Abstract—In this paper a novel blind method for watermarking of Farsi (Persian language) texts based on employing the existing common sloping letters is presented. The proposed method embeds the watermark data in four common sloping letters by readjusting the amount of their slopes. This approach can be categorized under character coding methods. The presented algorithm is so simple to implement and experiments have shown its higher capacity and transparency comparing to similar text watermarking methods. Noticing similarities of Arabic and Urdu characters with Farsi characters, the presented method can easily be implemented on those scripts.

Keywords—character coding; information security; sloping letters; text watermarking

I. INTRODUCTION

Nowadays more and more attention is paid to digital watermarking technology. Digital watermark is the information embedded into host digital multimedia such as image, video, audio, text, and software. The principles of the image, video and audio watermarking are similar in that they make use of the redundant information of their host media as well as the characteristics of human visual or auditorial system. Among these media text documents show very peculiar properties: binary nature, block/line/word patterning, and clear separation between foreground and background areas. In other words, in contrast to image, video and audio there is no redundant information in text documents. Therefore, text watermarking techniques are basically different from that of non-text watermarking. Therefore text watermarking algorithms have difficult task in satisfying the requirements for transparency, robustness and etc.

Data hiding on binary images can be done either on low level such as flipping pixels from black to white and vice versa or on a higher level such as modifying width of strokes and spacing between words and characters [1]. In this paper, our focus is on data hiding in Farsi and Arabic text images using a high level technique. Some text watermarking techniques have already been presented. Line-shift, word-shift and character coding are three main approaches adopted in text watermarking [2]-[9]. In line-shift coding each evenly numbered line is slightly shifted up or down according to the value of a specific bit in the payload [2]. If the bit is one, the corresponding line is shifted up; otherwise, the line is shifted down. The odd lines are considered as control lines, hence they remain unmoved. They are used as references for measuring and comparing the distances between lines in the decoding stage. Decoding is achieved by comparing the distances between the bases of the lines [3], or the distances between the centroids of the lines [4]. The baselines in the original document are normally uniformly spaced; hence, the original document is not needed for the detection process that uses baselines. However, the centroids are not necessarily uniformly spaced; therefore, the original document must be used in the detection process that uses centroids. In word-shifting method [5], each line is first divided into groups of words. Each group has a sufficient number of characters. Then, each even group is shifted to the left or the right according to the value of a specific bit in the payload. The odd groups are used as references for measuring and comparing the distances between the groups during the decoding stage. Correlation and centroid-based methods have been suggested for detecting the watermark data. Both of these methods require the use of the original document, especially when the inter-word spacing is variable. In [6] a word-shifting algorithm is developed that modifies the inter-word spaces to match a sine waveform. The signals are encoded in the phase, amplitude, and frequency of the sine waves. For signal insertion, spaces between characters should also be adjusted. Character coding approach is a class of techniques which embed a mark by altering a particular feature of an individual character. Examples of possible feature alterations include a change to an individual character’s height, or its position relative to other characters [7]. A watermarking technique for Chinese text based on the mathematical expression of a Chinese character and its automatic generation is proposed in [8]. Watermarking signals are embedded into some Chinese characters with occlusive components by readjusting the size of the closed rectangular regions in these components. Noticing the existence of too many points in Farsi and Arabic phrases, by vertical displacement of the points, watermark bits could be embedded in such texts [9].
In contrast to English and Chinese languages very few watermarking techniques have been presented for Farsi language based on the unique characteristics of its characters. In this paper we present a novel character coding method for Farsi text images. This technique is based on use of four very common Farsi characters which show specific slope in their shapes. The watermark bits are hidden in these letters by altering their slopes according to the bit to be hidden.

II. PROPOSED METHOD

Investigations in Farsi texts have revealed that there are four letters which show specific slope in their shapes. In this section, the basic technique which uses these letters to hide data is presented.

