Journal of Information Hiding and Multimedia Signal Processing ©2019 ISSN 2073-4212
Ubiquitous International Volume **10**, Number **3**, September **2019**

A Text Watermarking Algorithm based on Hidden Object

Dong Tian, Zhe-Ming Lu and Hang-Yu Fan

School of Aeronautics and Astronautics
Zhejiang University
Hangzhou, 310027, P. R. China
zheminglu@zju.edu.cn

Received July 2018; revised October 2018
(Communicated by Zhe-Ming Lu)

Abstract. *For digital copyright protection and traceability, many text watermarking algorithms have been proposed. However, the vast majority of them have one or more the following disadvantages: Insufficient embedded capacity, poor robustness, and large embedded noise affecting the expression of normal documents. To solve the above problems, this paper proposes a text watermarking scheme based on hidden objects. It solves problem from the level of programming objects in office and hides these objects that can record information, so as to realize the information hiding of the text. Compared with the existing methods, our proposed solution obtains higher robustness results against most of the extremely dangerous attacks, and at the same time possesses the faster embedded speed and the greater embedded capacity.*
Keywords: Text Watermark, Hidden Object, Embedded speed, Embedded Capacity, Robustness.

1. **Introduction.** In the digital information era, digitized and networked document data enables rapid information dissemination and convenient sharing. Therefore, how to effectively prevent digital documents from being abused and falsified, how to effectively protect the legitimate rights of copyright owners and how to trace the origin of documents urgently need to be resolved. The digital watermarking technology mainly solves the copyright and information security protection in the network environment. After the watermark is embedded, the host media can not be visually degraded or perceptible and after the inadvertent or malicious attacks, the watermark is difficult to remove, so as to the watermark can be detected. According to the use requirements of document information hiding, the document watermark should meet at least two benchmark characteristics. One is that the embedding noise is small enough so that the eyes of human are not aware that the document has been modified. The other is that the embedding of watermark information can not affect the normal use of the document. For example, it cannot change the original semantics of the document. In addition to these two points, robustness and capacity are also indicators for further evaluation of a watermarking scheme. In the past few decades, a series of document watermarking schemes that follow the above principles have been proposed [1, 2, 3]. They can be roughly divided into four types of method. The first one is based on text format. Its essence is to embed watermark information by changing line spacing, word spacing or other format information. Brassil et al. [5] proposed a variety of different methods for embedding hidden information in documents, including

inter-word embedding concealment information and line spacing embedding concealment information. Similar work [6, 7, 8] also appeared later. The second type of schemes [9, 10] is a document watermark based on character features and hidden information is recorded in features such as font size, font colour, and so on. There are two main issues with these methods mentioned above. One is that their embedded capacity strongly depends on the size of the text, The other is the poor robustness. For example, once delete a word, it will lose synchronization and can't extract any information. The third type of scheme is based on natural language, which is further divided into a natural language text watermark based on a grammatical structure [11, 12] and a semantic-based natural language text watermark [13]. The former can be achieved by adding a form subject, active changing to passive,and so on. The latter can be achieved by means of synonyms substitution, addition and deletion of punctuation. [11] proposed natural language watermarking method consisting of several steps. First, they construct a syntactic dependency tree of input text. Next, choose target syntactic constituents to move. Then, embed watermark bits. If the watermark bit does not coincide with the movement bit of the target constituent, move the syntactic constituent in the syntactic tree. Finally, from the modified syntactic tree, marked text formed. [13] proposed a method for selecting alternative vocabulary according to a quantitative resilience criterion when there are many words that can be replaced. These improvements have increased the concealment to some extent. But the biggest disadvantage of the natural language method is that the machine can not really understand the natural language and there will be a problem of dividing the words and sentences, so that the embedded information can also lost synchronicity, this makes it difficult to extract the watermark information. With the development of image watermark [14, 15, 16, 17], a series of text watermark based on image watermarking method are also proposed. Considering that the binary text images are featured complicated texture, little capability in date hiding and two values, [18] converts a text file into a binary image and embedding the watermark using a conventional DWT image watermarking method. Based on this, the embedded capacity of the text watermark rises to the same level as the image watermark embedding capacity. But due to the difficulty in accurately analyzing and using the features of the document image and the operation of the text is different from the image, this kind of watermarking method has the problems of poor robustness. Unlike image, audio and video, the text documents show very peculiar properties: binary nature, block/line/word patterning, clear separation between foreground and background areas [4]. So text watermark needs specific algorithms for its unique structure and embedded bottlenecks. The capacity issues have always been a challenge for text watermarking. Most existing watermark algorithms are more or less limited by the embedded capacity. On the other hand, the text watermarking schemes are generally sensitive to common file attacks such as deletion, format brush and so on. However, only by guaranteeing the ability to embed sufficient hidden information and resist common file attacks can the document watermark be put into practical use.

