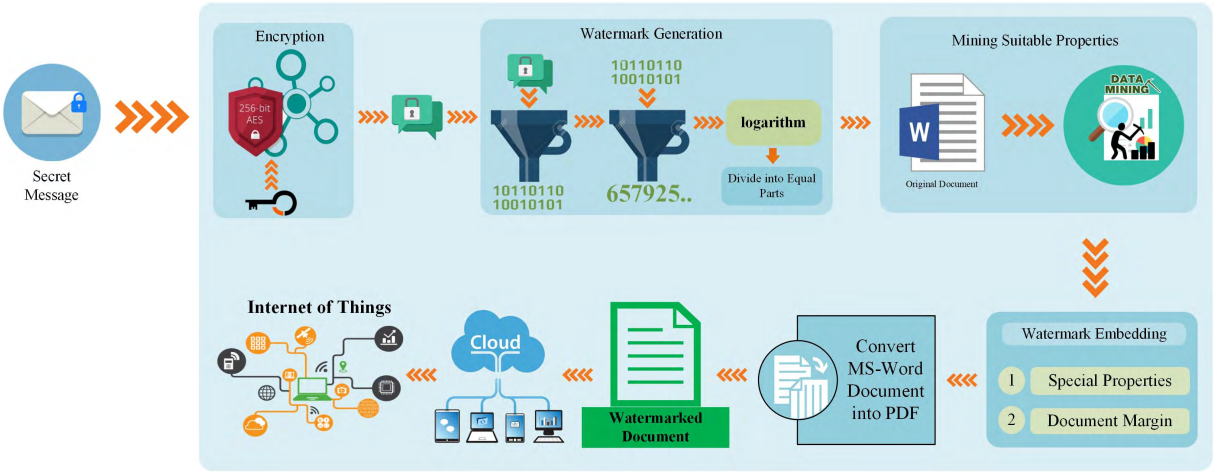



**FIGURE 7.** Proposed text watermarking model.

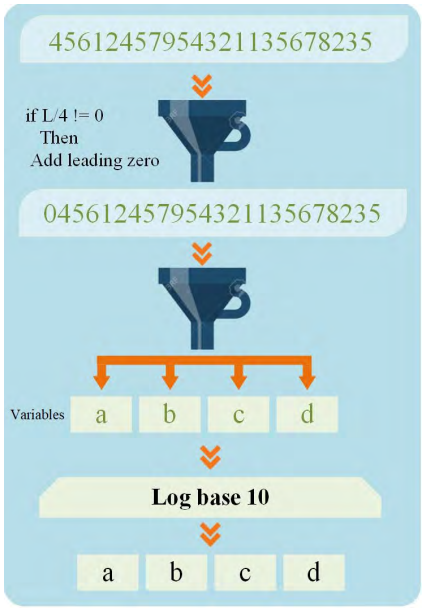



**FIGURE 8.** Watermark Group Generation.

After embedding the watermark, the watermarked document can store and share via the cloud. The authentication and accessibility of the original documents possible using smart devices on IoT.

### *A. WATERMARK EMBEDDING*

In this section, we discussed how watermark information embedded in the document. The secret message is encrypted AES algorithm using 256-bit; Encryption is applied to secure the message. The encrypted message is shifted into the next phase, where the watermark is generated. The encrypted message is converted into binary and then numbers. Algorithm 1 is applied to divide the numbers into four equal parts and store them into different variables (a, b, c, d). The length (L) of numbers is measured and then added zeroes at the beginning as shown in Fig. 8. Logarithm Base 10 is applied to further reduce the value of variables (a, b, c, d). An inverse function to the exponential that is used in mathematics to find the Anti-log. The complete procedure of the watermark group generation is described in Algorithm 1.

The logarithm of an "x" is the exponent to another number which is fixed, Base "b" need to be raised for the "x" number. The logarithm to the base 10 is natural logarithm which is applied using (1) and (2).

$$ln(x) = log_e(x) \tag{1}$$

$$e = lim_n(1+n)^n \tag{2}$$

The inverse of the logarithm also called anti-logarithm is calculated by raising the base "b" to the logarithm "y" using (3).

$$x = log^{(-1)}(y) = b^y \tag{3}$$

The purpose of using $\log_{10}$ (x) the output is always near 0 and 1 as shown in Fig. 9. After that, the original document taking as input and data mining is applied to find suitable properties from MS-Word document. MS-Word document contains a set of objects, where each object has a lot of attributes and methods. MS-Word document consists of two classes, one is application class and second is document class. Application class properties are modified through Visual Basic (VB) for embedding watermark, and it cannot affect the document class. The special properties of the MS-Word document are suitable for two reasons. First, a large amount of information is stored without affecting the whole document with imperceptibly. Second, any MS-Word mutual command will not affect the watermark information [19]. Table 2 presents suitable MS-Word properties, which are chosen for watermarking.

