IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 20, 2017, accepted January 16, 2018, date of publication January 24, 2018, date of current version March 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2796585

A Review of Text Watermarking: Theory, Methods, and Applications

NURUL SHAMIMI KAMARUDDIN[1], AMIRRUDIN KAMSIN[1], LIP YEE POR[1], AND HAMEEDUR RAHMAN[2]

[1]University of Malaya, Kuala Lumpur 50603, Malaysia
[2]University of Kebangsaan Malaysia, Bangi 43000, Malaysia

Corresponding authors: Nurul Shamimi Kamaruddin (shamimi.k@siswa.um.edu.my), Amirrudin Kamsin (amir@um.edu.my), and Lip Yee Por (porlip@um.edu.my)

This work was supported by the University of Malaya from 2016 to 2017 through the Fundamental Research Grant under Project FP003-2016.

ABSTRACT During the recent years, the issue of preserving the integrity of digital text has become a focus of interest in the transmission of online content on the Internet. Watermarking has a useful tool in the protection of digital text content as it solves the problem of tampering, duplicating, unauthorized access, and security breaches. The rapid development currently observable in information transfer and access is the consequences of the widespread usage of the Internet. When it comes to the different types of digital data, text constitutes the most complex and challenging type to which the method of text watermarking can be applied. Text watermarking constitutes a highly complex task, most of all, since only limited research has been done in this field. In order to ensure the successful evaluation, analysis, and implementation, a comprehensive research needs to be performed. This paper studies the theory, methods, and applications of text watermarking, which includes the discussion on the definition, embedding and extracting processes, requirements, approaches, and language applications of the established text watermarking methods. This paper reviews in detail the new classification of text watermarking, which is through embedding process and its related issues of attacks and language applicability. Open research challenges and future directions are also investigated, with a focus on its information integrity, information availability, originality preservation, information confidentiality, protection of sensitive information, document transformation, cryptography application, and language flexibility.

INDEX TERMS Information protection, information security, text analysis, text watermarking, watermarking.

I. INTRODUCTION

The rapid developments currently observable in the field of information technology in the form of storage devices, digital content and communications has created a vast electronic environment with the ability to transmit, duplicate, copy and distribute information through digital media without any loss of quality. Nevertheless, this technological revolution in the online propagation of digital multimedia also suggests that such data are vulnerable to attacks, unauthorized access, and other threats [1]. Hence the study of information security which includes not just encryption but also traffic security whose essence lies in hiding data are increasing in demand [2].

Digital watermarking belongs to the branch of information hiding methods. It involves the idea of using an algorithm to hide copyright information by embedding it in the digital data. This copyright information can be in the form of text, image or logo chosen by the owner. Digital watermarking seems to be the most suitable application to protect intellectual property rights, identify ownership, keep track of digital media content, and ensure authentication and security [3], [4]. Watermarking aims at protecting the rights of the owners of digital media. Even if an unauthorized copy is made or minor modifications are made to the watermarked file, the owner can still prove his original ownership. The goal of watermarking is to protect the cover file itself [5].

As of to date, digital data have come to invade almost all types of public media, be they in the form of audio, video, image and text. Among the media of digital content, text is the least discussed topic when it comes to information hiding. A digital text consists of any type of textual content in the form of monographs, articles, and webpages. These uploaded texts are prone to all kinds of attacks and copying, which in turn makes their security and protection a crucial matter. The protection of digital text has become difficult considering the lack of available techniques in handling the security of digital content [6]. Thus, it is very crucial to take control of the intellectual copyright over the text content, by first analyzing its nature of implementation and the theory behind it. Differences in text watermarking methods also contribute to the level of effectiveness in the protection of digital text [7].

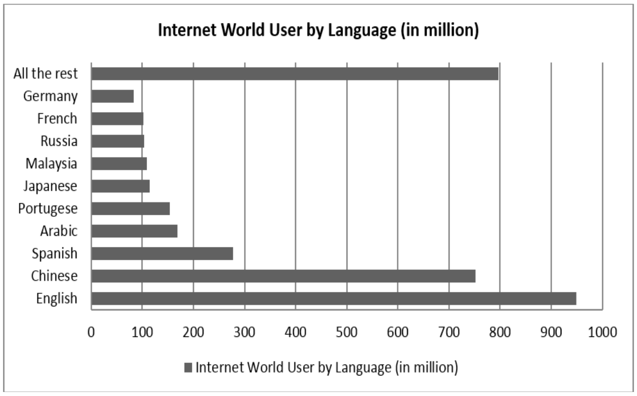

FIGURE 1. Top 10 languages in the Internet in millions of users [8].

The rate of information accessible over the Internet is staggering. Of a considerable challenge posed to researchers and practitioners is that this growth rate exceeds their ability to design appropriate text watermarking techniques in order to protect the text based on the language used. It is crucial to preserve the integrity of the data content and at the same time ensure the confidentiality and availability of information. The rise of the Internet has resulted in an outgrowth of many languages including English, Chinese, Spanish, and Arabic [8] (see Fig. 1). However, the one method of text watermarking that's applicable for text of any languages has yet to be studied thoroughly and in depth. Many of the proposed text watermarking focused on certain types of language due to the differences in the nature and properties of its text. More general text watermarking techniques that can be applied to any types of text needs to be developed and tested in order to overcome the difficulty and vulnerability in information transfer and information hiding due to language barrier.

Thus, the goal of this study is to carry out a comprehensive investigation of the current status of the development of text watermarking, which includes its theory, methods and applications. This comprises of the discussion on the text watermarking definition, embedding and extraction processes, requirements, approaches, and applications based on the languages used in the watermarking methods. This article also reviews in detail the new classification of text watermarking; which is according to the embedding process. Also discussed is the relationship between the embedding process against its related attacks and language applicability. Furthermore, open research challenges that require substantial research efforts are investigated, with focus on information integrity, information availability, originality preservation, information confidentiality, protection of sensitive information, document transformation, cryptography application, and language flexibility.

The remainder of this paper is organized as follows. Section 2 presents the definition, overview and process of embedding and extracting the watermark. Section 3 outlines the currently available approaches of text watermarking. The requirements and measurements for text watermarking are presented in Section 4. Section 5 presents the main findings related to the logical and physical embedding process. The applications of text watermarking based on the three most widely used languages are discussed in Section 6. Section 7 summarizes the classifications done in the form of taxonomy. Several issues and research challenges are discussed in Section 8 followed by the conclusions of this research in Section 9.

II. EMBEDDING AND EXTRACION OF WATERMARK

Watermarking is an information hiding method that can be used to hide digital data into a cover file such as text, audio, image, and video [9]. Besides hiding information, watermarking can be used as a reliable method to preserve the originality of data because the hidden data can be extracted and be used to validate its ownership [2], [10], [11].

Before a digital data (X) can be embedded, a key (K_1) is needs to be generated. After the key generation, X can be embedded with a watermark (W) using K_1. The embedding process is completed once the watermarked data (X') is produced (see Fig. 2).

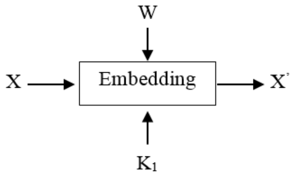

FIGURE 2. Process of embedding a digital data.

In general, there are three types of extraction methods, which are blind, semi-blind, and non-blind watermarking extraction methods [9].

A. BLIND WATERMARKING

In blind watermarking (also known as public watermarking), the extraction can be accomplished even in the absence of the embedded data (X) and the watermarked data (X'). The blind watermarking technique does not require the original data to detect the watermark [9]. It extracts the bits of the watermark (W) from the watermarked data (X') and produces the extracted watermark (W') for verification (see Fig. 3).

B. SEMI-BLIND WATERMARKING

In semi-blind watermarking (or semi-private watermarking), the extraction can be achieved without the presence of X,

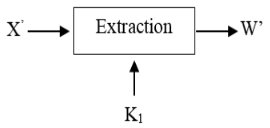

FIGURE 3. Process of blind extraction from the embedded watermark.

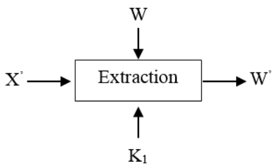

FIGURE 4. Process of semi-blind extraction from the embedded watermark.

but needs the presence of X' and W (see Fig. 4). Semi-blind watermarking is used to see whether the watermark can be detected [12].

C. NON-BLIND WATERMARKING

Using the non-blind watermarking method (or private watermarking), the presence of X is required [13]. The embedded watermark can be extracted from X' using K$_1$ (see Fig. 5). The extracted watermark (W') can be compared with W. If both watermarks are similar, the ownership is determined; otherwise the embedded data was altered.

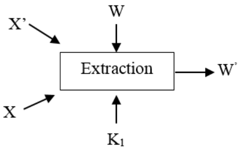

FIGURE 5. Process of non-blind extraction from the embedded watermark.

In terms of mathematical equation, the embedding and extraction process of watermark can be denoted as follows:

I. EMBEDDING : $E_M\ (K_1, W, X) \rightarrow X'$

II. BLIND EXTRACTION : $E_X\ \left(K_1, X'\right) \rightarrow W'$

III. SEMI-BLIND EXTRACTION : $E_X\ \left(K_1, W, X'\right) \rightarrow W'$

IV. NON-BLIND EXTRACTION :$E_X \left(K_1, W, X, X'\right) \rightarrow W'$

Where;

K_1: key
X: data
X': watermarked data
W: watermark
W': extracted watermark
E_M: embedding
E_X: extraction

III. APPROACHES USED IN TEXT WATERMARKING

Research on text watermarking began in 1997 where several text watermarking approaches were proposed for encoding information in text documents and copyright protection [14]. This initial research has prompted other researchers to give more serious attention to the study of text watermarking. These early proposed approaches of text watermarking included structural based watermarking, where the line, letters and spaces are shifted to embed the bits of the watermark [15]. Later, linguistic based watermarking approach was proposed [16]. For this method the language of the text is analyzed and edited to embed watermark bits. Since then a lot of other innovative approaches have been proposed and classified [17]–[19], [7]. In this study, we examine and evaluate the structural-based, the linguistic-based, and the image-based approaches.

A. STRUCTURAL-BASED APPROACH

The structural-based approach alters the structure or feature of the text in order to embed the necessary bits. It involves the general formatting of the properties of the cover text, in which the text content is modified using its words or sentences to hide the watermark information. The locations of words and letters or the writing style can also be altered to hide watermark bits. This includes repeating some letters or altering the features of the text. The general properties of the text are studied, and certain physical properties in the text layout are utilized in order to hide the watermark bits.