In order to solve the above problem, we introduce a novel idea of text watermarking standing on the perspective of programmable objects. A document is organized hierarchically by many objects. We find that some objects can carry information and be hidden at the same time. Their better attributes are that these hidden objects will not be deleted when the text is deleted and the format brush will not involve these objects. When copying a piece of text and pasting it into another text, the object can also be copied if it is within the selected range. Therefore, embedding objects redundantly and distributing them to all corners of the text will further increase robustness. These facts stated above imply that an object that can be hidden is an excellent tool for document watermarking.

TABLE 1. Mapping for inter-word and inter-sentence spacing

combination	size	combination	size
space	0000	Space+Three-Per-Em	0001
Three-Per-Em+Space	0010	Space+ Four-Per-Em	0011
Four-Per-Em+Space	0100	Space+ Six-Per-Em	0101
Six-Per-Em+Space	0110	Space+Figure	0111
Figure+Space	1000	Space+Thin	1001
Thin+Space	1010	Space+Hair	1011
Hair+Space	1100	Space+Punctuation	1101
Punctuation+Space	1110	Space+Narrow No-Break	1111

The watermarking scheme we proposed can be applied to various documents such as Excel, Word, PPT in which the embedded watermark is no longer limited by the embedded capacity, and has a high NC for extracted watermark and strong concealment simultaneously. More importantly, it has been verified through experiments that our proposed method can resist a variety of common text attacks.

The rest of the paper is organized as follows. We first introduce the related work with our method in Section 2 and then discuss the proposed document watermark scheme in Section 3. In Section 4, we describe the experiment to demonstrate the effectiveness of the scheme in the aspect of payload capacity and concealment. Our conclusions are summarized in Section 5.

2. Related Work.

A large body of literature and study exist about text watermark. Due to the space limitations, we are unable to discuss all of it in detail. So the following focuses on some of the methods associated with our algorithm in comparison experiments.

2.1. Watermark Scheme using Unicode Space Characters.

The scheme [21] is based on text format. It embeds watermark information into the mixture of inter-sentence, inter-word, end-of-line and inter-paragraph spacing. They selected 8 characters from all the spaces as the smallest units that can make up the new spaces. These smallest units are Three-Per-Em Space, Four-Per-Em Space, Six-Per-Em Space, Figure Space, Punctuation Space, Thin Space, Hair Space, and Narrow No-Break Space respectively. On this basis, all spacing can be replaced with the combination of these units to embed information. The specific regulations are an inter-word or inter-sentence spacing embeds four bits and a inter-paragraph or end of line spacing embeds three bits. The mapping for inter-word and inter-sentence spacing is shown in table 1.

The process of embedding the watermark in this scheme can be divided into two steps. First, traverse the text embedding watermark to retrieve all spacing including inter-word spacing, inter-sentence spacing, inter-paragraph spacing and the end of line spacing. The second step is replacing the retrieved spacing basing on the watermark information and the mapping relationship in table 1 until all watermark information is embedded in the text. During this process, if white space character is inter-word or inter-sentence, get four bits from secret data bit stream and replace the original spacing with the spacing the four bits correspond. Otherwise get three bits from secret data and embed corresponding space character.

2.2. Quantifiably Resilient Watermarking of Natural Language Text through Synonym Substitutions.

The scheme is a semantic-based natural language text watermark. They propose a lexical watermarking system that is based on substituting certain words with more ambiguous words from their synonym set. First, they build a graph G of