The watermark information divided into equal groups are embedded in these properties and then the second level

**Algorithm 1** Watermark Group Generation

**Input:** Secret Message (M)
**Output:** Watermarked Group $W_g$
**Start:**
**Data:** $W_n$, $D_n$, $W_l$, $W_g$, $W_{g1}$, $W_{g2}$, $W_{g3}$, $W_{g4}$, i
**Variable Declaration:**
$W_n$=Number String
$D_n$=No. of digits
$W_g$=Groups of digits
**Initialization:**
i=0, $W_{g1}$=0, $W_{g2}$=0, $W_{g3}$=0, $W_{g4}$=0, $W_{g4}$=0, $W_l$=0

```
for (Dn % 4 != 0 ) do
    0 + Wn
    Dn←Dn+1
end for
Wl ← Dn / 4
for (i = 1 to Dn ) do
    if (i < Wl ) then
        Wg1 = Wg1 + i
        if (i < 2 * Wl ) then
            Wg2 = Wg2 + i
            if (i < 3 * Wl ) then
                Wg3 = Wg3 + i
                if (i < 4 * Wl ) then
                    Wg4 = Wg4 + i
                end if
            end if
        end if
    end if
end for
Wg = ( Wg1,Wg2,Wg3,Wg4)
Wlog ← Log Wg 10
```

**end**

embedding starts. In second level embedding, MS-Word document margins from layout are targeted. The value of margin top, margin bottom, margin left, and margin right are modified and replaced with four variables respectively. The watermarked document is generated in Portable Document Format (PDF), and in the verification process when document format is changed the document margin cannot be altered and remain the same. When we convert MS-Word document into PDF or PDF to Word Document, the Margins and Layout of the document cannot be changed. After embedding watermark, MS-Word document is converted into PDF (Portable Document Format) and stored or share via the cloud. Algorithm 2 presents the complete procedure of watermark embedding.

### B. WATERMARK EXTRACTION

Watermarkextraction or verification process extract the watermark (secret information) from the atermarked

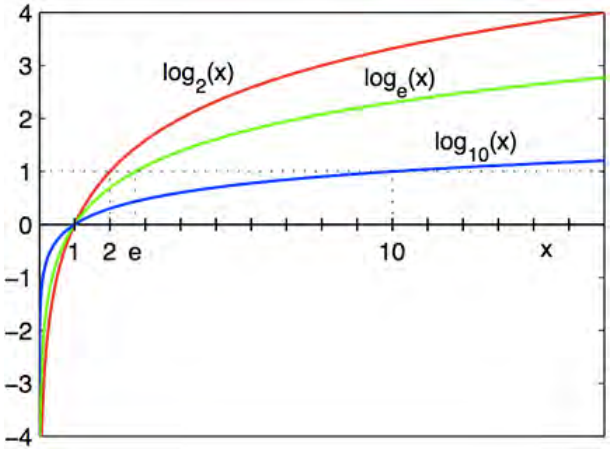



**FIGURE 9.** Logarithm functions commonly used [44].

**TABLE 2.** Microsoft word special properties.

| Object | Properties | Specification |
|---|---|---|
| Variable | Name | Macro settings in |
| | Value | between macro sessions. |
| Range | DisableCharacter | Read/write, Boolean. |
| | LanguageIDFarEast | Read/write, WdLanguageID. |
| | LanguageIDOther | Read/write, WdLanguageID. |
| | NoProofing | Read/write, Boolean. |
| Bookmark | Kerning | Read/write. Single. |

document. Watermark extraction is the reverse process of watermark embedding. A PDF file which is stored on the cloud given as input and system can convert it into MS-Word. In the first phase, the values are received from special properties, and anti-log is applied to retrieve the actual values. Four temporary variables are used to store values then concatenate all into sperate variable ''M''.

The system can automatically detect the top, bottom, left, and right margin of document layout and save them into separate variables like (''T'', ''B'', ''L'', ''R''). Anti-log is applied to retrieve the actual values and concatenate all into ''D''. Both variables ''M'' and ''D'' concatenate, and results are produced as number string. The number string is further converted into Binary's and then revert into characters. AES algorithm using 256bit that is used for encryption is applied to decryption the cover message. The same Key is used to decrypt the text which gives us the secret message which is hidden in the document. Algorithm 3 describes the complete process of watermark extraction or verification.

## V. RESULTS AND DISCUSSION

The evaluation criteria for text watermarking are classified into robustness, security, capacity, and imperceptibility. The proposed scheme is tested through series of experiments, the experiment setup includes: Core i3-3110 M CPU @ 2.40 GHz 2.40GHz RAM is 4.0GDDR, Window 10 operating system and development tools are VB 6.0 and MS-Word 2016. The watermark embedded in the document is ''This