The method of shifting the words and sentences upwards or downwards in order to embed watermark bits was first proposed by Brassil [15], [20], [21]. It may consist of a line shift algorithm in which the sentence moves upwards or downwards, a word shift algorithm that moves the words horizontally or feature coding where the feature of a certain text is altered in order to embed the watermark bits. A number of slightly different methods were subsequently proposed in order to improve the methods initially developed by Brassil. Analyzing the average word distance in each line also constitutes one of the structural-based methods that are being applied in watermarking, where the distance for the embedded watermark is based on some formulas [22]. Other methods using an algorithm based on word classifications has also been developed [23]. It is classified according to its features and based on class labels of the word within a segment. Another algorithm that has been proposed exploits the justified paragraphs and irregular spacing contained in the text in order to embed the desired watermark information [24].

A number of other new and innovative approaches have been proposed such as the method proposed by Jalil et al. [25] (2010). Their algorithm utilizes the content of the text in order to embed the watermark. A keyword is chosen by the text's author, and the watermark is generated based on the length of the preceding and the following word; to and from the keyword occurrence in the text. This process is illustrated in Fig. 6, where 'is' constitutes the keyword, and the watermark is generated based on the text content.

New methods for text watermarking for languages other than English, for example Chinese and Arabic, have also

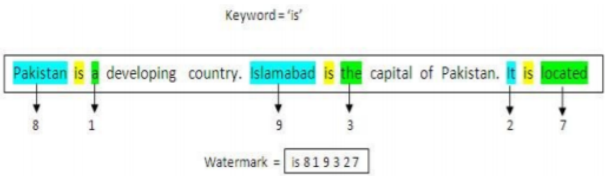
Keyword='is'
Pakistan is a developing country. Islamabad is the capital of Pakistan. It is located
8
1
9
3
2
7
Watermark = is 8 1 9 3 2 7

FIGURE 6. Watermark generation using word context [25].

found their way into text watermarking. The first Chinese text watermarking structural-based approach was developed by Li and Dong (2008) who proposed a method that utilized the features of the pictographic characters in Chinese script. In this algorithm, the Chinese pictographic characters are utilized by splitting the character, converting the selected character's coding, and then extending the watermark redundancy space [26]. Other Chinese text watermarking methods also make use of the properties of Chinese characters as they offer themselves as ideal objects for hiding information. Also, a watermarking algorithm can be constructed by adopting the pronunciation of polyphone characters and the features between two polyphones of Chinese characters [27]. Other method uses Chinese sentences entropy, where it is calculated based on word frequency and makes crucial sentence selection based on entropy [27], [28].

In the case of watermarks to be embedded in Arabic text content, the majority of the currently developed methods utilize the special text feature of the Arabic script characters. The characters and words are extended, letters are replace with identical character of different Unicode or the diacritics are manipulated to embed watermark bits [29]–[32].

The aforementioned structural-based algorithms use the text structure and features to manipulate the watermark bits and add them to the text. Those methods are not resistant to formatting-based attacks including copying and pasting, OCR and retyping. Some methods are resistant to printing and font changing, which depends on the technique's capability. In the case of content-based attacks, it also depends on the robustness of the proposed technique.

B. LINGUISTIC-BASED APPROACH

The linguistic-based approach is a natural language-based method, which works by making changes to the syntactic and semantic nature of the cover text in order to embed the watermark [33], [34]. The watermark is embedded in such a way that the structure and meaning of the text remain unchanged. Most of the linguistic-based watermarking methods make use of the semantic or syntactic transformation or a combination of both, depending on the language of the text [35].

According to the syntactic approach, words in the set are manipulated to hide data. The verbs, nouns, adjectives, pronouns, prepositions, synonyms, and other grammatical features of the text content are utilized in order to hide the watermark message. These grammatical alterations are done without affecting the original meaning of the text. The order of the words in the sentences can also be rearranged to hide bits. This can be done by altering the text structure and embed the watermark, such as moving the adverbial phrase, adding the subject or changing the sentence from active into passive clause. One of the early methods following the syntactic-based approach uses syntactic tree and transformation [16], [36], [37]. This pioneer technique has inspired other researchers to create other equally unique and reliable methods. Many other syntactic-based methods have since then been proposed, thus expanding the watermarking technique [38]–[42].

Following the semantic-based approach, the data are hidden by manipulating the words in the text as watermark. This includes methods such as synonym substitution, algorithms based on noun-verbs, algorithms based on typos, acronyms and abbreviations, algorithms based on linguistic approaches of presuppositions, and algorithms based on text-meaning representational strings. This line of approach depends on the type of language itself by using its vocabulary, grammar or structure in order to hide the watermark [33], [43]–[45].

C. IMAGE-BASED APPROACH

The image-based approach falls under both text and image watermarking categories. In the image watermarking, the text content is understood as a series of text images, where the watermark image or logo is embedded inside the cover text. The watermarked text is this regarded as a picture, and the text can no longer be copied and pasted and has to be retyped in order to be reproduced. As for the image-based approach in text watermarking, the watermark image or logo is converted into text string and paired with the characteristics of the cover text in order to generate watermark data. This string is embedded into the cover text using certain algorithms. In order to check whether the text has been tampered with or not, a watermark logo is used to generate a key and identify the original data ownership. This method of watermarking is considered safe from attacks on image watermarking as well as format-based attacks, depending on the algorithm used.

In recent years, the image-based approach in text watermarking has been proposed. One method is to use the pixels of letters in order to embed the watermark [46]. In this study, curvaceous letters are used for watermarking by changing the letters' curves according to respective bits. The process of bit embedding entails the curve of the watermarked letter changing slightly in direct relation to the bit that has been modified. The curve parameter of the letter is changed depending on the bits (see Fig. 7).

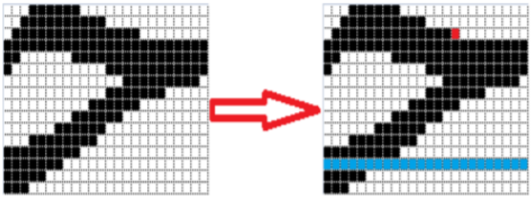

FIGURE 7. The image based watermarking, which changes the pixel of image to embed watermark [46].

IV. EVALUATION OF TECHNIQUES

The many issues that arise in the study of digital text watermarking techniques can be assessed and evaluated based on their requirements [2], [47]–[50]. The requirements for watermarking can be measured using the four variables; robustness, capacity, security, and imperceptibility. These requirements analyze the performance of the different text watermarking techniques. Below are the descriptions of the text watermarking requirements.

A. ROBUSTNESS

The robustness of a system can be defined as the ability and strength of the technique to resist any attack that aims at retrieving or modifying the hidden data [51]. The watermark data must be securely embedded and retrieved, so that they can survive any attack on the watermark. The watermark must withstand all attacks and must be retrievable for verification. Robustness constitutes an essential requirement in order to proof of ownership, copy control, identification, and fingerprinting where conceivable distortion is inevitable and there is a concern about the removal of the watermark. The most important feature of robustness is that even if a third party can detect the watermark, it cannot be destroyed without distorting or altering the cover file.

System robustness can be divided into three categories: robust, fragile, and semi-fragile based on the level of modification done on the hidden data [52], [53]. In the robust technique, the data are protected and hidden from unauthorized access and modification. On the other hand, a fragile data hiding system is not robust and cannot withstand an attack on the system, and a semi-fragile technique means that the hidden data can only resist some attacks. Thus, robustness can be used to measure and differentiate between good and bad text watermarking by assessing how many attacks it can resist.

No exact means are available to measure the robustness. However, in the case of text watermarking, the attacks attempted on breaking the watermarked data can be used in order to measure the robustness. The two types of attack that are usually directed at watermarked data are formatting and content attacks [6], [29], [34], [54]. Therefore, if a system is able to prevent those attacks, it is considered a robust system. If it can only prevent half or none of the attacks, it is considered a semi or non-robust system.

1) FORMATTING-BASED ATTACKS

1) *Copy and paste*
2) *Printing*
3) *OCR*
4) *Retyping*
5) *Font changing*

2) CONTENT-BASED ATTACKS

1) *Insertion/addition attack*
   Add some word into the text, so that the generated watermark is different from the original watermark.
   Localized insertion attack: A single word or sentence is added at random in the text content.
   Dispersed insertion attack: Multiple words or sentences are added at random in the text content.
2) *Deletion attack:*
   Some words or parts of the text content are deleted, which distorts the watermark in the text.
3) *Reordering attack:*
   Words in the text are rearranged without modifying the meaning.
4) *Syntactic transformation attack:*
   The text is transformed by altering the sentences without altering their meaning, for example by changing the preposition, the tense, passive and active and other syntax elements.
5) *Synonym substitution attack:*
   A certain word is substituted by a synonym, which does not alter the meaning of the text.
6) *Replacement attack [55]:*
   This is a new geometrical attack changing the content of a document without destroying the text structure. It maintains the location of the words in the text and is almost similar to the synonym substitution attack, yet uses more advanced techniques in selecting the word to make it invisible.

B. IMPERCEPTIBILITY

Imperceptibility is the ability of the watermark to remain unnoticed by the naked eye. This is achieved by the high level of similarity between the original data and the watermarked data. The original data should be almost the same after the data are hidden and should not degrade its content. It aims at adding invisible data hidden in the text, preserve its integrity and as copyright protection. Although it is common that watermarked data is imperceptible, digital watermark embedded can be visible or invisible [17], [56]. In visible watermarking, the embedded watermark is obvious and can be detected by the naked eye. This is usually done in images and videos in order to prevent their unauthorized distribution. The watermark signals clearly that the document is owned and that any form of tempering and copying is not permissible. Text content can be easily manipulated and copied, and therefore, the invisibility of the watermark data is important. The invisible watermark is embedded in the document in such a way that it cannot be seen by the naked eye. It can, however, be detected by algorithm.

The mathematical expressions as shown below signify that imperceptibility can be formulated in such a way that the watermarked data are approximately the same after being tampered with.

$$X \cong X'$$

Where;

X: data
X': watermarked data

C. CAPACITY

Capacity is the maximum amount of embedded bits or information hidden in the cover file over a specific period of time. It is usually measured in bit per second [7]. The number of hidden data bits reflects the capacity in the sense that the higher the number of hidden data bits are, the higher its capacity is. It can be calculated and measured by using the following equation:

$$C_r = \frac{T_h}{T_d} \times 100$$

Where;

C_r = Percentage capacity ratio,
T_h = Total of hidden secret data (bits), and
T_d = Total of data in cover file (Kb).

