J. Basic. Appl. Sci. Res., 3(4)125-131, 2013
© 2013, TextRoad Publication

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

A New Method to Persian Text Watermarking Using Curvaceous Letters

Vahid Yazdani[1], Mohammad Ali Doostari[1] and Hamid Yazdani*[2]

[1]Department of Computer, Information Technology Branch, Shahed University, Tehran, Iran
[2]Department of Electronics, Islamic Azad University, Nour Branch, Nour, Iran

ABSTRACT

Following the expansion of communication, hiding of confidential information has become a crucial requirement. Watermarking is one of the techniques used to hide information in image, sound and especially text. Comparing to English and Chinese texts, there has been little work on Persian texts, because of special characteristics of Persian texts. In this paper, a method is proposed, for the first time, to hide information in the curvatures existing in some Persian letters. One of the main steps in watermarking process is detection and separation of letters using the appearance features of the letters, and the way they stand within the text. Since that in the proposed method, information are embedded in Pixels of curved letters, the results of proposed method show the good quality, the capacity and the robustness of this method.

KEYWORDS: Security, Information Hiding, Curvaceous Letters, Persian Watermarking.

1. INTRODUCTION

Nowadays, due to increasing development of global communications and innovation of various communication channels such as internet network, satellite communications and other telecommunications, information is easily available to a wide range of people all over the world. One of the important issues of information security is the information hiding in data exchange. Therefore, information security has found special importance. In this regard, various techniques of hiding information are used such as cryptography, steganography, watermarking and etc. Watermarking is the art of keeping data against illegal copying and forging. In watermarking, in order to make security against abusing documents and making illegal copy out of them, limited and definite pieces of information are embedded into host media [1].

Based on the type of document to be watermarked, watermarking can be classified as image watermarking, video watermarking, audio watermarking, and text watermarking [1]. Unlike for text watermarking, there are varieties of methods in watermarking for the fields such as audio and video (either colored or grayscale). In general, text watermarking may be done through two ways. In the first method which is called low level watermarking and is not highly dependent on the kind of font (Calligraphy), watermark signal information are embedded in field of binary pixels inside binary text image. This method can also be used for non-textual images, because the method is somehow independent on the type of the involving media [2].

The second method which is called high level watermarking and is quite dependent on the type of text and font (Calligraphy) type, includes ways in which to place information obtained from the existing factors inside text image such as paragraphs, lines, words and characters. It also has a rather low capacity but a good robustness in comparison with the first method [2].

In this paper, a new method is proposed for Arabic and Persian text watermarking which uses the appearance features of curvaceous letters for confidential information embedding. The other parts of this paper are organized as follow:

In the second section, the differences between watermarking, cryptography and other methods have been discussed. The third section deals with different kind of watermarking applied to Persian and Arabic texts and. In section four, the features of curvaceous letters have been discussed and a suggested watermarking algorithm has been presented; and finally the results of suggested watermarking algorithm have been presented in the fifth section.

2. Comparison of Watermarking, Cryptography and Steganography

Cryptography refers to encoding the content of a message. Cryptography is derived from the Greek word "Crypto" meaning "secret". In order to make the operation of decrypting possible, the message sender should certainly send the decryption key for the recipient. The key should not be accessible for any party except for the intended recipient. The process of hiding the existence of a secret message is called steganography which is taken the Greek word "Stegano" meaning "hidden". In steganography, the existence of secret message is essentially denied. In steganography, the embedded secret message to the hosting environment should include the least amount of harm to the content of the host[3].

Inherent advantage of stenography is that, as embedded secret message is hidden from public view, similarly, the efforts and motivation to break the encrypted message will subside. It is worthy noticing that one should never think

***Corresponding Author:** Hamid Yazdani, Department of Electronics, Islamic Azad University, Nour Branch, Nour, Iran.
Email: eng.hamid.yazdani@gmail.com

stenography as a substitution for cryptography, but rather, it is an appropriate complement to cryptography. To increase security and reliability coefficient, steganography and cryptography can be used combinationally and synchronously.