**Algorithm 2** Watermark Embedding

**Input:** *Document (T), Secret message (SM), Key (K)*
**Output:** $\acute{T} \leftarrow$ Watermarked Document
**Start:**
**Data:** $E_m$, $B_s$, $N_n$, *WSP,SP*, $W_{log}$, $W_{log1}$, $W_{log2}$, $W_{log3}$, $W_{log4}$, $DM_{argin}$
**Variable Declaration:**
$E_m$=*Encrypted message*
$B_s$=*Binary string*
$N_n$=*Natural numbers*
*WSP=MS-Word special properties*
*SP=Suitable properties*
$DM_{argin}$=*Document margins*
$W_{log}$=*log variables*
**Initialization:**
SP=0
$E_m \leftarrow (SM, K)$
$B_s \leftarrow E_m$
$N_n \leftarrow B_s$
Data Mining (WSP)
**for** (*i to WSP* ) **do**
WSP = [SP]
SP $W_{log}$
**end for**
Check document margins($DM_{argin}$)
**Set** $DM_{argin}$
**Top** $W_{log1}$
**Left** $W_{log2}$
**Right** $W_{log3}$
**Bottom** $W_{log4}$
*Convert (T)* $\Longrightarrow$ *PDF*
*PDF* $\Longrightarrow$ *PDF* $\acute{T}$

**end**

**Algorithm 3** Watermark Extraction

**Input:** Watermarked Document ($\acute{T}$), Key (K)
**Output:** Secret message (M)
**Start:**
**Data:** $E_m$, $B_s$, $N_n$, *WSP, M, D, T, B, L, R*, $W_{log}$, $DM_{argin}$
**Variable Declaration:**
$E_m$=*Encrypted message*
$B_s$=*Binary string*
$N_n$=*Natural numbers*
*WSP=MS-Word special properties*
$DM_{argin}$=*Document margins*
$W_{log}$=*log variables*
Retrieve $\acute{T}$) from Cloud
PDF $\Longrightarrow$ MS-Word
Retrive from WSP
*Check(WSP)*
**for** (*i* $\Longrightarrow$ *WSP* ) **do**
$W_i$ = *[WSP]*
$M \leftarrow W_{log}$
**end for**
Check document margins($DM_{argin}$)
**T** $\Longrightarrow$ $DM_{argin}$ [Top]
**B** $\Longrightarrow$ $DM_{argin}$ [Bottom]
**L** $\Longrightarrow$ $DM_{argin}$ [Left]
**R** $\Longrightarrow$ $DM_{argin}$ [Right]
**D** $\Longrightarrow$ *T,B,L,R*
**Anto-log** $\Longrightarrow$ *D*
$E_m$=$M + D$
$N_n$=$E_m$
$B_s$=$N_n$
$C_{har}$=$B_s$
$D_m$=*AES* ($C_{har}$)
$SM$=$D_m$
*SM* $\Longrightarrow$ *Secret Message*

**end**

Document belongs to the University of Engineering and Technology, Taxila 45070, Punjab, Pakistan. Mr. Umair Khadam (Email: umair_khadim@live.com) is the original author of this document and have copyrights''.

### A. ROBUSTNESS

The content may experience several attacks of watermark earlier the watermark is recovered, an application that is distinct as any alteration in the content, which can damage the watermark [46]. The robustness of digital watermarking is computed using and Pattern Matching Rate (PMR), and Watermark Distortion Rate (WDR) which are formalized in (4) and (5).

$$PMR = \frac{No.\ of\ patterns\ matched\ correctly}{No.\ of\ watermark\ patterns} \quad (4)$$

$$WDR = 1 - PMR \quad (5)$$

The robustness of the proposed technique is tested, and the detection accuracy is calculated in each sample of text. The proposed technique is 99.9% robust against formatting attacks, as mention above MS-Word document special properties is utilized for watermark embedding. Any common MS-Word application command cannot interrupt the watermark. After applying different attacks on a watermarked document which includes cut, copy, paste, font size, font family and other changes as shown in Fig.10, the 100% watermark information is recovered from the document. Which shows the proposed technique is robust against the formatting attacks. The formatting attacks cannot interrupt or destroy the watermark. These attacks include font color, font families, font size, text background color, line spacing, and change case. As shown in Fig. 11, the proposed algorithm results are calculated which demonstrate that it is robust against formatting attacks with 99.9% detection accuracy. The proposed algorithm results are compared with the previous techniques.

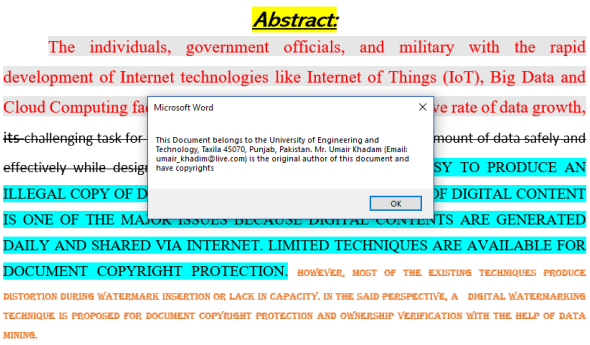



**FIGURE 10.** Applying formatting attack 100% watermark extracted.

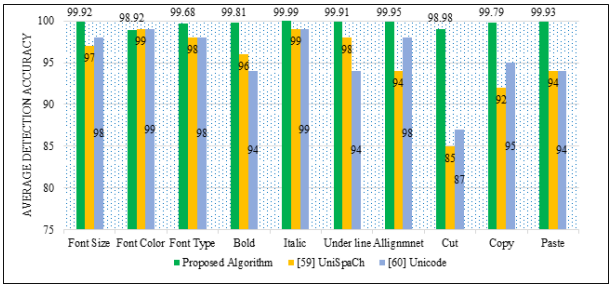



**FIGURE 11.** Formatting attacks under robustness.

### B. IMPERCEPTIBILITY

Imperceptibility is the primary and fundamental requirement of the watermark so that the watermark is embedded imperceptibly into the MS-Word document object. The watermark information could not affect the text. The original and watermark documents comparison are shown in Fig.13. Which illustrate that the proposed technique is imperceptible. The original text and watermarked text are the same because as mention above special properties are utilized for embedding the watermark information. To ensure the imperceptibility measure Peak Signal to Noise Ratio (PSNR) and Similarity (SIM) percentage of the proposed technique are calculated by using equation (6) and (7).