A high capacity means that the system is fully utilized when embedding the watermark. Although this is good to begin with, it may affect the transparency of the watermark in the watermarked data. Therefore, a good capacity system should be high and does not affect the visibility of the watermark. Generally, different watermarking applications may require different watermarking capacities.

D. SECURITY

Security refers to the ability of the system to prevent any unauthorized manipulation of the watermark [57]. The digital text should be secure enough to prevent any unauthorized removal of the hidden data and should be only modifiable by the owner [58]. Any potential attacker must not be able to modify or fabricate the hidden data without knowing the key [59]. Blind algorithms are useful in this respect as they ensure the security of the approach where the original cover data are not needed when validating the copyright owner of text. The unauthorized manipulation of the content of digital media can be prevented by upgrading the system efficiency and restricting access to authorized persons. In this case, it is important to ensure that the key is well hidden from any potential attacker. This can be done by securing it with cryptographic security. The system security can be enhanced by using a cryptographic application and digital signature, for which the data will be hashed using hash function.

In order to evaluate the security of a system, it has to be assumed that the attacker knows the approach used for the watermark embedding. However, the key remains hidden under the cryptography protocol to ensure its security. In this case, the attacker will try to analyze the watermarked data and destroy the watermark. He or she should not be able to tamper with the watermark data without also changing the content of the data. In such a case, the watermark is considered as well secured. It is important to identify the best security protocol and policy in order to improve the text watermarking effectiveness. The application of the hash function and cryptographic algorithm must be taken into consideration when building the watermarking system.

V. THE EMBEDDING PROCESS

The embedding process in text watermarking can be defined as the process in which the watermarking information is generated and implemented. The two types of embedding are logical and physical embedding. They identifies how the watermark is stored, either physically in the text itself or logically. Although there are three types of watermarking approaches as discussed, the embedding process of these methods might differ, whether it uses logical embedding or physical embedding. Both methods possess their advantages and disadvantages as discussed in the sections below.

A. LOGICAL EMBEDDING

Logical embedding is a completely invisible or imperceptible text watermarking method. It is a case of data hiding where the embedded data is completely invisible to the viewer due to the fact that it is not done on the text. This unique method entails embedding the watermark digitally into the text, without physically applying the watermark inside the text [56], [60]. This is achieved by generating the watermark data from the text. While the watermarked data are kept secret and ready for verification, the original texts are distributed to the public. In order to verify the authenticity of the attacked data, the watermarked information is gathered from the data and compared with the original watermarked data.

As the method of logical embedding is not done on the text, the image, structural or feature characteristics of the text are analyzed and certain embedding characteristics are gathered to obtain the watermark information. The information is kept safe with certification authority and is used for future authentication purposes. This process proves to be highly robust and is imperceptible, since the data distributed to the public do not contain the secret information. However, it has some drawbacks that deserve to be highlighted. Although it is safe from format-based attack or content-based attack, a potential attacker may easily change or alter a major part of the text so that the watermark information cannot be detected anymore. If the watermark is hard to be detected, the text information can be misused. Most of the proposed logical embedding process techniques use the term zero-watermarking approach.

In this work, we analyze some of the studies on the logical embedding method. The analysis of the methods in relation to the attacks and performance are summarized in Table 1.

1) HE et al. (2009)

The authors proposed a novel text zero watermarking algorithm which use words that correspond to one special part-of-speech (POS) tag subsequence pattern. Part-of-speech is a category of words that have similar grammatical properties. The sequences were extracted using the chaotic function to develop the watermark without modifying the cover data,

TABLE 1. Attacks and performance evaluation on logical embedding text watermarking.

No.	Author	Evaluated attacks					Other evaluation metric	Performance analysis
		Insertion	Deletion	Reordering	Synonym substitution	Syntactic transformation		
1	[61]	x	x	x	/	/	x	Almost impossible to destroy watermark by synonym substitution and sentence transformation. **Synonym substitution attack**: <5% for more than 10 text. **Sentences transformation attack**: <10% for more than 100 transformed sentences.
2	[28]	/	/	x	/	x	x	Good performance of experimented robustness attacks. **Addition and deletion attack**: for >50% of words added or deleted, more than 90% of watermark can be extracted. **Synonym substitution attack**: accuracy rate is lower for 50% of transformed words, but higher for 100% transformed words.
3	[62]	/	/	x	x	x	x	The accuracy was evaluated with 5%, 10%, 20%, and 50% insertion attack (IA) and deletion attack (DA), it shows that watermark is more sensitive towards deletion attack, as watermark accuracy is least for 5%IA, 50% DA and 20% lA, 50%DA.
4	[63]	/	/	x	x	/	x	Pattern matching rate clearly indicate low, moderate and high state of tampering where tampering with text always gets detected.
5	[64]	/	/	/	x	x	x	Better result from previous methods
6	[25]	/	/	x	x	x	x	Watermark distortion rate is very high even when insertion and deletion volume is low
7	[56]	/	/	/	x	x	x	The ratio of successfully detected watermark shows up to 100% of detection rate.
8	[65]	/	/	x	x	x	x	The ratio of successfully detected watermark from insertion and deletion attacks shows up to 100% of detection rate.
9	[66]	x	/	x	/	/	x	Better performance and usability. Method is improved in terms of anti-aggressive, robust and erroneous recognition. **Sentences transformation attack**: >70% transformation gives 0% watermark detection. **Synonym substitution attack**: >20% synonym replacement gives lower robustness. **Deletion attack**: >20% synonym replacement gives lower robustness.
10	[67]	/	/	x	x	x	x	High percentage rate of watermark detection for both insertion and deletion attack
11	[68]	/	/	/	/	/	x	Method has better robustness and is proved to be better than some identified previous watermarking methods **Syntactic transformation attack**: <20% transformation gives 100% %watermark existence. **Synonym replacement attack**: >20% replacement gives <80% watermark detection **Addition, deletion and reordering attack**: accuracy of extracted watermark >70%
12	[69]	x	x	x	x	X	Computational time and percentage of watermark key.	Lower computational time but no improvement on the percentage of watermark key after tampering as compared to previous methods.
13	[70]	x	x	x	x	x	Development of the system	Detection of any tampering and minimal hardware resource requirements
14	[71]	x	x	x	x	x	Computational time to encode and decode watermark	Computation time to encode is higher than to decode.
15	[72]	/	/	x	/	/	x	Experiment proves this approach has better robustness and anti-attack. **Synonym substitution attack**: <10% gives accurate watermark detection. **Sentences transformation attack**: >24% transformation still enable watermark detection. **Addition and deletion attack**: <15% replacement gives higher accuracy
16	[73]	/	/	x	/	x	x	Robust against experimented attacks and sensitive to malicious tampering **Synonym replacement attack**: Does not show result. **Addition and deletion attack**: <20% gives high tamper detection.
17	[74]	/	/	/	x	x	x	Good percentage of extracted watermark with result of higher than 90%.
18	[74]	/	/	/	x	x	x	The method shows highest performance against reordering attack, followed by deletion and insertion attack.
19	[60]	x	x	x	x	x	Computational time to encode and decode watermark, features and benefits	Shows lower computational time. Decoding time is higher than encoding time.
20	[76]	/	/	x	/	/	x	**Syntactic transformation attack**: <10% transformation gives 100% watermark generation. **Synonym replacement attack**: <50% replacement gives >90% watermark similarity. **Addition and deletion attack**: for attack rate of <40%, similarity to watermark is >80%.
21	[77]	/	/	x	x	x	x	**Adding attack**: At 8%, intact rate began to decrease. **Deleting attack**: at 6%, intact rate began to decrease.

which resolves the problem of imperceptibility. This method offers good security since any attacker does not know the sequence of the selected POS tag while the chaotic function has unpredictable properties. Robustness is also assured and includes protection against reformatting and converting document type, synonym substitution, and sentence transformation attack with a very low success percentage [61].

2) MENG et al. (2010)

The team discussed a method where the watermark key is calculated based on the sentence entropy. Entropy is the average expected value of the information contained in each message. The entropy of the sentence is calculated based on word frequency and crucial selections, where the watermark is constructed based on the order of the crucial sentences. Given the complexity of Chinese text semantics, a sentence with a high word frequency possesses bigger entropy. This method is tested for some known attacks which includes 'adding and deleting' attacks and synonym substitution attacks. Tests have shown that this method is robust and withstands this kind of attacks which shows very low success rate. This method offers imperceptibility and security where the watermark key is registered with the authority [28].

3) JALIL et al. (2010)

They proposed a method that logically embeds the watermark in the text in order to generate a watermark key. It first analyzes the non-vowel ASCII character occurrence in each partition to find the non-vowel character that occurs most often. The key letters from the author and the maximum occurrence of the non-vowel are used to generate the watermark. The watermark is then registered with the certification authorities in order to provide security. In an insertion and deletion attack, the accuracy of the extracted watermark on attacked text is analyzed. The insertion and deletion rate of 5%, 10%, 20% and 50% are tested and the result of the watermark accuracy are evaluated. The accuracy is shown to be lowest when the percentage of insertion and deletion is highest. Since the essential parts of the text are used to embed the watermark, it is impossible to completely destroy the watermark without degrading the text content [62].

4) JALIL et al. (2010)

This team also proposed another zero watermarking algorithm which offers protection against tempering including syntactic transformation attacks in the form of passivizing, clefting, topicalizing or rephrasing the text content. The method identifies all the words with more than four letters. The words' initials are used to generate a watermark key for each sentence. In a simulated insertion and deletion attack, the pattern matching and watermark distortion rate are evaluated indicating a good result in which low, moderate and high levelled tampering attacks are detected at a consistent rate [63].

5) JALIL et al. (2010)

They proposed an image to text-based watermarking algorithm that first converts a watermark image to an alphabetical watermark. The key is generated from the first letter that occurs most often (MOFL) after using the propositions in the text as the separator. The separater are used to form groups and each group containing group size (GS) partitions. Each group's first double letter occurrences are used to create the MOFL list. Finally, the watermark key is generated from the watermark letters and the list. The watermark has been proven to withstand and resist insertion, deletion and re-ordering attacks. It is also computationally efficient. As compared to previous methods, this algorithm gets higher percentage of successfully detected watermark after the attacks [64].