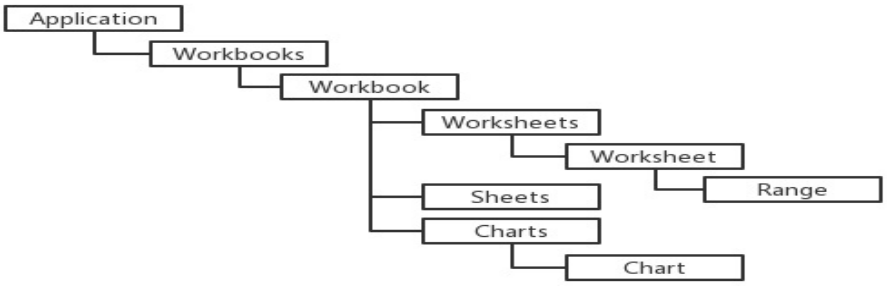

FIGURE 1. A simple object hierarchy in Excel

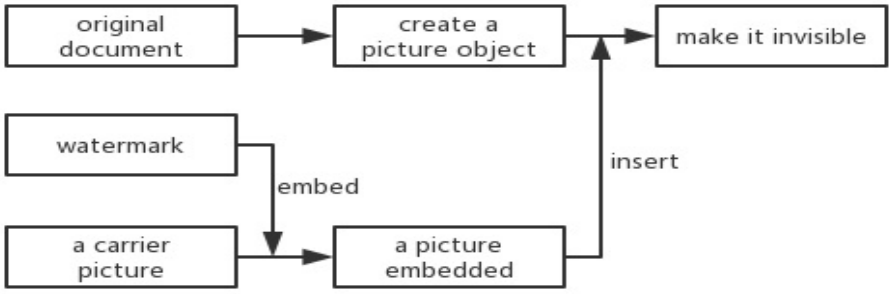

FIGURE 2. Embedding process

(word, sense) pairs. Connect different senses of the same word with a special edge in order to follow the links to every neighbor of a word independent from its senses. Secondly, calculate differences between the (word, sense) pairs, using a similarity measure. Assign these values as edge weights in G. Thirdly, select a subgraph G^W of G using the secret key k. Then, Color the graph G^W in such a way that approximately half of the homograph neighbors of a non-homograph word are colored with blue to represent the encoding of ?1?, and the other half are colored with green to represent the encoding of ?0?. At last, for each word w_i in the cover document S, replace w_i with the neighbor that carries the color that encodes bits of watermark if $w_i \in G^W$.

3. **Proposed Document Watermark Scheme.**

3.1. **The choice of objects.** Our proposed text watermarking scheme is from the perspective of programming object level. Taking an Excel document as an example, it is composed of many object levels. A simple composition example is shown in Fig.1. The Application object represents the entire Excel application and the workbooks object represents the collection of Excel workbook files. Below the workbook, there are worksheet objects, chart objects, range objects that set cell properties, front objects that control font properties, and so on. The pivotal task of this algorithm is to select a suitable object to embed information that can be hidden. The selected object should satisfy two basic characteristics. One is that the object can store hidden information, and the other is that the object can have invisible related attributes. Therefore, objects like Shape and Picture meet the above two requirements, and these two objects exist in Excel, Word, PPT and other types of texts. This makes it universal to all documents.

3.2. **The Embedding Scheme.** Take the object of picture as an example, the specific watermark embedding system is shown in Fig.2.

First, the scrambling generation algorithm is applied to the watermark to be embedded. Therefore, a key K and a scrambled binary watermark sequence are obtained. At the