$$PSNR = 20\log_{10}\frac{O_{doc}(Max)}{RMSE} \tag{6}$$

$$SIM = [1 - \frac{RMSE}{O_{doc}\ (Max)} \times 100] \tag{7}$$

where Odoc(Max) is a maximum pixel value in the document image, Root Mean Squared Error (RMSE) is calculated equation (8).

$$PMSE = \sqrt{\frac{1}{m \times n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[O_{doc}\ (\mathbf{i},\mathbf{j}) - W_{doc}(\mathbf{i},\mathbf{j})]^2} \tag{8}$$

The proposed technique more imperceptible as compared to previous techniques. The MS-Word special properties did not affect the documents. The circos graph in Fig.12 represents the comparison of the proposed algorithm with Unicode [46] and UniSpaCh [47] on the bases of PSNR and

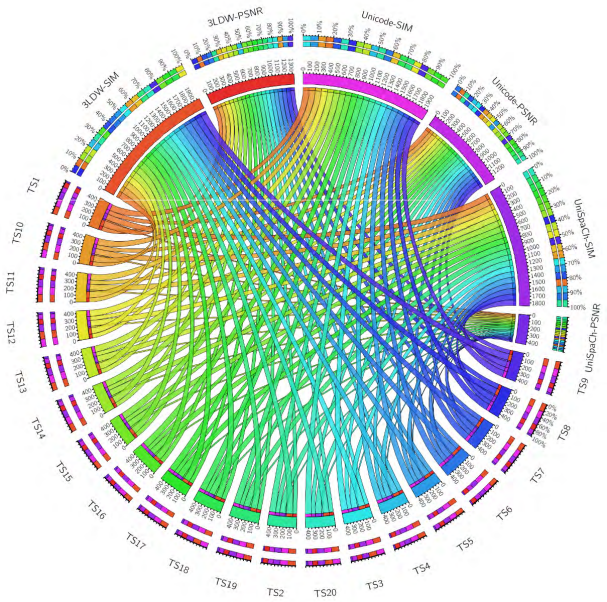



**FIGURE 12.** PSNR and SIM comparison with Unicode [45] and UniSpaCh [46].

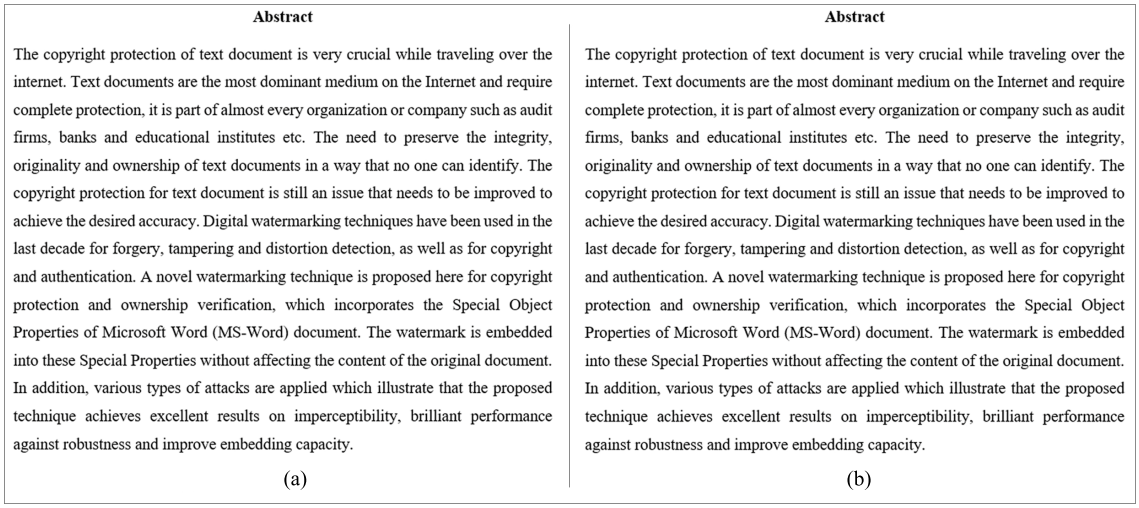
SIM. Twenty document samples with different variations are used for experiments to measure the PSNR and SIM of the proposed algorithm. In the proposed algorithm, the obtained Peak Signal to Noise Ratio (PSNR) values are between 64.67% and 71.03% and the Similarity (SIM) percentage is between 99.92% and 99.99%. The acceptable value of PSNR should be above 30 [47] and attained a high-level of imperceptibility. The result of the proposed algorithm is also compared with the previous techniques. Whereas UniSpaCh [47] introduced to much deterioration in the original document, that's why PSNR value is below than 30 and their SIM percentage between 88.96% and 93.13%. The PSNR value of Unicode [45] between 63.15 and 70.88 and their SIM percentage is between 99.93% and 99.97%.