6) JALIL et al. (2010)

Another method proposed analyzes the characteristics of the text to generate the watermark. A keyword occurring multiple times is selected, and a watermark is generated based on the length of the word preceding and following the keyword. A numeric watermark key is obtained and registered with the certifying authority for authentication. The experiment done to evaluate this approach is based on insertion and deletion attacks to calculate the watermark distortion rate. The test result has shown that this method is very sensitive to distortion. The accuracy of the tampered cover file is calculated and found to be feasible [25].

7) JALIL AND MIRZA (2010)

This proposed algorithm uses preposition and double letters to generate the watermark key has been evaluated. In this approach the partition of data in the cover file are analyzed by its frequency of repeating letters. The key is developed based on the count of those letters in an interval of time. This method uses image to text conversion to produce the hidden data to be integrated into the cover file. The experiment assesses the degree of closeness between the original and the altered watermark. This algorithm appears to be effective, secure and more robust in terms of resisting any insertion, deletion and reordering attack [56].

8) JALIL et al. (2011)

The same team proposed another algorithm which embeds the watermark image logically into the text by generating a key. This approach uses text constituents, double letters and most used English word to create the watermarking key. The algorithm is tested against insertion and deletion attacks, and other experiments are used to evaluate the effect of localised and dispersed alteration by way of insertion and deletion. The result shows that it is secure and robust against those attacks [65].

9) MENG et al. (2011)

Zero watermarking based on space models was also proposed by Meng et al. whereby the zero watermark was constructed from a 3-D model using the 2-D coordinates of word level and the sentence weights of the sentence-level. The 2-D word space structure consists of the length and frequency of the words and is extended into a 3-D model. The text watermark was constructed by mapping the sentence to the 3-D model. This algorithm is tested with three most common attacks; syntactic transformation attack, synonym replacement attack, and deleting attack, in order to test its performance. In the syntactic transformation attack, the higher percentage of transformation yields a lower watermark detection rate. The same is the case with the synonym replacement attack and the deletion attack. This method improves robustness and provide good imperceptibility and security [66].

10) AL-WESABI et al. (2012)

The research team proposed a watermarking approach based on Markov's model where the probabilistic features of the cover file are used to generate the watermark key. The watermark information of the text is analysed using the hidden Markov model where it is stored for authentication of the document. It offers protection against attacks with a watermark distortion percentage rate of greater than one for all insertion and deletion attacks [67].

11) MENG et al. (2012)

Meng and colleagues create a three-dimensional (3-D) space model by combining the general principle and methods of the traditional zero watermark with the syntactic and semantic features of the text. The algorithm generates a watermark based on the abstract set which can be extracted later by comparing the distance of each sentence point. Its effectiveness and feasibility was proven through algorithm simulation. The performance tests included attacks in the form of syntactic transformation, synonym replacement, addition, deletion and reordering attack. The syntactic transformation and synonym replacement attacks were resisted well while the deletion and reordering attack revealed the model's sensitivity to changes. It offered generally a good level of security by providing a certifying authority to register the watermark key [68].

12) TAYAN et al. (2013)

Two methods are proposed in this work. The first method uses data sequencing of the watermark logo and a unique key from the text to embed data in file content. The Unicode binary values of each word character are summed up in a data sequence to produce the registered watermark key. The second method utilizes the characteristics of text content in order to obtain the watermark key. By summing the Unicode values of each character, the watermark key is obtained. The proposed approaches are compared and tested in terms of their computational time and percentage of watermark key change after tampering. The first method is shown to have lower computational time due to its lower algorithm complexity as compared from the second method. For the percentage modification of watermark, it doesn't show a lot of improvement from previous techniques for both methods. Nevertheless, the ratio is considered small [69].

13) TAYAN et al. (2013)

Here the watermark key is produced in numerical values using the Unicode standard of characters. This algorithm inserts one watermark bit per character of the word set. A system is built to test the authentication of sensitive digital text of the Holy Qur'an. This method allows the detection of any alterations to the original text content and requires only minimal hardware resource requirements [70].

14) TAYAN et al. (2013)

This approach uses an image-to-text converter in order to generate the watermark key. The data sequence from the image logo is embedded in the duplicated cover file where it is processed and classified into word sets and the key is generated based on its characteristics. This method offers protection against attempts of forgery and unauthorized content manipulation. The watermark key is secured using a blind and fragile watermark extraction approach. The algorithm is tested in order to evaluate its computational time to encode and decode watermark, which produces good results [71].

15) YINGJIE et al. (2013)

A text zero-watermarking is constructed based on the Chinese edit distance using Chinese machine code and information strategy. The weighing terms in paragraphs are computed and the text feature selected formally expressed by the Chinese machine code. The edit distance is computed, and the commutation position is marked to complete the watermark construction. This method shows good robustness and resistance to addition and deletion attacks, synonym substitution attacks, and syntactic transformation attacks. It offers security by providing a date stamp and copyright holder ID for each key that is generated [72].

16) QI AND LIU (2013)

The proposed algorithm uses POS tag frequency in order to obtain information on the text feature. The zero watermarking algorithm is proposed based on the cloud model, the watermark being generated by a forward cloud model generator using the features extracted from the frequencies of the POS tags. This algorithm is very sensitive to insertion and deletion attacks. Less than 20% of insertion and deletion allows for a very high rate of tamper detection as the cloud model generated after the attack possesses different POS. Synonym substitution attacks do not have any effect on the watermark similarity level [73].

17) JALIL AND MIRZA (2013)

The proposed embedding algorithm uses the occurrence of preposition and double letters in the text to generate a key based on the most frequently used words in the English language. It offers robust copyright protection of plain text and secures it against insertion deletion and reordering attacks. This blind watermarking algorithm can also be used for ASCII characters and records a good percentage of proposed extracted watermarking with results above 90% [74].

18) BA-ALWI et al. (2014)

This method uses natural language processing to extract the probabilistic pattern based on third order 3-gram of the Markov model. It analyses the content of any English text document and extracts the probability features of the interrelationships between these contents. This approach shows a better performance of robustness in insertion, deletion and reordering attacks as compared to other approaches [75].

19) TAYAN et al. (2014)

This research upgrades the image-based approach by using a hybrid technique based on zero watermarking and digital signatures to further secure the data. It aims at protecting all types of text including sensitive text by modifying the content of the cover file and thus ensuring the content integrity and originality. The proposed approach uses logical embedding of the watermark data in the cover document, where the image of the watermark is converted into a character sequence. It uses Unicode to numerate words into binary values. When compared to the other proposed watermarking methods, it is found better in terms of tamper detection, robustness, capacity ratio, perceptibility, document authenticity verification, and language independence. It has its drawback as it requires large storage in the certification authority to store the keys [60].

20) LIU et al. (2015)

The proposed algorithm is based on merging features of Chinese text sentences. The text is segmented into sentences, where the semantic code of every word is used to calculate sentence entropy. The weight of each sentence is obtained using the sentence entropy, relevance, length, and weighing function. The key is then encrypted and registered with a trusted third party, the Certified Authority (CA). This method offers security as well as robustness in terms of attacks prevention [76].

21) ZHU et al. (2016)

The proposed algorithm is based on the connection between syllable parts of the Chinese phonetic alphabet. Every syllable is assigned an initial and final value, and the frequency is counted in accordance with the sum of the values. A sequence is formed in correspondence to the value of the sums and transformed using the logistic chaotic equation. The text watermarking key represents the result of the transformed sequence. One dimensional logistic chaotic equation is tested since these features of the chaos meet the demand of the sequence key to determine its robustness against content attacks (adding and deleting) and format attacks. The good result reflects the algorithm's capability in performing anti-adding attacks. It shows strong robustness and high resistance to tampering attacks. The key is secured with registered information date and license sent to registration center in order to prevent illegal access [77].

B. PHYSICAL EMBEDDING

Physical embedding can be defined as a process where the watermark is physically embedded into the cover text. The watermark is embedded in the form of a linguistic, image or structural manipulation. The watermark can be visible, slightly visible or invisible to public depending on the method used. In visible watermarking, the embedded watermark is visible to the end user [78] and thus discourages the copying or reusing of the data. It is also useful in advertising the owner of the work. Usually this type of embedding is done in the form of an image watermark, yet not common for text content as text content can be easily retyped and manipulated. However, visible data can also be more easily removed or altered by the public user [79]. Thus, text watermarking is not as robust against attacks and it promotes the unauthorized copying and redistribution of data. Another drawback of physical watermarking is that it may alter the content of the text.

The changes made to the text vary from being slightly visible to completely invisible text watermarking depending on the method used. In invisible or slightly visible physical embedding, the watermark is only known to authorities. Therefore in the case of copying and tampering, the watermark data are not destroyed and deleted as the end user does not realize that there is watermark embedded in the text. The invisible watermark may hinder the manipulation of the text, and any sort of tampering can be detected. Physical embedding can be done using the three approaches of text watermarking as described previously, although the level of visibility differs. Below discussed are the related works on the physical embedding of text watermarking together with an analysis according to the attacks and performances in Table 2.

1) ZHANG et al. (2010)

A text watermarking approach for word document is proposed by Zhang and colleagues. In this work, a novel method of robust watermarking for Word documents is proposed to protect copyright and dissemination control. A Word document consists of a lot of word objects arranged in a hierarchical order. Since every object has its own properties and functions, they cannot be modified via the interface of word application and only via programming. The author's and the legal user's information is embedded in the special attributes of word objects after encryption, grouping and packing into the message. The experimental result shows that this watermarking scheme performs excellently in terms of robustness and capacity. After all kinds of attacks (including adjusting the features, deletion, insertion and replacement), the watermark can still be extracted from the document. It can be widely used in copyright protection and protection of plain text and secures it against insertion deletion and reordering attacks. This blind watermarking text delivery on the internet and applies to both English and Chinese language. This method, however, shows low imperceptibility and is unable to withstand retyping and font changing attacks [80].

2) SHIRALI-SHAHREZA AND SHIRALI-SHAHREZA (2010)

In this text watermarking technique for Arabic and Persian script, the common characters in Arabic and Persian is used to hide the watermark's bits by exchanging the Arabic characters to Persian in the Arabic text and vice versa. This method shows good capacity but gives low imperceptibility and robustness. It is prone to many kinds of attacks including retying, reformatting, and replacement [81].

TABLE 2. Attacks and performance evaluation on physical embedding text watermarking.