Watermarking is a branch of steganography which can be considered as a suitable method for protection of copyright and guarding the data and documents against illegal copy. The difference between watermarking and steganography is that in steganography, appropriateness of subject and content between secret and media host is not required, and content between code and host are quite independent. But in watermarking, subject and content between secret and media host should be interdependent. For example, in a file content that is the passport information of individuals can be used individual characteristics such as fingerprint, eye color, the structure of iris as the watermark embedded to the main file. As another example, in a file associated with the production of a science center, copyright information of the center, customers list, logo plan of the center, history and … can be used as watermark. In fact, Watermarking can be seen as one of steganography methods that more focus on robustness and less on security [4] [6].

2. KINDS OF WATERMARKING METHODS

Text watermarking is the most difficult kind of watermarking; this is due largely to the relative lack of redundant information in a text file as compared with a picture or a sound file [5]. Many related watermarking techniques have been proposed for text watermarking. However, most of them are suitable for English and cannot be generalized to different other languages such as Arabic and Persian. Generally there are a few methods to information hiding in text which are as follow:

1 – Line Shift Coding
2 – Word Shift Coding
3 – Character Method
 a) Feature Coding
 b) Character Shift
4 –Open Space Coding

In Line Shift method, by replacing the lines of a text downwards or upwards, watermark information bits are stored in the text, which are suitable in text image [6]. The same process is done in Word Shift Coding. In this method, the information by shifting words horizontally and changing the distance between words, is hidden in text. The Word Shift Coding is suitable for the texts which the spaces between words are changeable in them. Implementation of this method is very time consuming and there is a great chance to find hidden information [7] [8]. The algorithm for this method is shown in Fig. 1.

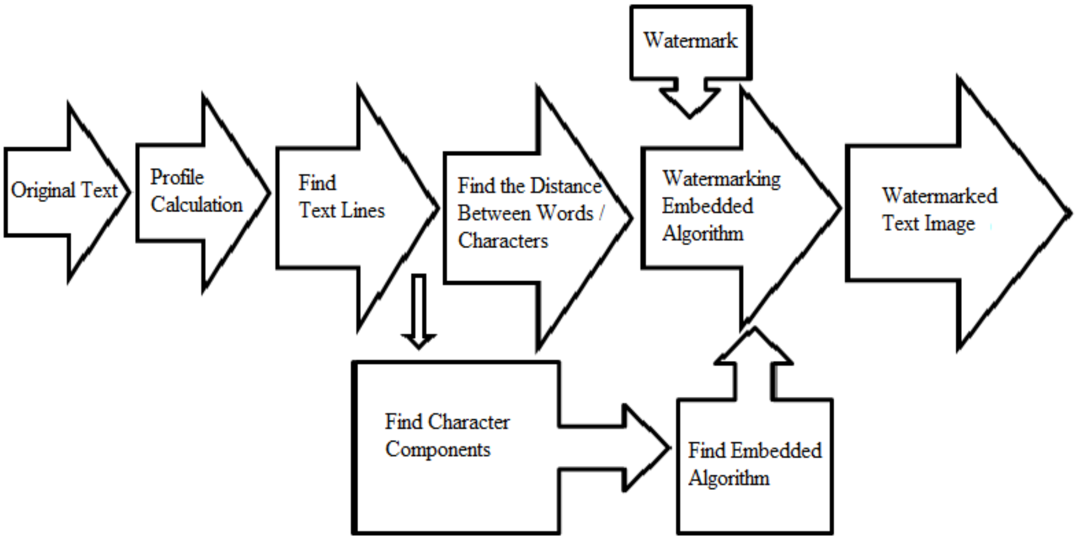


Fig. 1. Block diagram for data embedding process using the space between words and characters [4]

In watermarking method by using the characters in the text, Watermarking can be done in two ways. In the first case, like Line Shift Coding and Word Shift Coding, shifting of character can be used for information embedding [4]. The latter case is Feature Coding method, which is used in this paper. Feature Coding is completely dependent on the font (Calligraphy) of text. One example of the Feature Coding method is points relocation method. In this method to insert a bit with value of one, the point is slightly shifted up towards in pointed letters Otherwise, if the value of hidden bit is Zero, points location remains unchanged. This point shifting process is shown in Fig. 2 for the Arabic letter “ن” [Noon] [5]. Adding kashideh character to some letters [9], the change in the slope of the letters [2], identification and embedding some letters such as setting "لـا" [La] instead of "لا" [La] [10] and ... are other examples of Feature Coding method. Fig. 3 also shows watermarking by using the change of slope "ر" [Ra] for secret information embedding.