A. Baselines of Farsi Texts

Following digitization a page image is represented by a two-dimensional array with elements

\[ f(x, y) = 0 \text{ or } 1, \quad x = 0, 1, ..., W, \quad y = 0, 1, ..., L \]

where \( f(x, y) \) represents the intensity of the pixel at position \((x, y)\). For a black and white image, \( f(x, y) \in \{0, 1\} \). Here, \( W \) and \( L \), whose values depend on the scanning resolution, are the width and length of the page in pixels, respectively. The image of a text line is simply presented by a function which describes the pixels restricted to the region of that line, that is:

\[ f(x, y) = 0 \text{ or } 1, \quad x = 0, 1, ..., W, \quad y = t, t + 1, ..., b \]

where \( t \) and \( b \) are the top and bottom ‘boundaries’ of the text line, respectively. For instance, we may take \( t \) or \( b \) to be the mid-point of the interline spacing. The horizontal profile of a text line can be described by:

\[ h(y) = \sum_{x=t}^{W} f(x, y), \quad y = t, t + 1, ..., b \]

is the total number of 1’s along that specific horizontal scan lines \( y \). Fig. 1 shows the horizontal profile of a Farsi text. As shown in this figure, each text line demonstrates a distinctive peak in its profile which can be taken as the baseline for that line.

B. Sloping Letters in Farsi Text

There are 32 letters in Farsi script. In this paper four letters \( \{\text{ض، ش، ن، }\} \) are singled out as letters with specific slopes in their shapes which can be used for embedding watermark data (among these, three letters \( \{\text{ش، ن، }\} \) are also used in Arabic texts). We briefly call these letters as ‘sloping letters’.

C. Popularity of Sloping Letters in Farsi Text

The introduced letters \( \{\text{ش، ن، }\} \) are among the mostly used letters in Farsi texts. Table I shows that in 20 typical randomly selected texts, the share of these letters in the texts is about 21%, while they just make 12.5% of the whole Farsi
letters. This indicates the capacity provided by using these letters for hiding data.

### TABLE I. Popularity of the Sloping Letters in 20 Samples of Farsi Pages

<table>
<thead>
<tr>
<th>No. of words</th>
<th>No. of sloping letters</th>
<th>No. of letters</th>
<th>Percentage of sloping letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>10680</td>
<td>39740</td>
<td>8293</td>
<td>20.87%</td>
</tr>
</tbody>
</table>

#### D. Computing the slopes

**Slope of letter \( \{ j \} \):** The slope of letter \( \{ j \} \), as depicted in Fig. 3 can be calculated as:

\[
m_r = \tan(\hat{R}) = \frac{L_r}{W_r}
\]

where \( \hat{R} \) is the angle of the line connecting the beginning and the end of letter \( \{ j \} \) with horizon and \( L_r \) and \( W_r \) denote the vertical length and the horizontal width of the letter \( \{ j \} \), respectively.

**Slope of letter \( \{ k \} \):** The slope of the letter \( \{ k \} \) is described in Fig. 4 and is calculated as:

\[
m_v = \tan(\hat{V}) = \frac{L_v}{W_v}
\]

where \( \hat{V} \) is the angle of the line connecting the beginning and the end of the letter \( \{ k \} \) with horizon and \( L_v \) and \( W_v \) denote the vertical length and the horizontal width of the letter \( \{ k \} \), respectively.

#### E. Detection and Extraction of Sloping Letters in Farsi Text

As described earlier in the paper, the algorithm for detection and extraction of sloping letters is limited to detection of the two letters \( \{ j \} \) and \( \{ k \} \). After separating text image lines and specifying the top and the bottom boundaries of each text line by tracing the matrix of each line block, sloping letters could be detected using a special algorithm which traces their shapes. Since the main bodies of sloping letters \( \{ j \} \) and \( \{ k \} \) are located under the baseline, tracing operation is started from the baseline in each line block matrix. In this way, detection and extraction of the sloping letters consist of the following steps:

1) Text lines separation.
2) Specifying the top and the bottom boundaries of each text line.
3) Tracing the line block matrix and detection of the sloping letters \( \{ j \} \) based on some given conditions for the shape of the letters \( \{ j \} \).