No.	Author	Evaluated attacks					Other evaluation metric	Performance analysis
		Insertion	Deletion	Reordering	Synonym substitution	Syntactic transformation		
1	[80]	/	/	/	x	x	Capacity ratio	High robustness and capacity but average imperceptibility. Cannot withstand retyping and copy and paste attacks.
2	[81]	x	x	x	x	x	Capacity ratio	Good capacity with average of 33 bit/kb but low robustness and imperceptibility. Prone to retying, reformatting and replacement attacks.
3	[32]	x	x	x	x	x	Average capacity	Good capacity of 2.8. Lower imperceptibility and robustness. Prone to formatting attacks.
4	[82]	x	x	x	x	x	Robustness against visual attacks	Imperceptible and high efficiency, but vulnerable to statistical analysis attack. Robust against visual attacks.
5	[83]	x	x	x	x	x	Average capacity	Not completely robust, low capacity technique which are 0.3. Good imperceptibility. Prone to formatting attacks.
6	[84]	x	x	x	x	x	x	High imperceptibility, good capacity but lower robustness. The method is vulnerable due to the usage of only 2 bits.
7	[85]	x	x	x	x	x	Watermark detection	High imperceptibility but lower robustness and security.
8	[86]	x	x	x	x	x	x	Good security and imperceptibility but low robustness.
9	[87]	/	/	x	x	x	x	Imperceptible, high robustness and can authenticate the copyright. However, changing of font style, copy and paste and retyping might damage the watermarked data.
10	[30]	x	x	x	x	x	Average capacity	Better capacity ratio than previously proposed with 1.3-1.5. Lower imperceptibility and robustness.
11	[88]	x	x	x	x	x	Memory complexity	Good imperceptibility but high memory complexity
12	[89]	/	/	x	x	x	Imperceptibility and capacity ratio	High imperceptibility with ratio between 63.15 and 70.88 and similarity between 99.93% and 99.97%. The capacity is 2 bits/word. The method is vulnerable to retyping, reordering font changing attacks.
13	[90]	/	/	x	x	x	Capacity ratio	High imperceptibility and good capacity which are 32% for method 1 and 74% for method 2. Good robustness but vulnerable to retyping attack.

3) GUTUB et al. (2010)

This kashida-based approach is a unique type of technique where the extended Arabic characters are used to hide the watermark bits. It does not affect the text and can be placed before and after certain letters. This method randomises the location of each kashida based on the sequence of random value. A pseudorandom number generator is used to obtain the random number allocated to each position. The capacity ratio is counted and shown to be higher than previous methods. It does not change or alter writing content and claims to be featured with security, capacity and robustness [32].

4) POR et al. (2012)

This open space method in hiding secret data constitutes a novel text-based data hiding method called UniSpaCh. It is suitable for embedding information in Microsoft Word documents using Unicode space characters. Eight types of space characters are utilised and embedded into the text. This method is highly imperceptible and suitable to be implemented and shows good efficiency. However, this method remains vulnerable to attacks using statistical analysis on Unicode characters. In order to protect the hidden information from such attacks, a secret key is suggested and changes in periodic mapping of spaces as well as encryption of the hidden data. Furthermore, expert readers can notice the abnormality in some places where the number of spaces between words is too high [82].

5) ALHINAHI et al. (2013)

This proposed method an enhanced kashida technique where the kashida are inserted in front of specific characters. A kashida is placed for bit 1 and omitted for bit 0. Certain rules followed in the process of embedding the kashida in order to enhance its security. This method is evaluated on the capacity ratio and it is emphasised on the fact that higher imperceptibility gives lower capacity [83].

6) RUI et al. (2013)

The proposed algorithm can watermark text content containing mixed English and Chinese text. LanguageIDOther and Noproofing property is used to hide the watermark. This method shows a higher capacity than others approaches and ensures good security by using the hashing method. It provides high imperceptibility but lower robustness. However, the algorithm uses only 2 binary bits, which makes it vulnerable. Its robustness against tampering attacks needs to be evaluated to show its ability to resist tampering attacks [84].

7) JAISWAL AND PATIL (2013)

The proposed method aims to watermark HTML webpages. The watermark is changed into HEX form, which then is converted into HTML tag. The tag is inserted into the source code of the web page. This method shows high imperceptibility since none of the content of the web page is changed. However, it has lower robustness since the source code is easily accessible to remove the tags. The security and capacity has not yet been evaluated [85].

8) MIR (2014)

Mir developed a watermarking algorithm that uses the analysis of semantics and syntax to generate the structural-based watermark. The watermark is embedded in the white space throughout the text content. The method is suitable for web pages and offers some level of security to protect the watermark. The robustness of this method is fairly low due to the vulnerability of tampering attacks and formatting attacks [86].

9) ZHANG et al. (2014)

The proposed method uses a text watermarking algorithm based on Word Software (Microsoft Word or WPS Word) for document and copyright protection. The complete imperceptibility is achieved by the Font.Hidden attribute of Office Word. This algorithm in C++ language uses a binary image as the watermark in the embedding algorithm. The binary bits from the image are used as the watermark data to be hidden in the text using the Font.Hidden attribute of Office Word. The experimental result shows that this algorithm achieves complete imperceptibility, high robustness and can authenticate the copyright. However, changing of font style and copying the text may damage the watermarked data [87].

10) ALGINAHI et al. (2014)

This proposed method of frequency recurrence properties uses kashida inserted in front of specific characters. The kashida are placed for bit 1 and omitted for bit 0. This approach offers good security, capacity and robustness. However, too many kashida characters may give rise to suspicion and might lower the security and imperceptibility. Also, not all letters contain secret watermark bits, thus lowering its capacity. The kashida method is also not applicable to short texts. The robustness of the proposed method has not been tested against attacks. Although this approach is almost invisible, it is unprotected against copy and paste actions, retyping and OCR attacks [30].

11) RIZZO et al. (2016)

This proposed method allows the embedding of the password-based watermark in short texts while strictly preserving the content. It preserves the appearance and the content without converting the text into image. It shows invisibility, content preserving properties as well as it is blind watermarking. This method uses alternative Unicode symbols to ensure visual indistinguishability and length preservation (content preservation). The selected symbols are chosen and transformed into identical symbols analysed and encrypted through Unicode. The method uses 64 bits for 46 characters, which is quite high if applied to longer words and sentences [88].

12) AL-MAWERI et al. (2016)

This method uses Unicode extended characters in watermarking documents. The watermark bit size is set to 80 bits and converted to binary bits, which are embedded in the text using certain extendable characters. This algorithm shows high imperceptibility as well as robustness for conversion, copying, and addition and deletion attacks. The robustness evaluation proves that the proposed algorithm tolerates most of the possible attacks and is able to extract the watermark with high accuracy. The capacity evaluation shows that the algorithm has a good payload capacity of about 2 bits/word. Improving the proposed algorithm to recover re-ordering attack and investigating other Unicode properties constitutes one possible direction of future research. The watermark cannot withstand retyping and font changing attacks, and only does show some level of resistance to reordering attacks [89].

13) ALOTAIBI & ELRAFAEI (2017)

Another method consists of using pseudo-space in Arabic text. Pseudo-space makes connected letters seem isolated. The space is utilized to hide the watermark bits. The first method adds a watermark to the text by inserting the pseudo-space based on the dot character contained in the Arabic text. The second method adds the pseudo-space with normal space, which increases the capacity. This method is imperceptible and robust when it comes to copy and paste attempts, formatting and tempering attacks, yet perceptible to retyping attacks [90].

VI. APPLICATIONS ON LANGUAGES

Text watermarking techniques are built to protect text documents. However texts consist of many languages and characters. Due to this, usually certain methods are developed to cater a certain types of text language only, and it is not applicable to other types of text. The ability of the text watermarking methods to be applied to certain languages is called a flexible method, and it usually does not contain a linguistic and language-based watermarking process. It embeds using the formatting of the text itself. If a technique is not flexible, it usually implies that the method uses either the language's linguistic properties or the language structural properties.

A. ENGLISH

English text uses alphabetic letters taken from the original Latin script. The Latin alphabet contains 26 letters, two of them containing dots. Text watermarking of English text can also be applied to other text that uses Latin scripts including Spanish. Spanish or Latin-American is the second most used language on the internet, and minor modifications can be done on the text watermarking methods to be implemented (the Spanish alphabet contains one extra letter ''ñ''). Only certain semantic and syntactic characteristics need to be adjusted in order to be used for Latin script languages other than English.

More and more studies on watermarking of English text are conducted, and various techniques are proposed following all three major approaches. It can be applied to many other languages that are written in the Latin script. A lot of of logical embedding approaches for English text also have been developed which can be apply to either the image or the structural-based approach.

B. CHINESE

Chinese script constitutes the most widely used writing systems in the world. Other Asian languages such as Japanese, Korean, and Vietnamese adopted this writing system. These characters can be divided into the traditional and simplified style. There are 26 letters in the Chinese alphabet which can be combined according to its phonetic system. 23 consonants and 24 vowels are represented using the combinations of the 26 letters. They can be written horizontally or vertically, yet are more commonly written vertically. The direction of writing is similar to that of Arabic, namely from right to left. The Chinese language system is based on three basic elements which are grapheme, pronunciation and signification [26].

Chinese text watermarking mainly uses the features of the text. Since the Chinese characters contain many different symbols and complicated structures, it allows the creation of many embedding spaces for Chinese text watermarking [26]. One part of the Chinese character denotes the meaning or abstract classification, and the other part with another separate meaning, and when the two parts combine together a complete Chinese character are formed with a new meaning. For example, the character "瞎" is formed by the character "目" and "害", and through the meaning of "目" and "害", we can know the meaning of "瞎" [26].

A number of studies have already been concluded on the watermark methods in Chinese text. Due to the different types of symbols and characters of Chinese text, the methods used to develop the watermarking are different from the text watermarking methods used in other languages. The approaches proposed for Chinese text are usually based on the structure and characteristics of the Chinese letters. No image-based approach is currently proposed for this text. Extensive research has been conducted on logical embedding techniques in digital natural language documents of Chinese text.