In Open Space Coding, the distance between the sentences, distance the existing in the end of lines and distance between the words in the texts which are aligned, are used to insert confidential information [12].

In some cases, information can be hidden by the aid of some symbols in Persian writing such as "." [Full stop] or "," [Coma] in appropriate places. This method will require the identification of suitable locations for placing the symbols. In this case, the amount of hidden data is quite small [13].

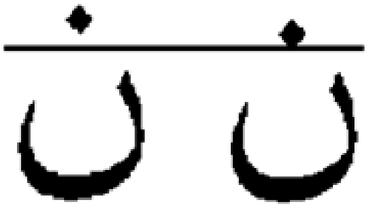


Fig. 2. A sample of watermarking by using relocation of points [5]

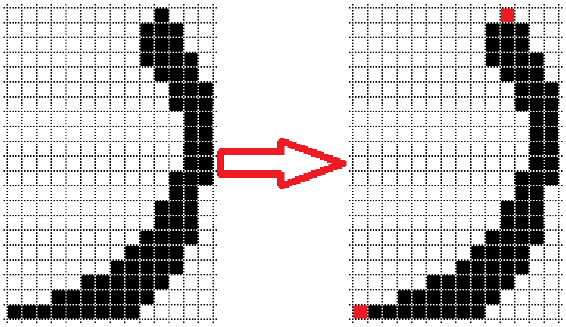


Fig. 3. A sample of watermarking by using sloping character [2]

4. The Proposed Method

In this paper, curvaceous letters "چ ، ج ، خ ، ح" [h, kh, g, ch] respectively are used for watermarking. In this section, Persian text specifications, curvaceous letters features, Persian texts Base line and how to insert and extract information according to proposed algorithm will be described.

4.1.Characteristics of Persian Alphabet

Persian alphabet consists of 32 letters that some of them may be written in four different forms. For example, the letter "ح" at the beginning of word will be written "حـ" at the middle of a word will be "ـحـ" and at the end of a word will be either "ـح" or "ح" form. The specific property of Persian and Arabic unlike English that of is that Not only they are written from right to left side, but the letters also are stuck together in the texts of both languages [11].

Another characteristic of Persian language is the abundance of points in its letters. Although in English letters there are points, there is a huge difference between the two languages in this respect. In English, only two letters of small "i" and small "j" have point while in Persian 18 letters out of 32 alphabet letters have points. These pointed letters in these two languages have one to three points. One point can be placed, on top of a letter such as "ت", or in the middle of a letter such as "ج", or at the bottom of a letter such as "ب" [11].

Curvaceous letters are also of high frequency letters in Persian texts. Four letters "ح", خ","ج" and "چ" have the same appearance, and the only difference is in number and location of points. Therefore, the features of curvaceous letters can be used to secret information embedding. Table 1 shows frequency of curvaceous letters in a sample Persian text.

Table 1. Number of curvaceous letters in a sample Persian text

Number of Words in The Text	653
Number of Letters in The Text	2243
Number of Curved Letters in The Text	457
Percentage of Curvaceous Letters in The Text	20.37%

In general, the Persian letters can be classified into curvaceous and non-curvaceous letters as shown in Table 2.

Table 2. Category of Persian Alphabet Letters Based on Their Curving Shape

Curved Letters	**Curve Toward Right**	ج – چ – ح – خ – ع – غ
	Cure Upward	س – ش – ص – ض – ق – ل – ن – ی
Non-Curved Letters	الف – ب – پ – ت – ث – د – ذ – ر – ز – ژ – ط – ظ – ف – ک – گ – م – و – ه	

4.2.Basic lines of Persian Texts

After digitalizing the image of a text, it is possible to show it in form of two dimensions array f as in relation (1).

$$f(x,y) \in \{0,1\} \qquad x = 0,1,\dots,w-1$$

y=0,1,…,$l-1$ (1)

In relation (1), f (x, y) is the lighting intensity of located pixels at coordinates (x, y). The amount of "f" is always either equal to one or zero, because the image of the texts has always a black and white nature and are considered binary. X-Axis is the horizontal line of text image (along the lines) and Y-Axis shows the vertical line of the text image. "W" is the number of pixels along horizontal axis (the width of the image), and "L" is the number of pixels along vertical axis (the length of the image). The amount of "W" and "L" depends on the image resolution. The horizontal profile function that is used to identify and separate the text lines, can be defined as relation (2).