#### F. Embedding Algorithm

**Input:** a host Farsi text image, \( m_r \) and \( m_v \) slope parameters in original text image, \( m'_r \) and \( m'_v \) equivalent slope parameters in watermarked image, watermark signal bits \( W \).

**Output:** the watermarked Farsi text image in which the watermark \( W \) is embedded.

**Data embedding process:** extraction of the sloping letters is started from the first line in the original text image. For each extracted letter, if the corresponding bit of \( W \) is 1, then \( m_r \) is changed to \( m'_r \) (or \( m_v \) to \( m'_v \)) for any of the four detected letters \( \{ j \} \) and \( \{ k \} \), respectively. If the watermark bit is 0, no change is imposed on the slope.

#### G. Extracting Algorithm

**Input:** a watermarked Farsi text image and the slope parameters \( m_r \) and \( m_v \).

**Output:** watermark signal bits.
Data extracting process: at first, all sloping letters are extracted from the watermarked text image and the slope of each letter is computed separately. Calling \( m_{rw} \) and \( m_{w} \) as the slopes of letters \( \{ j \} \) and \( \{ j \} \) in watermarked text image, the data extracting algorithm can be described as:

\[
\begin{align*}
| m_{rw} - m_r | > \epsilon, & \quad \text{Watermarked bit is 1} \\
| m_{rw} - m_r | = 0, & \quad \text{Watermarked bit is 0}
\end{align*}
\]

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

The proposed method was implemented on a sample Farsi text image. The results are illustrated in Fig. 5. The slope parameters which are used in this implementation are:

\[
m_r = \tan(\hat{\theta}) = \frac{21}{12} = 1.75 \quad m_w = \tan(\hat{\theta}) = \frac{22}{11} = 2
\]

where \( \hat{\theta} \) and \( \hat{\theta}^r \) are the equivalent angles in watermarked text image. It can be seen that the distortion produced in the watermarked image is reasonably imperceptible.

Comparing with conventional line-shifting and word-shifting methods [2]-[6], the advantages of the proposed method can be summarized as:

Imperceptibility: The experimental results show that the proposed method is more transparent than the existing techniques such as line-shift coding and word-shift coding, basically since it uses some, and not all, letters while changing the characteristic which is fairly imperceptible to human visual system.

Capacity: The capacity of the proposed method is fairly higher when compared to other conventional methods. Table II shows capacities of line-shift, word-shift, and our proposed method in the sample Farsi text.

Blindness: Our proposed method can be classified as a blind method, i.e. the decoder does not need an original copy of the document to extract the hidden data. Legal receivers of the watermarked documents can extract the watermarked data just by using slope parameters. Furthermore, the extraction of the watermark doesn’t require the control line or control word, which makes the method more immune to the errors occurred in page segmentation.

IV. CONCLUSIONS

In this paper, a blind text watermarking method for Farsi texts was presented which uses the unique characteristics of some of very common letters used in Farsi texts. This method is based on the existence of four sloping letters in majority of words in Farsi texts. Information was hidden in text by changing the slope of these sloping letters. Noticing the similarities of Urdu and Arabic scripts with Farsi, this method can also be implemented on Urdu and Arabic texts. Experiments show that this method presents higher capacity and imperceptibility comparing with line-shifting and word-shifting methods. Hence, it can easily be used in many applications such as copyright protection, document authentication, digital notarization to name a few.

<table>
<thead>
<tr>
<th>Type of coding</th>
<th>Capacity in the sample Farsi text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line-shift coding</td>
<td>15 bits</td>
</tr>
<tr>
<td>Word-shift coding</td>
<td>222 bits</td>
</tr>
<tr>
<td>Proposed method</td>
<td>335 bits</td>
</tr>
</tbody>
</table>

REFERENCES

Figure 5. Hiding effects. (a) Original host image, (b) Watermarked image by proposed method.