C. ARABIC

In Arabic text watermarking, the method used is more creative and interesting due to the nature of the language. Arabic has 27 letters with eight main diacritical signs. Arabic differs from English in terms of the letters' shapes and properties. Writing is from right to left and does not differentiate between uppercase and lowercase. Each letter is written slightly differently depending on its position in the word. Arabic characters use Unicode standard, which is the international character encoding standard for displaying text in computer systems. This standard uses UTF (8/16/32) encoding, which allocates space for 65,000 characters. Hence it is easy to define all the characters in different formats including digits and symbols. There are some other languages that use almost the same alphabet such as Urdu, Persian, and Kurdish. These languages can use the same method as that used for the Arabic text watermarking.

Six established methods are available to watermark texts in Arabic based on Unicode, Kashida, letter dots, pseudo code, image and diacritics [91]. Only certain approaches are applicable for highly sensitive Arabic texts which are the Holy Qur'an. This religious scripture contains sensitive text structure and cannot be changes. This religious scripture sensitivity to modification is due to its diacritics or letters whereby even the slightest change will potentially alters the meaning.

D. APPLICABILITY TO OTHER WRITING STYLES

Although text watermarking is usually proposed for English, Chinese and Arabic text, it can also be applied to other languages that have common characteristics. Table 3 shows the applicability of English, Chinese and Arabic text to other languages.

TABLE 3. Language applicability of text watermarking.

Language	Applicability to other languages
English	Languages that uses Latin script like Spanish, German, Swedish, French, Italian, Malay language and others.
Arabic	Urdu, Persian, Kurdish, Pashto and other Semitic language
Chinese	Japanese, Korean, Taiwan, Hong Kong, Macau, and other agglutinative language

Table 4 and 5 summarize the recent works of text watermarking to the applicable languages. Table 4 shows the application of logical embedding and Table 5 shows the physical embedding methods.

VII. SUMMARY OF TEXT WATERMARKING: TAXONOMY

Determining whether a text watermarking system meets its usability and security goals can be challenging. This section discusses the classification and evaluation method that is used in this study by proposing a taxonomy of text watermarking. Texts are more vulnerable and sensitive to modifications [92], and therefore the classification and evaluation approach is different. General watermarking classification focuses on other media (image, video and audio). This classification and evaluation is not suitable for text because text watermarking methods do not have the same challenges, implementation methods and approaches as other media. Attacks done on text watermarking is different, and there are different types of text and languages available, which contribute to its vulnerability.

Previously, researchers have analyzed text watermarking depending on their chosen criteria, usually based on their respective requirements and approaches. The classifications differ for each researcher, therefore it is difficult to compare and analyze the method. This develop a gap in the area of text watermarking, where the precise and comprehensive classification are in need. The above reviewed methods and

TABLE 4. Applicability of logical embedding methods to the relevant languages.

No.	Author	Applicability to languages			
		English	Chinese	Arabic	Others
1	[61]	x	/	x	x
2	[28]	x	/	x	x
3	[62]	/	x	x	x
4	[63]	/	x	x	x
5	[64]	/	x	x	x
6	[25]	/	x	x	x
7	[56]	/	x	x	x
8	[65]	/	x	x	x
9	[66]	x	/	x	x
10	[67]	/	x	x	x
11	[68]	x	/	x	x
12	[69]	/	/	/	/
13	[70]	/	/	/	/
14	[71]	/	/	/	/
15	[72]	x	/	x	x
16	[73]	x	/	x	x
17	[74]	/	x	x	x
18	[74]	/	x	x	x
19	[60]	/	/	/	/
20	[76]	x	/	x	x
21	[77]	x	/	x	x

TABLE 5. Applicability of physical embedding methods to the relevant languages.

No.	Author	Applicability to languages			
		English	Chinese	Arabic	Others
1	[80]	/	/	x	x
2	[81]	x	x	/	x
3	[32]	x	x	/	x
4	[82]	/	/	/	/
5	[83]	x	x	/	x
6	[84]	/	/	x	x
7	[85]	/	/	/	/
8	[86]	/	x	x	x
9	[87]	/	/	x	x
10	[30]	x	x	/	x
11	[88]	/	x	x	x
12	[89]	/	x	x	x
13	[90]	x	x	/	x

processes needs to be defined analysed in order to help future research select the most appropriate technique. In this study, the taxonomy developed is based on the study's criteria that can help assess and evaluate the proposed text watermarking techniques (Fig. 8).

VIII. OPEN RESEARCH CHALLENGES AND FUTURE RESEARCH DIRECTION

Although the process of watermarking has been broadly studied, the research on text watermarking has remained in its early stage. Several important issues have remained unsolved. In addition, new challenges continue to emerge from applications, and text watermarking has yet to be implemented by many organizations. The subsequent sections discuss some the key research challenges and future research direction that require further investigations.

A. INFORMATION INTEGRITY

One of the key aspects of text watermarking is integrity. Information integrity relates to information reliability, relevance, usability, quality and value. It can be defined as the assurance of accuracy and consistency of data [93]. The proliferations of information access over the internet provide users the opportunity to access vast amounts of information, which in turn calls for integrity. The issue is how to ensure the information integrity on the internet and how the owner can ensure that the information content is protected through text watermarking.

B. INFORMATION AVAILABILITY

Availability refers to the ability of a user to obtain access to information in an easy and secure manner. When a system is not secure and easily available, information security is affected. In the context of text watermarking, the information must be easily accessible and securely watermarked in order to prevent any alterations of the text content. The authorization service of the information must be easily available and implemented. In addition, given the millions of internet users worldwide, the generated and shared information must be fully protected to prevent misuse or manipulation. An effective system must be introduced in which the user obtains access information after an independent self-check of the system to ensure that the data available are well secured.

C. ORIGINALITY PRESERVATION

The originality of data accessible to the public user must be preserved. The data obtainable online may originate from all sorts of databases, and it is hard to identify whether the quality and originality are preserved well in all cases. Through text watermarking, information can be protected from inside the text, yet the scale of the technique's implementation is still low given its high processing time and lack of imperceptibility. The challenge is how to make sure the right text watermarking technique is properly implemented and offers high imperceptibility, so that the originality and quality can be preserved.

D. INFORMATION CONFIDENTIALITY

Information confidentiality or secrecy is the act of making the property or information not available or disclosed to unauthorized individuals, entities, or organizations. There are certain measures that need to be taken in order to prevent the protected information from reaching unauthorized

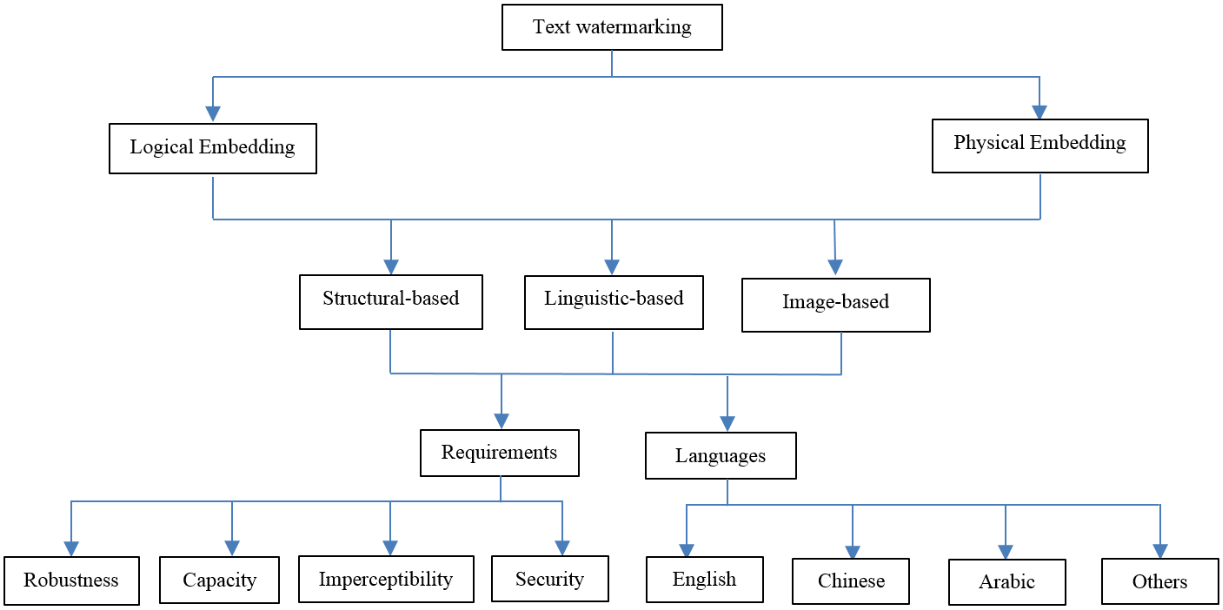

FIGURE 8. Taxonomy of text watermarking.

third parties (restricted access). In text watermarking, some techniques need to be implemented in order to take control of the content confidentiality. The challenge in this case is to find the suitable approach that promotes secrecy in text watermarking.

E. PROTECTION OF SENSITIVE INFORMATION

There are certain types of text content that is sensitive and require additional protection. Sensitive text is the type of text that cannot tolerate even the slightest form of modification such as a small alteration in a word or character [94]. It contains sensitive issues and format which may change the text meaning or alter the original purpose of the text. This is usually the case for religious scriptures, political documents, financial data, and government data. When it comes to text watermarking, such issues have only been addressed in relation to sensitive Arabic text in order to protect the religious scripture. Although quite a number of studies address the issue of text watermarking, yet sensitive text watermarking not in particular. Safeguarding the sensitive nature of certain information requires a special type of text watermarking, which is a challenge that needs to be worked on.

F. DOCUMENT TRANSFORMATION

The transformation of a digital file into other format entails the risk of losing the embedded watermark. For example; from word to pdf and vice versa. It is important to identify the text watermarking method that can withstand format transformation in order to ensure that the watermark remains secure and safe in any given file format.

G. CRYPTOGRAPHY APPLICATION

The security of embedded data needs to be further secured using the data security technique of cryptography. Cryptography helps safeguard the key and watermark from reaching the unauthorized user. A new framework is needed to test the implementation level of the security so that the text watermarking technique is trusted by those organizations that may depend on it.

H. LANGUAGE FLEXIBILITY

Text watermarking usually applies to a certain alphabet only. This reduces the applicability and usability of the technique. A good text watermarking method should be implementable to any type of text language. It represents a major challenge for the researchers to identify and propose the adequate techniques to solve this matter. Many criteria of the text languages need to be taken into consideration to successfully apply this method.