$$h(y) = \sum_{x=0}^{w-1} f(x,y) \quad y = b, b+1, \dots, e \qquad (2)$$

In relation (2),"b" and "e" are the beginning and ending points of a line block of text image respectively. According to relation (2), horizontal profile is composed of the sum elements of the f array for each row of y. The curve of horizontal profile for a Persian text with 8 rows, is shown in Fig. 4. As it is observed, the curve of each line block has a maximum peak which is seen almost at the middle of the curve of the line. It lays on the fact that the stretch of the most letters and words in Persian is along horizontal line, and the physical shape of the letters are so that the density of pixels in a point on Y- axis (along a row of pixels per line of text) has the highest amount. The place of this maximum peak is considered the same baseline in Persian texts lines [2].

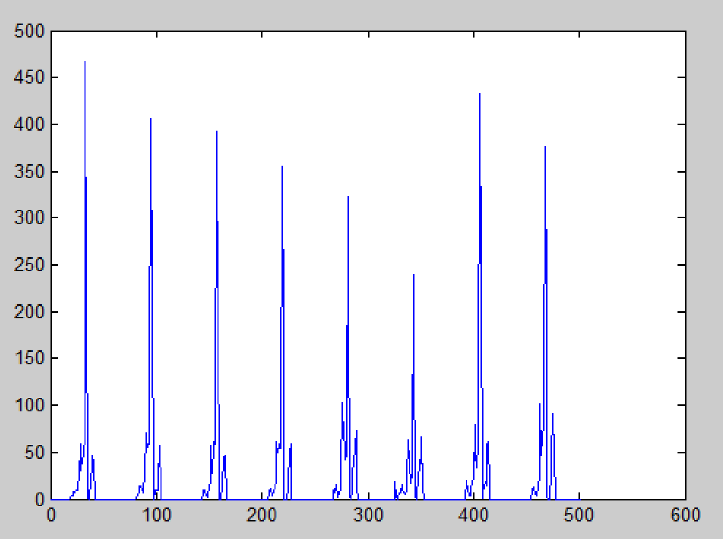


Fig. 4. Horizontal profile curve for a Persian text, including 8 lines

4.2.Curvaceous Letters Properties

The set of Letters "ح" ,"خ" ,"ج" and"چ" are the letters which are used them to watermark information embedding in this paper. These four letters can be called the curvaceous letters with a curve toward right side. Using the specific properties of the four mentioned letters, these letters can be scanned easily in the image of a text and can be used for water mark data. The following brief description of some of properties related to the four mentioned letters:

1- The four letters "ح" ,"خ" ,"ج" and"چ" are only different in the number of their points, thus separation is limited only to recognize the shape of "ح".

2- The way the curvaceous letters with Curve toward right stand is such that parts of the pixels of these letters are located above the base line and the other are below the base line. In Fig. 5 base line (red line) and the location of the curvaceous letters with curve toward right side is shown, compared to the base line. The expressions in the Fig.8 is An example of Persian lyrics

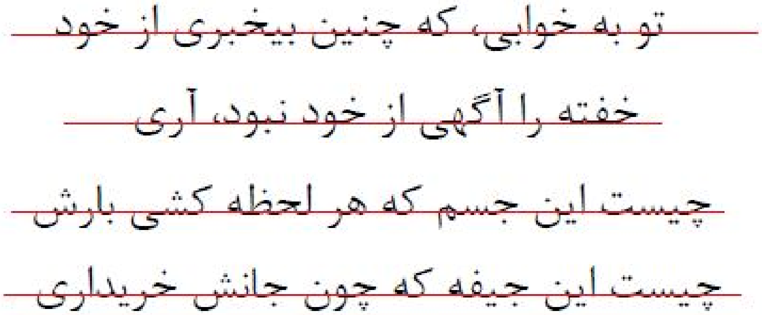

تو به خوابی، که چنین بیخبری از خود
خفته را آگهی از خود نبود، آری
چیست این جسم که هر لحظه کشی بارش
چیست این جیفه که چون جانش خریداری

Fig. 5. The placement of the curvaceous letters with curve toward right compared to the base line

3 - The curvaceous letters which curve to right side in Persian texts appear in two forms of joined and separated; but their shapes don't differ a lot in either case.