### C. CAPACITY

The capacity of digital watermarking signifies the maximum amount of data can be stored and measured the length of a digital watermark. The capacity of the proposed technique is measured by (9).

$$Capacity = \frac{Total\ \text{no bits (Secret Data)}}{\text{Total no of cover file data (Kb)}} \times 100 \tag{9}$$

Capacity indicates the upper limit of watermark length. The proposed technique has a high capacity as compared with Liang *et al.* [20], where 197 characters are embedded. The proposed technique can embed 206 characters. The proposed technique improves the embedding capacity as compared with existing techniques, and the length of the watermark information can be seen in Fig.10. The capacity comparison of watermark information is presented in Table 3, where the proposed technique has a higher embedded capacity as compared with [17], [20], [27], [35]. The proposed technique



**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.

(a)

**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.

(b)

**FIGURE 13.** Comparison of original and watermarked Text.

**TABLE 3.** Capacity comparison of the proposed algorithm with existing techniques.

| Sr No | Authors | Words | Characters | Size in Bits |
|---|---|---|---|---|
| 1 | Cheng *et al.* [35] | 6 | 32 | 256 |
| 2 | Alotaibi *et al.* [27] | 21 | 130 | 1040 |
| 3 | Taha *et al.* [17] | 28 | 186 | 1490 |
| 4 | Liang *et al.* [20] | 32 | 197 | 1576 |
| 5 | Proposed technique | 28 | 206 | 1648 |

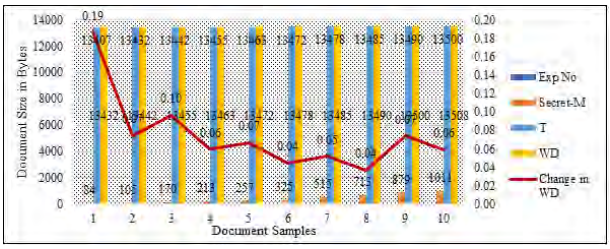



**FIGURE 14.** Capacity analysis document file size.

can embed up to 206 characters in the document. The series of ten experiments are performed with different text document sizes, in the first experiment original document size is 13,407 bytes, and 84 bytes of the secret message is embedded into the original document. After embedding the secret message, the watermarked document size is increased by 0.19. After embedding 1011 bytes of the secret message, the change in the watermark size is 0.06 which can be revealed in Fig. 14 the capacity of the secret message is increased, but the document file size has a slight change. The results are verified by conducting several experiments to check the performance of the proposed technique. The watermark information is extracted correctly after applying different attacks. We compared our results with a previous technique for robustness, imperceptibility, and capacity. We concluded that our technique gives better results and achieved high accuracy. The results demonstrate that it is suitable for document copyright protection and ownership verification and will be appropriate for text document security in smart cities.

## VI. CONCLUSION

A robust and secure watermarking algorithm is proposed to authenticate the digital contents in smart cities. The performance of the proposed technique is compared and verified with the previous techniques to confirm the imperceptibility, security, robustness, and capacity. Several techniques have been proposed in this field, but still need a technique which is applicable for the cloud, IoT devices, and smart cities. Through experiments, the proposed algorithm is highly imperceptible and achieve about 99.99 similarity factor. After applying formatting attacks such as cut, copy, paste, font size, font color, and alignment proposed algorithm proves that it is robust and tolerate most of possible attacks and watermark is extracted with high accuracy. The capacity of the proposed algorithm is also increased as compared with previous techniques. In the cloud computing environment, the proposed technique gives the same results which are suitable in smart cities to ensure the security of text documents. In the future, the proposed solution will be extended for the printed text documents copyright protection.

## Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016R1D1A1A09919551).

## REFERENCES

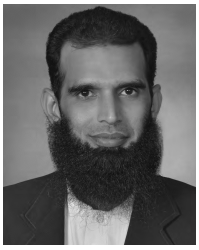
[1] M. Zeeshan, S. Ullah, S. Anayat, R. G. Hussain, and N. Nasir, "A review study on unique way of information hiding: Steganography," *Int. J. Data Sci. Technol.*, vol. 3, no. 5, p. 45, 2017.

[2] S. Huh, S. Cho, and S. Kim, "Managing IoT devices using blockchain platform," in *Proc. 19th Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2017, pp. 464–467.

[3] A. S. Panah, R. Van Schyndel, T. Sellis, and E. Bertino, "On the properties of non-media digital watermarking: A review of state of the art techniques," *IEEE Access*, vol. 4, pp. 2670–2704, 2016.