IX. CONCLUSION

This paper reviews the recent studies and issues on text watermarking. Text watermarking is becoming more and more popular, and many new methods and innovative techniques have been proposed and tested, which requires a new system of classification. A new model for evaluating text watermarking methods is developed that is readily and easily accessible and consulted by the research community and the relevant organizations. This model identifies the requirements, approaches, embedding processes and applications of text watermarking. Text watermarking faces many challenges, foremost in improving the implementation and precision of text detection. This article has discussed the most common challenges and problems in text watermarking. A brief overview of the existing research in this field is given, the possible limitations assessed, and their findings are evaluated. As for future research in the area of text watermarking, the primary tasks have been marked and await further research.

REFERENCES

[1] O. Thonnard, L. Bilge, A. Kashyap, and M. Lee, "Are you at risk? Profiling organizations and individuals subject to targeted attacks," in *Proc. 19th Int. Conf. Financial Cryptograph. Data Secur.*, 2015, pp. 13–31.

[2] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding—A survey," *Proc. IEEE*, vol. 87, no. 7, pp. 1062–1078, 1999.

[3] What is Intellectual Property? *World Intellectual Property Organization*. Accessed: May 11, 2017. [Online]. Available: http://www.wipo.int/edocs/pubdocs/en/intproperty/450/wipo_pub_450.pdf

[4] P. Singh and R. S. Chadha, "A survey of digital watermarking techniques, applications and attacks," *Int. J. Eng. Innov. Technol.*, vol. 2, no. 9, pp. 165–175, 2013.

[5] M. Kaur and V. K. Sharma, "Encryption based LSB steganography technique for digital images and text data," *Int. J. Comput. Sci. Netw. Secur.*, vol. 16, no. 9, p. 90, 2016.

[6] X. Zhou, W. Zhao, Z. Wang, and L. Pan, "Security theory and attack analysis for text watermarking," in *Proc. Int. Conf. E-Bus. Inf. Syst. Secur.*, 2009, pp. 1–6.

[7] 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.

[8] Miniwatts Marketing Group. (2016). *Top Ten Internet Languages—World Internet Statistics*. Accessed: Apr. 17, 2017. [Online]. Available: http://www.internetworldstats.com/stats7.htm

[9] O. Jane, E. Elbaşi, and H. G. İlk, "Hybrid non-blind watermarking based on DWT and SVD," *J. Appl. Res. Technol.*, vol. 12, no. 4, pp. 750–761, 2014.

[10] P. Chaubey and A. Singhadia, "An art to protect originality of digital data: Digital watermarking," *Int. J. Sci. Res. Eng. Trends*, vol. 1, no. 6, pp. 112–115, 2015.

[11] M. A. Qadir and I. Ahmad, "Digital text watermarking: Secure content delivery and data hiding in digital documents," in *Proc. 39th Annu. Int. Carnahan Conf. Secur. Technol.*, 2005, pp. 101–104.

[12] S. Mohammadi, "A semi-blind watermarking algorithm for color images using chaotic maps," in *Proc. 2nd Int. Conf. Knowl.-Based Eng. Innov. (KBEI)*, 2015, pp. 106–110.

[13] A. John, "Text watermarking using combined image and text for authentication and protection," *Int. J. Comput. Appl.*, vol. 20, no. 4, pp. 8–13, 2011.

[14] N. F. Maxemchuk and S. Low, "Marking text documents," in *Proc. Int. Conf. Image Process.*, vol. 3, 1997, p. 13.

[15] 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.

[16] M. J. Atallah *et al.*, *Natural Language Watermarking and Tamperproofing*. Berlin, Germany: Springer, 2003, pp. 196–212.

[17] Z. Jalil and A. M. Mirza, "A review of digital watermarking techniques for text documents," in *Proc. Int. Conf. Inf. Multimedia Technol.*, 2009, pp. 230–234.

[18] M. Kaur, "An existential review on text watermarking techniques," *Int. J. Comput. Appl.*, vol. 120, no. 18, pp. 29–32, 2015.

[19] R. A. Alotaibi and L. A. Elrefaei, "Arabic text watermarking?: A review," *Int. J. Artif. Intell. Appl.*, vol. 6, no. 4, pp. 1–16, 2015.

[20] 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.

[21] J. T. Brassil, S. Low, N. F. Maxemchuk, and L. O'Gorman, "Hiding information in document images," in *Proc. Conf. Inf. Sci. Syst. (CISS)*, vol. 95. 1994, pp. 482–489.

[22] D. Huang and H. Yan, "Interword distance changes represented by sine waves for watermarking text images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 12, pp. 1237–1245, Dec. 2001.

[23] Y.-W. Kim and I.-S. Oh, "Watermarking text document images using edge direction histograms," *Pattern Recognit. Lett.*, vol. 25, no. 11, pp. 1243–1251, Aug. 2004.

[24] A. M. Alattar and O. M. Alattar, "Watermarking electronic text documents containing justified paragraphs and irregular line spacing," *Proc. SPIE*, vol. 5306, pp. 685–695, Jan. 2004.

[25] Z. Jalil, A. M. Mirza, and M. Sabir, "Content based zero-watermarking algorithm for authentication of text documents," *Int. J. Comput. Sci. Inf. Secur.*, vol. 7, no. 2, pp. 212–217, 2010.

[26] Q.-C. Li and Z.-H. Dong, "Novel text watermarking algorithm based on Chinese characters structure," in *Proc. Int. Symp. Comput. Sci. Technol.*, 2008, pp. 348–351.

[27] W. Fei and X. Tang, "A Chinese text watermark algorithm based on pOLYPHONE," in *Proc. Cross Strait Quad-Regional Radio Sci. Wireless Technol. Conf.*, 2011, pp. 1215–1218.

[28] Y. Meng, T. Guo, Z. Guo, and L. Gao, "Chinese text zero-watermark based on sentence's entropy," in *Proc. Int. Conf. Multimedia Technol.*, 2010, pp. 1–4.

[29] R. A. Alotaibi and L. A. Elrefaei, "Utilizing word space with pointed and un-pointed letters for Arabic text watermarking," in *Proc. UKSim-AMSS 18th Int. Conf. Comput. Modeling Simulation (UKSim)*, 2016, pp. 111–116.

[30] 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. Electr. Eng.*, vol. 6, no. 5, pp. 381–392, 2014.

[31] M. H. Shirali-Shahreza and M. Shirali-Shahreza, "A new approach to persian/Arabic text steganography," in *Proc. 5th IEEE/ACIS Int. Conf. Comput. Inf. Sci.*, Jul. 2006, pp. 1–6.

[32] 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.

[33] J. Chen, F. Yang, H. Ma, and Q. Lu, "Text watermarking algorithm based on semantic role labeling," in *Proc. 3rd Int. Conf. Digit. Inf. Process., Data Mining, Wireless Commun. (DIPDMWC)*, 2016, pp. 117–120.

[34] M. Topkara, C. M. Taskiran, and E. J. Delp, III, "Natural language watermarking," *Proc. SPIE*, vol. 5681, pp. 441–452, Jan. 2005.

[35] M. L. Mali, N. N. Patil, and J. B. Patil, "Implementation of text watermarking technique using natural language watermarks," in *Proc. Int. Conf. Commun. Syst. Netw. Technol.*, 2013, pp. 482–486.

[36] M. J. Atallah, C. J. McDonough, V. Raskin, and S. Nirenburg, "Natural language processing for information assurance and security," in *Proc. Workshop New Secur. Paradigms (NSPW)*, 2000, pp. 51–65.

[37] M. J. Atallah *et al.*, "Natural language watermarking: Design, analysis, and a proof-of-concept implementation," in *Proc. Int. Workshop Inf. Hiding*, 2001, pp. 185–200.

[38] H. Wang, X. Sun, Y. Liu, and Y. Liu, "Natural language watermarking using chinese syntactic transformations," *Inf. Technol. J.*, vol. 7, no. 6, pp. 904–910, 2008.

[39] H. M. Meral, E. Sevinç, E. Ünkar, B. Sankur, A. S. Özsoy, and T. Güngör, "Syntactic tools for text watermarking," in *Proc. 9th Conf. Secur. Steganigraphy, Watermarking Multimedia Contents*, 2007, p. 65050X.

[40] M.-Y. Kim, "Text watermarking by syntactic analysis," in *Proc. 12th WSEAS Int. Conf. Comput.*, 2008, p. 904.

[41] H. M. Meral, B. Sankur, A. S. Özsoy, T. Güngör, and E. Sevinç, "Natural language watermarking via morphosyntactic alterations," *Comput. Speech Lang.*, vol. 23, no. 1, pp. 107–125, 2009.

[42] M.-Y. Kim and R. Goebel, "Adaptive-capacity and robust natural language watermarking for agglutinative languages," *Secur. Commun. Netw.*, vol. 5, no. 3, pp. 301–310, Mar. 2012.

[43] M. Lou and J. Liu, "Watermarking text document based on structure and semantic of chinese characters," in *Proc. 7th Int. Conf. Syst. Eng.*, 2012, pp. 866–869.

[44] N. Mir, "Zero watermarking for text on www using semantic approach," in *Proc. 2nd Int. Conf. Softw. Eng. Comput. Syst. (ICSECS)*, 2011, pp. 306–316

[45] O. Vybornova and B. Macq, "A method of text watermarking using presuppositions," in *Proc. 9th Conf. Secur., Steganigraphy, Watermarking Multimedia Contents*, 2007, pp. 1–10.

[46] H. Yazdani, V. Yazdani, and M. A. Doostari, "A new method to persian text watermarking using curvaceous letters," *J. Basic Appl. Sci. Res.*, vol. 3, no. 4, pp. 125–131, 2013.

[47] 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.

[48] F. Hartung and M. Kutter, "Multimedia watermarking techniques," *Proc. IEEE*, vol. 87, no. 7, pp. 1079–1107, Jul. 1999.

[49] N. Tiwari, "Digital watermarking applications, parameter measures and techniques," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 17, no. 3, 2017.

[50] I. J. Cox, M. L. Miller, J. A. Bloom, I. J. Cox, M. L. Miller, and J. A. Bloom, "Applications and properties," in *Digital Watermarking*. Morgan Kaufmann, 2002, pp. 11–40.

[51] L. R. Matheson, S. G. Mitchell, T. G. Shamoon, R. E. Tarjan, and F. Zane, "Robustness and security of digital watermarks," in *Proc. Int. Conf. Financial Cryptograp.*, 1998, pp. 227–240.