4- The Fig. 6 illustrates some forms of letter "ح" in various fonts.

Since most of Persian users prefer “B Nazanin” font to type their articles, theses, and etc, the focus in this paper will be on the mentioned font.

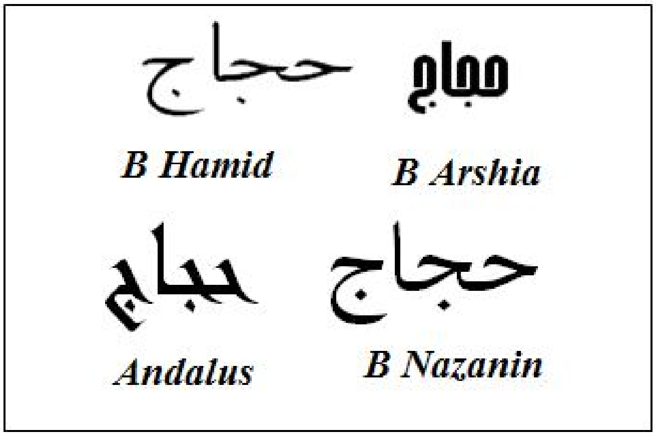


Fig. 6. Various forms of letter "ح"[h] in different fonts

4.3.Data Embedding Algorithm

The inputs include the main Persian text image, the curving of letter "ح" in main image, curve parameter of "ح" letter after watermarking and the purposed watermark signal bits. The output is watermarked image of Persian text which includes information of purposed watermark. Process of embedding of bits is such that after scanning the curvaceous letters with right curving in the main image, the process of embedding watermark signal bits begins from the first line. If watermark bit is one, the curve of the letter to certain amount according to the bit which is modified. In the case that watermark signal bit is zero curve parameter of the letter remains unchanged.

4.4.Data Extraction Algorithm

The inputs are image of watermarked Persian text, curve parameter of curvaceous letters with right curve. The output is sender's purposed watermark signal data bits. Process of extraction bits is such that after separating curvaceous letters (curvaceous letters with right curving) and calculating curve parameters of these letters in watermark image, the bits extraction begins from the first line what is done in the process of data embedding. In figure 7 embedding one bit into "ح" letter is shown. The embedding action is done for a popular font type, i.e. BNazanin in size of 14. In figure 7 blue line represents base line and the red pixel represents the hidden bit in "ح" letter.

5. RESULTS AND DISCUSSION

In this method, regarding the fact that the hidden bit is set in the curvaceous part of letter "ح" in comparison with a letter such as "ر" which the purposed pixel is set at the end point of the letter [2] and is embedded as linear form, and also in comparison with way that uses point moving in the letter "ن" which by having the desired font and comparing it with the main letter, point moving is clearly understandable; the proposed method in this paper has a better watermarking power, so it is not easily discoverable. Fig. 7 shows Shape of letter "ح" before and after watermarking according to proposed algorithm. Furthermore, since the letters "ح" ,"خ" "ج" and"چ" form a considerable Percentage of Persian words, a big capacity of watermarking should be expected. It is noticeable that the detection results have been obtained under the condition of no noise.