[4] M. Pal, "A survey on digital watermarking and its application," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 1, pp. 153–156, 2016.
[5] U. Khadim, A. Khan, B. Ahmad, and A. Khan, "Information hiding in text to improve performance for word document," *Int. J. Technol. Res.*, vol. 3, no. 3, p. 50, 2015.
[6] J. T. Brassil, S. Low, N. F. Maxemchuk, and L. O'Gorman, "Electronic marking and identification techniques to discourage document copying," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 8, pp. 1495–1504, Oct. 1995.
[7] J. Brassil, S. Low, N. Maxemchuk, and L. O'Gorman, "Hiding information in document images," in *Proc. Conf. Inf. Sci. Syst. (CISS)*, 1995, pp. 482–489.
[8] Y. Hao, Q. F. L. Chuang, and D. Rong, "A survey of digital watermarking," *J. Comput. Res. Develop.*, vol. 7, pp. 1093–1099, 2005.
[9] M. Kaur and K. Mahajan, "An existential review on text watermarking techniques," *Int. J. Comput. Appl.*, vol. 120, no. 18, pp. 1–4, 2015.
[10] S. G. Rizzo, F. Bertini, and D. Montesi, "Content-preserving text watermarking through unicode homoglyph substitution," in *Proc. 20th Int. Database Eng. Appl. Symp.*, 2016, pp. 97–104.
[11] O. Tayan, M. N. Kabir, and Y. M. Alginahi, "A hybrid digital-signature and zero-watermarking approach for authentication and protection of sensitive electronic documents," *Sci. World J.*, vol. 2014, Aug. 2014, Art. no. 514652.
[12] K. Thongkor and T. Amornraksa, "Digital image watermarking for printed and scanned documents," *Proc. SPIE*, vol. 10420, Jul. 2017, Art. no. 104203O.
[13] M. T. Ahvanooey *et al.*, "A comparative analysis of information hiding techniques for copyright protection of text documents," *Secur. Commun. Netw.*, vol. 2018, pp. 1–22, 2018.
[14] Y. Liu, Y. Zhu, and G. Xin, "A zero-watermarking algorithm based on merging features of sentences for Chinese text," *J. Chin. Inst. Eng.*, vol. 38, no. 3, pp. 391–398, Apr. 2015.
[15] M. Yingjie, L. Huiran, S. Tong, and T. Xiaoyu, "A zero-watermarking scheme for prose writings," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery (CyberC)*, Oct. 2017, pp. 276–282.
[16] J. T. Brassil, S. Low, and N. F. Maxemchuk, "Copyright protection for the electronic distribution of text documents," *Proc. IEEE*, vol. 87, no. 7, pp. 1181–1196, Jul. 1999.
[17] A. Taha, A. S. Hammad, and M. M. Selim, "A high capacity algorithm for information hiding in Arabic text," *J. King Saud Univ.-Comput. Inf. Sci.*, to be published.
[18] F. M. Ba-Alwi, M. M. Ghilan, and F. N. Al-Wesabi, "Content authentication of english text via Internet using zero watermarking technique and Markov model," *Int. J. Appl. Inf. Syst.*, vol. 7, no. 1, pp. 25–36, 2014.
[19] Y. Zhang, H. Qin, and T. Kong, "A novel robust text watermarking for word document," in *Proc. 3rd Int. Congr. Image Signal Process. (CISP)*, vol. 1, Oct. 2010, pp. 38–42.
[20] O. W. Liang and V. Iranmanesh, "Information hiding using whitespace technique in Microsoft word," in *Proc. 22nd Int. Conf. Virtual Syst. Multimedia (VSMM)*, Oct. 2016, pp. 1–5.
[21] B. Usmonov, O. Evsutin, A. Iskhakov, A. Shelupanov, A. Iskhakova, and R. Meshcheryakov, "The cybersecurity in development of IoT embedded technologies," in *Proc. Int. Conf. Inf. Sci. Commun. Technol. (ICISCT)*, 2017, pp. 1–4.
[22] G. Suciu *et al.*, "Big data, Internet of Things and cloud convergence—An architecture for secure E-health applications," *J. Med. Syst.*, vol. 39, no. 11, p. 141, 2015.
[23] C. Xiao, C. Zhang, and C. Zheng, "FontCode: Embedding information in text documents using glyph perturbation," *ACM Trans. Graph.*, vol. 37, no. 2, p. 15, 2018.
[24] Z. Jalil, "Copyright protection of plain text using digital watermarking," FAST Nat. Univ. Comput. Emerg. Sci., Islamabad, Pakistan, Tech. Rep. 1059, 2010.
[25] R. A. Alotaibi and L. A. Elrefaei, "Improved capacity Arabic text watermarking methods based on open word space," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 30, no. 2, pp. 236–248, 2018.
[26] A. M. Hamdan and A. Hamarsheh, "AH4S: An algorithm of text in text steganography using the structure of omega network," *Secur. Commun. Netw.*, vol. 9, no. 18, pp. 6004–6016, 2017.