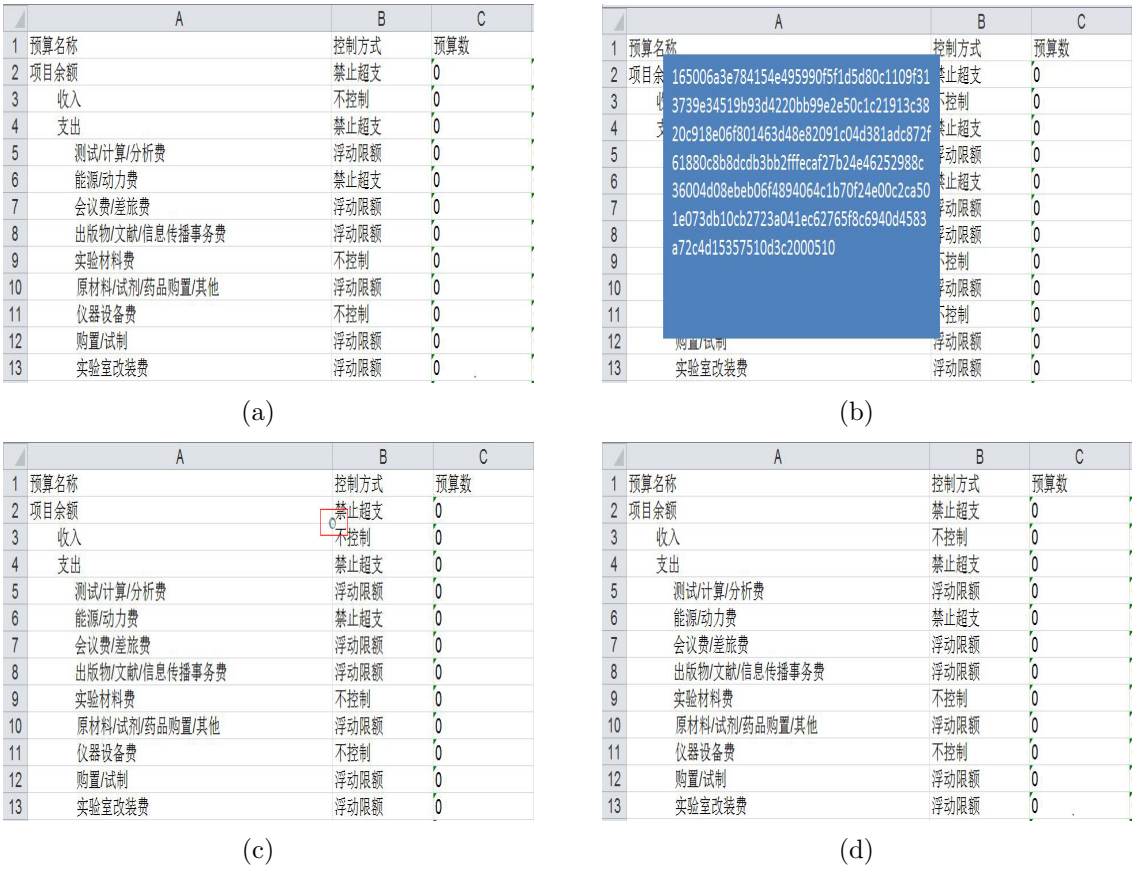

FIGURE 3. Excel interface during embedding process

same time, a local picture is used as a carrier picture, and the watermark is embedded into the carrier picture using a well-developed image watermarking algorithm (such as DCT transform domain embedding [19]). After that, a picture object is programmed to generate and insert the carrier image with embedded information into this object. Finally set this object size to zero and the invisible property to true. This completes the basic watermark embedding process. If choose the shapes object to embed information, it will go through roughly the same process. The specific difference is that no longer insert pictures into the document. Instead, create the shapes object. Add scrambled binary watermark information to the object. For security, it can also encode binary information before adding it to objects. Similarly set the object size to zero and the invisible property to true at the end. For example, we chose the shapes object of Excel. The Excel interface during embedding process is shown in Fig.3. Subgraph (a) is the original Excel file. Subgraph (b) shows the result of encoding a scrambled binary watermark into hexadecimal and inserting it into the created shapes object. Subgraph (c) shows the result of setting the size of the shapes object to zero. We can see that there is still a small visible dot in the red box of (c), which is the shapes object that previously embedded information. Finally, the embedded process is completed by setting the property of shapes object to invisible. Subgraph (d) shows the final result.

3.3. **The Extraction Scheme.** Take the object of picture as an example. The specific watermark embedding system is shown in Fig.4.

The extraction process for this document watermarking scheme is very simple. First, program to extract the hidden picture object in the document. Then use the corresponding

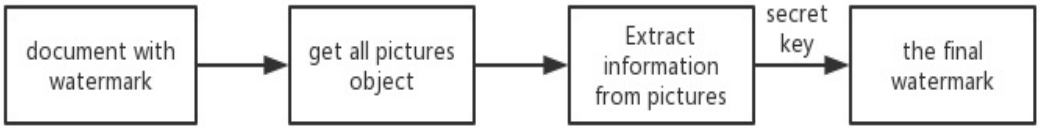

FIGURE 4. Extraction process

FIGURE 5. Watermark

image watermark extraction algorithm to extract the watermark information in the image. The key is inversely scrambled to obtain the final watermark. If the embedded selection is a shapes object, the extraction process is roughly the same. First, program to extract the hidden shapes object in the document. The anti-encoding and anti-scrambling of the information in the Shapes object is directly performed, and finally the final watermark information can be obtained.