[52] X. Zhou, S. Wang, W. Zhao, and R. Peng, "A semi-fragile watermarking scheme for content authentication of Chinese text documents," in *Proc. 2nd IEEE Int. Conf. Comput. Sci. Inf. Technol.*, Aug. 2009, pp. 439–443.

[53] G. Sharma and D. Coumou, "Watermark synchronization: Perspectives and a new paradigm," in *Proc. 40th Annu. Conf. Inf. Sci. Syst.*, 2006, pp. 1182–1187.

[54] A. A. Ali and A.-H. S. Saad, "New text steganography technique by using mixed-case font," *Int. J. Comput. Appl.*, vol. 62, no. 3, pp. 6–9, 2013.

[55] M. Bashardoost, M. S. M. Rahim, T. Saba, and A. Rehman, "Replacement attack: A new zero text watermarking attack," *3D Res.*, vol. 8, no. 1, p. 8, Mar. 2017.

[56] Z. Jalil and A. M. Mirza, "An invisible text watermarking algorithm using image watermark," in *Innovations in Computing Sciences and Software Engineering*. Dordrecht, The Netherlands: Springer, 2010, pp. 147–152.

[57] S. Tyagi, H. V. Singh, R. Agarwal, and S. K. Gangwar, "Digital watermarking techniques for security applications," in *Proc. Int. Conf. Emerg. Trends Elect. Electron. Sustain. Energy Syst. (ICETEESES)*, 2016, pp. 379–382.

[58] B. Kaur and S. Sharma, "Digital watermarking and security techniques: A review," *Int. J. Comput. Sci. Technol.*, vol. 8, no. 2, pp. 44–47, 2017.

[59] F. Cayre, C. Fontaine, and T. Furon, "Watermarking security: Theory and practice," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3976–3987, Oct. 2005.

[60] 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, pp. 1–14, Aug. 2014.

[61] L. He, L. Zhang, G. Ma, D. Fang, and X. Gui, "A part-of-speeach tag sequence text zero-watermarking," in *Proc. ISCSCT*, 2009, pp. 187–190.

[62] Z. Jalil, H. Aziz, S. Bin Shahid, M. Arif, and A. M. Mirza, "A zero text watermarking algorithm based on non-vowel ASCII characters," in *Proc. Int. Conf. Edu. Inf. Technol.*, Sep. 2010, pp. V2-503–V2-507.

[63] Z. Jalil, A. M. Mirza, and H. Jabeen, "Word length based zero-watermarking algorithm for tamper detection in text documents," in *Proc. 2nd Int. Conf. Comput. Eng. Technol.*, 2010, pp. V6-378–V6-382.

[64] Z. Jalil, A. M. Mirza, and T. Iqbal, "A zero-watermarking algorithm for text documents based on structural components," in *Proc. Int. Conf. Inf. Emerg. Technol.*, 2010, pp. 1–5.

[65] Z. Jalil, M. A. Jaffar, and A. M. Mirza, "A novel text watermarking algorithm using image watermark," *Int. J. Innov. Comput.*, vol. 7, no. 3, pp. 1255–1271, 2011.

[66] 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.*, 2011, pp. 1–5.

[67] F. N. Al-Wesabi, A. Z. Alshakaf, and K. U. Vasantrao, "A zero text watermarking algorithm based on the probabilistic weights for content authentication of text documents," in *Proc. MPGI Nat. Multi Conf. Int. J. Comput. Appl.*, 2012, pp. 26–31.

[68] Y. Meng, L. Gao, M. Liu, and L. Bai, "Chinese text zero-watermark based on three-dimensional space model," *J. Comput.*, vol. 7, no. 8, pp. 2063–2070, 2012.

[69] O. Tayan, Y. M. Alginahi, and M. N. Kabir, "Performance assessment of zero-watermarking techniques for online arabic textual-content," *Life Sci. J.*, vol. 10, no. 4, pp. 93–100, 2013.

[70] Y. M. Alginahi, O. Tayan, and M. N. Kabir, "A zero-watermarking verification approach for Quranic verses in online text documents," in *Proc. Taibah Univ. Int. Conf. Adv. Inf. Technol. Holy Quran Sci.*, 2013, pp. 42–46.

[71] O. Tayan, Y. M. Alginahi, and M. N. Kabir, "An adaptive zero-watermarking approach for authentication and protection of sensitive text documents," in *Proc. Int.*, 2013, pp. 1–4.

[72] M. Yingjie, W. Xianlong, L. Wenjun, and C. Wei, "Text zero-watermark based on Chinese edit distance," in *Proc. Int. Conf. Comput. Inf. Sci.*, 2013, pp. 686–689.

[73] X. Qi and Y. Liu, "Cloud model based zero-watermarking algorithm for authentication of text document," in *Proc. 9th Int. Conf. Comput. Intell. Secur.*, 2013, pp. 712–715.

[74] Z. Jalil and A. M. Mirza, "A robust zero-watermarking algorithm for copyright protection of text documents," *J. Chin. Inst. Eng.*, vol. 36, no. 2, pp. 180–189, Mar. 2013.

[75] 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.

[76] 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.

[77] P. Zhu, G. Xiang, W. Song, A. Li, Y. Zhang, and R. Tao, "A text zero-watermarking algorithm based on Chinese phonetic alphabets," *Wuhan Univ. J. Nat. Sci.*, vol. 21, no. 4, pp. 277–282, Aug. 2016.

[78] S. Dhiman and O. Singh, "Analysis of visible and invisible image watermarking—A review," *Int. J. Comput. Appl.*, vol. 147, no. 3, pp. 36–38, 2016.

[79] A. Johnson and M. Biggar, "Invisible digital watermarks," U.S. Patent 7 269 734, Sep. 11, 2007.

[80] Y. Zhang, H. Qin, and T. Kong, "A novel robust text watermarking for word document," in *Proc. 3rd Int. Congr. Image Signal Process.*, 2010, pp. 38–42.

[81] M. H. Shirali-Shahreza and M. Shirali-Shahreza, "Arabic/persian text steganography utilizing similar letters with different codes," *Arabic J. Sci. Eng.*, vol. 35, no. 1, pp. 213–222, 2010.

[82] 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.

[83] 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)*, 2013, pp. 301–304.

[84] X. Rui, C. XiaoJun, and S. Jinqiao, "A multiple watermarking algorithm for texts mixed Chinese and English," *Procedia Comput. Sci.*, vol. 17, pp. 844–851, Jan. 2013.

[85] 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)*, 2013, pp. 69–72.

[86] N. Mir, "Copyright for web content using invisible text watermarking," *Comput. Hum. Behav.*, vol. 30, pp. 648–653, Jan. 2014.

[87] S. R. Zhang, X. C. Meng, X. F. Liu, and W. Y. Chen, "A digital text watermarking for word document," *Appl. Mech. Mater.*, vol. 614, pp. 347–351, Sep. 2014.

[88] S. G. Rizzo, F. Bertini, and D. Montesi, "Content-preserving text watermarking through unicode homoglyph substitution," in *Proc. Eur. Intell. Secur. Informat. Conf.*, 2016, pp. 97–104.

[89] 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 character," *Indian J. Sci. Technol.*, vol. 9, no. 48, pp. 1–14, 2016.

[90] 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.*, pp. 1–14, Jan. 2017.

[91] N. S. Kamaruddin, A. Kamsin, and S. Hakak, "Associated diacritical watermarking approach to protect sensitive arabic digital texts," in *Proc. AIP Conf.*, vol. 2014, 2017, p. 20074.

[92] N. N. Patil and J. B. Patil, "Performance analysis of a novel text watermarking technique for Devanagari text," in *Proc. Int. Conf. Intell. Commun., Control Devices*, 2017, pp. 325–333.

[93] J. E. Boritz, "IS practitioners' views on core concepts of information integrity," *Int. J. Accounting Inf. Syst.*, vol. 6, no. 4, pp. 260–279, Dec. 2005.

[94] S. Hakak, A. Kamsin, O. Tayan, M. Y. I. Idris, and G. A. Gilkar, "Approaches for preserving content integrity of sensitive online Arabic content: A survey and research challenges," *Inf. Process. Manag.*, pp. 1–14, Aug. 2017.

NURUL SHAMIMI KAMARUDDIN received the B.Sc. degree in mathematics from the Faculty of Science, University of Malaya (UM), Kuala Lumpur, Malaysia, and the M.Sc. degree in applied mathematics from the Faculty of Science and Technology, University of Kebangsaan Malaysia. She is currently pursuing the Ph.D. degree in computer system & technology with the Faculty of Computer Science and Information Technology, UM. From 2013 to 2015, she was a tutor and a lecturer in several institutions.

She is currently involved in text watermarking, as a Graduate Research Assistant. Her current research interests include network security, cryptography, artificial intelligence, watermarking, information hiding, authentication, applied mathematics, and graphical user authentication subjects.

AMIRRUDIN KAMSIN received the BIT degree in management from the University of Malaya, U.K., and the M.Sc. degree in computer animation from Bournemouth University, U.K., and the Ph.D. degree from University College London. He is currently a Senior Lecturer with the Faculty of Computer Science and Information Technology, University of Malaya, Malaysia.

His research interests include human computer interaction, authentication systems, e-learning, mobile applications, serious game, augmented reality, and mobile health services.

LIP YEE POR received the Ph.D. degree from the University of Malaya, Malaysia, in 2012. He is currently an Associate Professor with the Department of System and Computer Technology, Faculty of Computer Science and Information Technology, University of Malaya.

His research interests include neural network (such as supervised and unsupervised learning methods such as support vector machine, extreme learning machine), bioinformatics (such as biosensors, pain research), computer security (such as information security, steganography, authentication (graphical password)), grid computing, and e-learning framework.

HAMEEDUR RAHMAN received the bachelor's degree in software engineering with multimedia from the Limkokwing University of Creative Technology, Malaysia, and the master's degree in computer science and the Ph.D. degree from the Faculty of Information Science and Technology, University of Kebangsaan Malaysia, Bangi, Malaysia. He was involved in industry for five years in core java technology.

He is currently a Senior Research Member with the Center for Artificial Intelligence Technology and also a Manager with the MyXLab. He received some international awards including the Itex Bronze Medal of Innovation, Malaysia and the Winner of NASA Space App and Virtual Reality Arena.

He is currently pursuing the Ph.D. degree in breast cancer visualization using mobile augmented reality technology. His research interests include the area of augmented reality, virtual reality, imagine processing, cryptography, e-commence, e-learning, mobile technology, artificial intelligent, automation, and medical radiological technology.