[27] G. Zhang, L. Kou, L. Zhang, C. Liu, Q. Da, and J. Sun, "A new digital watermarking method for data integrity protection in the perception layer of IoT," *Secur. Commun. Netw.*, vol. 2017, Oct. 2017, Art. no. 3126010.
[28] A. A.-A. Gutub, F. Al-Haidari, K. M. Al-Kahsah, and J. Hamodi, "E-Text watermarking: Utilizing 'Kashida' extensions in arabic language electronic writing," *J. Emerg. Technol. Web Intell.*, vol. 2, no. 1, pp. 48–55, 2010.
[29] Y.-W. Kim, K.-A. Moon, and I.-S. Oh, "A text watermarking algorithm based on word classification and inter-word space statistics," in *Proc. ICDAR*, 2003, pp. 775–779.
[30] H. Yang and A. C. Kot, "Text document authentication by integrating inter character and word spaces watermarking," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, vol. 2, Jun. 2004, pp. 955–958.
[31] Y. M. Alginahi, M. N. Kabir, and O. Tayan, "An enhanced Kashida-based watermarking approach for Arabic text-documents," in *Proc. Int. Conf. Electron., Comput. Comput. (ICECCO)*, Nov. 2013, pp. 301–304.
[32] Y. Meng, T. Guo, Z. Guo, and L. Gao, "Chinese text zero-watermark based on sentence's entropy," in *Proc. Int. Conf. Multimedia Technol. (ICMT)*, Oct. 2010, pp. 1–4.
[33] Z. Jalil and A. M. Mirza, "Text watermarking using combined image-plus-text watermark," in *Proc. 2nd Int. Workshop Educ. Technol. Comput. Sci. (ETCS)*, vol. 1, Mar. 2010, pp. 11–14.
[34] R. J. Jaiswal and N. N. Patil, "Implementation of a new technique for web document protection using unicode," in *Proc. Int. Conf. Inf. Commun. Embedded Syst. (ICICES)*, Feb. 2013, pp. 69–72.
[35] W. Cheng, H. Feng, and C. Yang, "A robust text digital watermarking algorithm based on fragments regrouping strategy," in *Proc. IEEE Int. Conf. Inf. Theory Inf. Secur. (ICITIS)*, Dec. 2010, pp. 600–603.
[36] N. Mir, "Copyright for web content using invisible text watermarking," *Comput. Hum. Behav.*, vol. 30, pp. 648–653, Jan. 2014.
[37] Y. Meng, L. Gao, X. Wang, and T. Guo, "Chinese text zero-watermark based on space model," in *Proc. 3rd Int. Workshop Intell. Syst. Appl. (ISA)*, May 2011, pp. 1–5.
[38] Y. M. Alginahi, M. N. Kabir, and O. Tayan, "An enhanced Kashida-based watermarking approach for increased protection in Arabic text-documents based on frequency recurrence of characters," *Int. J. Comput. Elect. Eng.*, vol. 6, no. 5, p. 381, 2014.
[39] Q. Wen, Y. Wang, and P. Li, "Two Zero-Watermark methods for XML documents," *J. Real-Time Image Process.*, vol. 14, no. 1, pp. 183–192, 2018.
[40] M. Kuribayashi, T. Fukushima, and N. Funabiki, "Data hiding for text document in PDF file," in *Advances in Intelligent Information Hiding and Multimedia Signal Processing*, vol. 81. Cham, Switzerland: Springer, 2018, pp. 390–398.
[41] L. Tan, K. Hu, X. Zhou, R. Chen, and W. Jiang, "Print-scan invariant text image watermarking for hardcopy document authentication," *Multimedia Tools Appl.*, pp. 1–23, 2018.
[42] M. Topkara, G. Riccardi, D. Hakkani-Tür, and M. J. Atallah, "Natural language watermarking: Challenges in building a practical system," *Proc. SPIE*, vol. 6072, Feb. 2006, Art. no. 60720A.
[43] N. A. S. Al-Maweri, R. Ali, W. A. W. Adnan, A. R. B. Ramli, and S. M. S. A. A. Ahmad, "State-of-the-art in techniques of text digital watermarking: Challenges and limitations," *J. Comput. Sci.*, vol. 12, no. 2, pp. 62–80, 2016.
[44] T. Becker *et al.*, "Benford's law and continuous dependent random variables," *Ann. Phys.*, vol. 388, pp. 350–381, Jan. 2018.
[45] N. A. S. Al-Maweri, W. A. W. Adnan, A. R. Ramli, K. Samsudin, and S. M. S. A. A. Rahman, "Robust digital text watermarking algorithm based on unicode extended characters," *Indian J. Sci. Technol.*, vol. 9, no. 48, pp. 1–14, 2016.
[46] L. Y. Por, K. Wong, and K. O. Chee, "UniSpaCh: A text-based data hiding method using unicode space characters," *J. Syst. Softw.*, vol. 85, no. 5, pp. 1075–1082, 2012.
[47] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electron. Lett.*, vol. 44, no. 13, pp. 800–801, Jun. 2008.
[48] L. Tan, K. Hu, X. Zhou, R. Chen, and W. Jiang, "Print-scan invariant text image watermarking for hardcopy document authentication," *Multimedia Tools Appl.*, pp. 1–23, 2018.