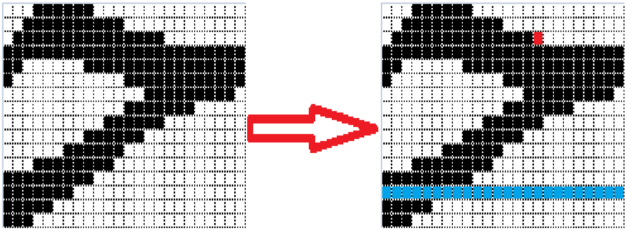
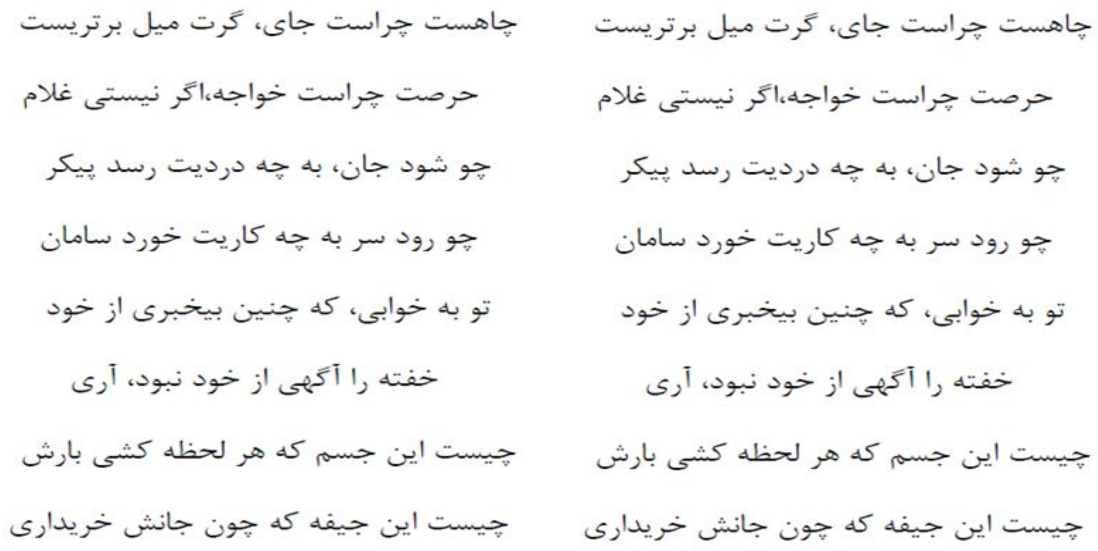


Fig. 7. Shape of letter "ح"[h] before and after watermarking according to proposed algorithm

The proposed watermarking algorithm is performed on a give Persian text image which includes 32 lines, and a section of the text image before and after watermarking has been shown in Fig. 8.



چاهست چراست جای، گرت میل برتریست
حرصت چراست خواجه،اگر نیستی غلام
چو شود جان، به چه دردیت رسد پیکر
چو رود سر به چه کاریت خورد سامان
تو به خوابی، که چنین بیخبری از خود
خفته را آگهی از خود نبود، آری
چیست این جسم که هر لحظه کشی بارش
چیست این جیفه که چون جانش خریداری

چاهست چراست جای، گرت میل برتریست
حرصت چراست خواجه،اگر نیستی غلام
چو شود جان، به چه دردیت رسد پیکر
چو رود سر به چه کاریت خورد سامان
تو به خوابی، که چنین بیخبری از خود
خفته را آگهی از خود نبود، آری
چیست این جسم که هر لحظه کشی بارش
چیست این جیفه که چون جانش خریداری

a) Original Text Image

b) Watermarked Text Image

Figu. 8. Part of the Original Text Image and Watermarked Text Image by Proposed Algorithm

In this project, we check the pages of some Iranian newspapers and news sites for computing the capacity of an article for hidden data. Table 3 shows the comparison of the capacity of proposed watermarking algorithm with other methods. All the articles selected on 16 September 2012.

Table3. Comparison of the capacity of proposed watermarking algorithm with other methods

Method	Dot[5] (Bit)	"La" [10] (Bit)	Word Shift Coding (Bit)	Line Shift Coding (Bit)	Proposed (bit)
Jame Jam newspapers	7581	59	1370	187	1617
Eghtesad newspapers	6496	87	1180	167	1348
Resalat newspapers	6184	40	980	109	1293
Fars news sites	4357	32	778	102	1076

It is noticeable that since that curvaceous letters are high frequency letters in Persian texts the proposed algorithm on this letters has more capacity in comparison with "La" method, line shift coding and word shift coding methods in storing data.

The Advantages of the proposed method include:

1. By this method, a large volume of information can be hidden in text, because a large number of curvaceous letters
2. Due to the lack of a strong OCR program for Persian and Arabic languages, the printed text cannot be easily converted into a simple text thus destroying the hidden information is difficult.
3. The text containing hidden phrases is not specific to computer and the hidden information can also extracted from printed text. In order to recover the information in case of printed text, the text should be scanned and then subjected to the relevant program.
4. The hidden text is resistant to enlargement or downsize and these changes do not destroy the hidden information.
5. since that in proposed method, information are embedded in Pixels of curved letters, the proposed method has a reasonable invisibility in comparison with the previous methods.