4. **Experiment.** The experiments use Fig.5 as the watermark information. Its resolution is 32*32. In order to explain the solution presented in this paper has better performance than most text watermark. In the experiment, this scheme was compared with the Unicode space characters embedding method [21], the synonym substitutions embedding method [22], and the font size embedding method.

4.1. **Robustness.** In order to eliminate the influence of subjective and objective factors such as experimental observer's experience, physical conditions, and experimental conditions, the normalized correlation coefficient NC was used to quantitatively evaluate the similarity between the extracted watermark and the original watermark. Its expression is shown in formula (1).

$$N_{NC} = \frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} w(i,j)w^*(i,j)}{\sqrt{(\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} w^2(i,j))(\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} w^{*2}(i,j))}} \tag{1}$$

where w is the original watermark, w^* is the extracted watermark, M and N are the number of rows and columns of the watermark, respectively. Larger N_{NC} indicates better extraction effect. The following five attacks were performed on the documents embedded with the methods mentioned earlier. The experimental results about robustness are shown in table 2. The value in the table is the NC value calculated from the original watermark and the extracted watermark after the attack.

- Tamper with document content
- Delete content
- Save as other file(Eg: .doc files save as .docx files, .exl files save as .exls files)
- Format brush attack

TABLE 2. Anti-attack comparison of different algorithms

algorithm	tamper	delete	copy	format brush	save as
our method	1.00	1.00	1.00	1.00	1.00
Unicode space characters	0.49	0.58	0.67	0.50	1.00
synonym substitutions	0.91	0.66	1.00	1.00	1.00
font size	1.00	0.62	0.63	0.45	1.00

TABLE 3. Time for different algorithms to embed information

algorithm	time
our method	2.1s
Unicode space characters	8.75s
synonym substitutions	21.5
font size	10.2s

- Copy: copy the document and poste it to a new document

From table 2, it can be concluded that the proposed watermarking method is highly robust. It can well resist common document attacks. There are two main reasons why traditional methods cannot resist these attacks more or less. One is that for format-based embedding, once the format has been changed, the embedded watermark information in the format will be lost. Both Format brush and Copy and paste cause the format to be changed completely. The other is that when the characteristics of the carrier embedded in the watermarking bit are added or deleted, the synchronization [20] is lost. This is the most serious distortion so that most of the watermark information cannot be extracted properly. The method of this article does not have the above problems. Its watermark information is not embedded in the format and it embeds all the watermark information as a whole, so there is no problem of destroying synchronization.

4.2. **Capacity and Invisibility.** Because the carrier of watermark information is the object without capacity constraints, the method proposed in this paper is not limited in terms of capacity. However, the embedding capacity of the other schemes is closely related to the size of the articles. More concretely, the capacity of the scheme based on synonym substitutions is related to the number of times words in the custom index table appears in the document. The capacity of the scheme based on Unicode space characters is related to number of sentences and paragraphs and the capacity of the scheme based on font size is related to the number of words and so on. Embedding capacity is not limited by the text size is a highlight of this scheme. Since the embedded information object of this method can be hidden through programming, good invisibility can be guaranteed. For the synonym substitutions methods, there will be a chance that the machine does not correctly understand the semantics and segment words error, which leads to semantic irrelevance after the synonyms are replaced. Since the watermarking schemes that embed information based on the format need to change the format of the text, their invisibility will also be affected more or less. In contrast, the watermarking scheme proposed in this paper is better than other watermarking schemes.

4.3. **Time efficiency.** We also compared the time efficiency between different document watermarking algorithms in experiments. We embed the watermark once in the document which is shown in Figure.5, a total of 1024bit information. The speed of embedding information between different algorithms is shown in table 3.

From the above table, it can be concluded that the watermarking scheme proposed by us is faster than the general document watermarking scheme in the embedding speed. This is because it only needs to open an object interface for insertion during programming. Its time complexity is O(1). The scheme of text watermark based on the font size needs to call the font object n times, so its time complexity is O(n), where n is the number of words that need to change their size to embed information. Synonym substitution methods take a lot of time to segment words, find and replace synonyms.

5. **Conclusion.** We proposed an efficient text watermark scheme based on hidden object. It converts the task of embedding watermark to find objects that can store the information and have invisible attributes. Experiments show that the robustness and embedded speed of our proposed method are obviously better than other existing document watermarking schemes. At the same time, there is no capacity limit for the embedding method. This series of advantages ensures that the document watermark is applied in the actual scene.