**UMAIR KHADAM** received the B.S. degree in computer science from the Department of Computer Science, University of Azad Jammu and Kashmir, Pakistan, in 2011, and the M.S. degree in computer science from IQRA University Islamabad, Pakistan. He is currently the Ph.D. degree with the Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan. He has been actively involved in teaching and research activities throughout his university educational period. His research interests include digital watermarking, data mining, and the Internet of Things.



**MUHAMMAD MUNWAR IQBAL** received the B.Sc. degree in double math and chemistry from Islamia University Bahawalpur, the M.Sc. degree in computer science from the University of the Punjab, Lahore, the M.S. degree in computer science from the COMSATS Institute of Information Technology, Lahore, in 2011, and the D.Phil. degree from the Department of Computer Science & Engineering, University of Engineering and Technology, Lahore, Pakistan, under the supervision of Dr. Y. Saleem. He was an Associate Professor with the Department of Computer Science & Engineering, University of Engineering and Technology at Lahore. He is currently an Assistant Professor with the Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan. He is bearer responsibilities in the Computer Science Department as the Director Academic Cell, the Head of the Semester Committee, the Head of the Scholarship Committee, HEC Laptop Scheme (focal person), Security Focal Person, Exam Scrutiny Committee, and Curriculum Revision Committee, and an Advisor COMPTECH Society, BSCS 2016 session, and Prospectus Amendment Committee. His interests include machine learning, databases, semantic web, e-learning, artificial intelligence, the Internet of Things, information-centric networking, ambient intelligence, wireless sensor, machine learning, databases, data science, data mining, semantic web, social media analysis, and artificial intelligence.

**MUHAMMAD AWAIS AZAM** was the Head of the Academics in Cromwell College of IT & Management, London, U.K. He is currently an Assistant Professor with the Computer Engineering Department, University of Engineering and Technology, Taxila. He is currently active research collaborations in technically advance countries including, U.K., USA, South Korea, and Germany, to find the innovative and remarkable solutions for the problems in hand. His research interests include network architecture, the IoT, network security, ambient intelligence, wireless communications, opportunistic networks, and recommender systems. He has received several national and international funding grants for projects and presenting his research at different international forums. He has also served as a Reviewer for various peer reviewed journals and technical program committee member of international conferences.

**SHEHZAD KHALID** received the degree from the Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan, in 2000, the M.Sc. degree from the National University of Science and Technology, Pakistan, in 2003, and the Ph.D. degree from the University of Manchester, U.K., in 2009. He is currently a Professor and the Head of Department with the Department of Computer Engineering, Bahria University, Pakistan. He is also a qualified Academician and a Researcher with more than 60 international publications in various renowned journals and conference proceedings. He is also the Head of the Computer Vision and Pattern Recognition Research Group which is a vibrant research group undertaking various funded research projects. His research interests include but are not limited to shape analysis and recognition, motion based data mining and behavior recognition, medical image analysis, ECG analysis for disease detection, biometrics using fingerprints, vessels patterns of hands/retina of eyes, ECG, Urdu stemmer development, short and long multi-lingual text mining, and Urdu OCR. He was a recipient of the Best Researcher Award for the year 2014 from Bahria University, the Letter of Appreciation for Outstanding research contribution, in 2013, and the Outstanding Performance Award, from 2013 to 2014. He was the Reviewer for various leading ISI indexed journals.

**SEUNGMIN RHO** received the B.Sc. degree in computer science from Ajou University, South Korea, in 2001, and the M.Sc. and Ph.D. degrees in information and communication technology from the Graduate School of Information and Communication, Ajou University, in 2003 and 2008, respectively. Before he joined the Computer Sciences Department, Ajou University, he spent two years in industry. He visited the Multimedia Systems and Networking Laboratory, The University of Texas at Dallas, from 2003 to 2004. From 2008 to 2009, he was a Postdoctoral Research Fellow with the Computer Music Lab, School of Computer Science, Carnegie Mellon University. From 2009 to 2011, he was a Research Professor with the School of Electrical Engineering, Korea University. In 2012, he was an Assistant Professor with the Division of Information and Communication, Baekseok University. From 2013 to 2018, he was an Assistant Professor with the Department of Media Software, Sungkyul University. He is currently a Faculty of the Department of Software, Sejong University, South Korea. His current research interests include database, big data analysis, music retrieval, multimedia systems, machine learning, knowledge management, and computational intelligence. He has published more than 180 papers in refereed journals and conference proceedings in these areas. He has been involved in more than 20 conferences and workshops as various chairs and more than 30 conferences/workshops as a program committee member.

**NAVEEN CHILAMKURTI** received the Ph.D. degree from La Trobe University, Melbourne, Australia, where he is currently a Senior Lecturer with the Department of Computer Science and Computer Engineering. He has published about 100 journal and conference proceeding papers. He currently serves on the editorial boards of several international journals. His current research areas include wireless multimedia, wireless sensor networks, nanocommunications, vehicle-to-infrastructure and vehicle-to-vehicle communications, multicast congestion control, multicast security, transmission control protocol/internet protocol congestion control, and crosslayer techniques. He is an Inaugural Editor-in-Chief for the *International Journal of Wireless Networks and Broadband Technologies*.