The disadvantages of the proposed method include:

1. The information is lost in case of retyping.
2. The output text has a fixed frame due to the use of only one font.
3. Due to the lack of good OCR program for Persian and Arabic languages, using this method in texts that are printed and then scanned is difficult.

6. Conclusion

In this paper, a new method is proposed for Arabic and Persian text watermarking which has used the appearance features of curvaceous letters for confidential information embedding. In comparison with the conventional method of Persian text watermarking, the advantages of the proposed method which have been run by the aid of MATLAB software are summed up as follow:

Since it doesn't require the original text while information extraction, the proposed method will be considered as a sub method of blind methods. The proposed method has a reasonable invisibility in comparison with the previous methods. Because the subtle changes created in curvature of some curvaceous letters which are placed among numerous of words and letter, are completely unclear to human eye; and that only the intended recipient of watermarked text who has appropriate watermarking algorithm will be able to scan the hidden data. Since there is no a good and reliable OCR for Persian language, the printed text is not easily convertible to a simple text so destroying the hidden data will be quite a hard job. Meanwhile, the hidden information is quite robust against shrinking or enlarging the image text size and will not be lost. Since that curvaceous letters are high frequency letters in Persian texts, the capacity of the proposed algorithm is good. Moreover, considering the similarity of letters in Persian, Arabic and Urdu languages, this method may be applied for few languages other than Persian.

REFERENCES

1. Jaseena, K. U., and A. John, 2011. Text Watermarking using Combined Image and Text for Authentication and Protection, International Journal of Computer Application, pp: 8-13.
2. Davarzani, R., K. Yaghmaie, 2009. Farsi Text Watermarking Based on Character Coding, International Conference on Signal Processing Systems, pp:152-156.
3. Gutub, A., 2010. Improved Method of Arabic Text Steganography Using The Extension 'Kashida' Character, Bahria University Journal of Information & Communication Technology, pp: 67-72.
4. Huijuan Y., and A. C. Kot, 2004. Text Document Authentication by Integrating Inter Character and Word Space Watermarking, IEEE International Conference on Multimedia and Expo, pp: 955– 958.
5. Shirali-Shahreza, M.H and M. Shirali-Shahreza, 2006. A New Approach to Persian/Arabic Text Steganography, International Conference on Computer and Information Science (ICIS), pp: 310-315.
6. Culnane, C., and H. treharne, 2007. Improving Mutli-Set Formatted Binary Text Watermarking Using Continuous Line Embedding, Second International Conference on Innovative Computing, Information and Control, pp: 287-292.
7. Low, S.H., N.F.Maxernchuk, J.T.Brassil, and L.O.Gorman, 1995. Document Mrrking and Identification Using Both Line and Work Shifting, Proceedings of the Fourteenthannual Joint Conference of the IEEE Computer and Communications Societies, pp: 853 – 860.
8. Kim, Y. W., K. A. Moon, and I. S. Oh, 2003. Text Watermarking Algorithm Based on Word Classification and Inter Work Space Statistics, Proceeding of Seventh International Conference on Document Analysis and Recognition (ICDAR), pp: 775–779.
9. Adnan, A., and M. Fattani, 2007. A Novel Arabic Text Seteganography Method Using Letter Points and Extensions, WASET International Conference on Computer Information and Systems Science and Engineering (ICCISSE).
10. Shirali-Shahreza, M., 2007. A New Persian/Arabic Text Steganography Using "La" Word, International Joint Conference on Computer, Information, Systems Sciences and Engineering (CISSE), pp: 339–342.
11. Shirali-Shahreza, M., 2008. Pseudo-Space Persian/Arabic Text Steganography, Computers and Communications, ISCC 2008. IEEE Symposium, pp: 864-868.
12. Singh, K.U. H., P.K. Singh and K. Saroha, 2009. A Survey on Text Based Steganography, In Proc. 3rd National Conference, Indiacom-2009 Computing for Nation Development India.
13. Bender, W., D. Gruhl, N. Morimoto, and A. Lu, 1996. Techniques for Data Hiding, IBM Systems Journal, pp: 313-336.