REFERENCES

[1] Bhaya W. Text steganography based on front type in msword documents, *Journal of Computer Science*, pp.898-904, 2013.

[2] S. G. Rizzo, F. Bertini, and .Montesi. Content-preserving Text Watermarking through Unicode Homoglyph Substitution, *in Proceedings of the 20th International Database Engineering amp, Applications Symposium*, pp.97C104, 2016.

[3] S. G. Rizzo, F. Bertini, D. Montesi, and C. Stomeo. Text Watermarking in Social Media, *in Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pp.208-211, 2017.

[4] Kim Y-W, Moon K-A, Oh I-S. A text watermarking algorithm based on word classification and inter-word space statistics, *in Proc. of the Seventh International Conference on Document Analysis and Recognition*, pp.775-779, 2013.

[5] S.Low , N.Maxemchuk , J. Brassil and L.OGorman. Document Marking and Identification Techinques to Discourage Document Copying, *Proceedings of Infocom94*, pp.1278 1287, 1994.

[6] R. A. Alotaibi and L. A. Elrefaei. Improved capacity Arabic text watermarking methods based on open word space, *Journal of King Saud University - Computer and Information Sciences*, pp.1-13, 2016.

[7] N.Chotikakamthorn. Document Image Data Hiding Technique using Character Spacing Width Sequence Coding, *International Conference on Image Processing*, pp.250 254, 1999.

[8] Wang Haichun, Qiu Jifan, Qiu Dunguo. Design and implementation of a steganography system based on word document, *Computer*, pp.47-49, 2006.

[9] XX Xing. A text digital watermarking algorithm based on the value of character attribute, *Information Technology*, pp.199-204, 2016.

[10] Chen Qing, Huang Pengbo. A Text Watermarking Algorithm Based on Character Color of PDF, *Electronic Science and Technology*, pp.96-100, 2016.

[11] M. Y. Kim. Text watermarking by syntactic analysis, *in Proceedings of the 12th WSEAS International Conference on Computers*, pp.904C909, 2008.

[12] Petitcolas F A P , Anderson R J , Kuhn M G. Information hiding-a survey, *Proceedings of IEEE*, pp.1062-1078, 1999.

[13] U. Topkara, M. Topkara, and M. J. Atallah. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural language text through synonym substitutions, *in Proceedings of the Multimedia and Security Workshop*, pp.164C174, 2006.

[14] Feng-Hsing Wang, Kang K. Yen, Lakhmi C. Jain, and Jeng-Shyang Pan. Multiuser-Based Shadow Watermark Extraction System, *Information Sciences*, Vol.177, pp.2522-2532, 2007.

[15] Lin C C, Shiu P F. High Capacity Data Hiding Scheme for DCT-based Images, *Journal of Information Hiding & Multimedia Signal Processing*, pp.220-240, 2010.

[16] Hsiang-Cheh Huang, Shu-Chuan Chu, Jeng-Shyang Pan, Chun-Yen Huang and Bin-Yih Liao. Tabu Search Based Multi-watermarks Embedding Algorithm with Multiple Description Coding, *Information Sciences*, Vol.32, pp.3379-3396, 2011.

[17] Benhocine A, Laouamer L, Nana L, et al. New Images Watermarking Scheme Based on Singular Value Decomposition, *Journal of Information Hiding & Multimedia Signal Processing*, pp.9-18, 2015.
[18] Li Jingbing, Huang Xiyue, Zhou Yaxun. A Robust Watermark Algorithm of Binary Text Images, *Computer Engineering*, vol.32, No.22, pp.23-25, 2006.
[19] B. Chen and G. w. Wornell. Digital Watermarking and Information Embedding Using Dither Modulation, *IEEE Second Workshop on Multimedia Signal Processing*, pp.273-278, 1998.
[20] Lee S J,Jung S H. A survey of watermarking technigues applied to multimedia, *In: IEEE Conf of the ISIE*, pp.272-277, 2001.
[21] Rajeev Kumar, Satish Chand, and Samayveer Singh. An efficient text steganography scheme using Unicode space characters, *The International Journal of FORENSIC COMPUTER SCIENCE*, pp.8-14, 2015.
[22] Umut Topkara, Mercan Topkara, and Mikhail J. Atal-lah. The hiding virtues of ambiguity: quantiably resilient watermarking of natural language text through synonym substitutions, *In Proceedings of the 8th Workshop on Multimedia and Security*, pp.164C174, 2006.