Hindawi Publishing Corporation
Journal of Engineering
Volume 2013, Article ID 465469, 11 pages
http://dx.doi.org/10.1155/2013/465469

Research Article

Printed Persian Subword Recognition Using Wavelet Packet Descriptors

Samira Nasrollahi and Afshin Ebrahimi

Faculty of Electrical Engineering, Sahand University of Technology, Tabriz, Iran

Correspondence should be addressed to Afshin Ebrahimi; aebrahimi@sut.ac.ir

Received 30 May 2013; Accepted 7 August 2013

Academic Editor: Yangmin Li

In this paper, we present a new approach to offline OCR (optical character recognition) for printed Persian subwords using wavelet packet transform. The proposed algorithm is used to extract font invariant and size invariant features from 87804 subwords of 4 fonts and 3 sizes. The feature vectors are compressed using PCA. The obtained feature vectors yield a pictorial dictionary for which an entry is the mean of each group that consists of the same subword with 4 fonts in 3 sizes. The sets of these features are congregated by combining them with the dot features for the recognition of printed Persian subwords. To evaluate the feature extraction results, this algorithm was tested on a set of 2000 subwords in printed Persian text documents. An encouraging recognition rate of 97.9% is got at subword level recognition.

1. Introduction

Optical character recognition (OCR) is one of the oldest subfields of pattern recognition with a rich contribution for the recognition of printed documents. The ultimate goal of OCR is to imitate the human ability to read—at much faster rate—by associating symbolic identities with images of characters [1]. In recent years, many OCR researches have extracted subword features. In the number of these researches, holistic shape information from subwords is extracted for modeling subwords [2]. In this work, we want to extract holistic shape features of printed Persian subwords using wavelet packet transform to build a pictorial dictionary.

Feature extraction is a vital step for pattern recognition and optical character recognition systems [3], especially for printed Persian OCR as there are varieties of characters depending on the fonts and sizes.

One of the main concerns of designing every OCR system is to make it robust to the font and size variations [4]. It is clear that OCR of multifont documents is more difficult than OCR of single-font documents. Design of an OCR engine which can recognize subwords independent of their font types and size variations is not impossible, but certainly it is very difficult and inefficient, because subwords take different shapes in different fonts [5].

There has been a great attempt to produce Omnifont OCR systems for Persian/Arabic languages, but the overall performance of such systems is far from being perfect. Persian written language which uses modified Arabic alphabet is written cursively, and this intrinsic feature makes it difficult for automatic recognition [6].

Many feature extraction methods have been reported such as various moment features, gradient-and distance-based features, geometric features, and transform domain features. Wavelet transform has been widely used in image processing and signal processing for image/signal enhancement, denoising, texture segmentation [7] based on its properties of short support, orthogonality, symmetry, higher order of vanishing moments, and more importantly its multiresolution decomposition analysis. Its application to pattern recognition, especially to OCR, is a relative new research field. In this paper, we propose a new technique based on feature extraction using wavelet packets for the recognition of Farsi text.

There are two approaches to the automatic recognition of cursive scripts: holistic and segmentation-based [1, 8, 9].

In the first approach, each word is treated as a whole, and the recognition system does not consider it as a combination of separable characters. The second strategy, which owns the majority in the literature, segments each word to containing characters as the building blocks and recognizes each character then. In this work, the first approach is used.

There are many works reported on the recognition of Arabic and Farsi texts [1, 10–23]. There are few works based on holistic recognition of Farsi/Arabic subwords by their shape information.

Albadr and Haralick have described the design of an Arabic word recognition system that does not segment word into characters [1]. This system recognizes the input word image by detecting a set of shape primitives on the word as features for describing the recognized word. For recognizing the words in Arabic texts, each region of the word has been modeled with a set of special signs as features. Thus, the shape descriptor is obtained from the recognized word.

The feature extraction stage, characteristic loci features are extracted from printed Persian subword images to describe their shapes [10]. In feature extraction stage, the number of crossings with subword bodies is restricted to 2. Using PCA algorithm, 12 uncorrelated features are selected.

In [11], subword features have been extracted from upper contour of each subword image. These features are given into a clustering stage to construct a dictionary. The subwords with same features or entry constitute their own neighbors within the dictionary.

Azmi and Kabir have proposed three methods to construct pictorial dictionaries [12]. They used loci characteristic features, Fourier descriptors, and upper contour labels of printed Persian subwords to construct three pictorial dictionaries. In each method based on its description the entry for that subword is defined.

In [13], a two-step method for the recognition of printed Farsi subwords has been proposed. Printed Farsi subwords using the loci characteristic features and the k-means method have been clustered to 300 clusters, and 10 closest clusters have been assigned. Then, Fourier descriptors of the subword contour have been used to classify the input subword images into the members of these 10 clusters. Postprocessing has been done using the dot features of the subwords to recognize correct subword.

Ebrahimi and Kabir have reported on the use of characteristic loci features to cluster 113,340 printed Farsi subwords of 4 fonts and 3 sizes to 300 clusters, based on their holistic shapes [14]. The feature vectors are compressed using PCA. A pictorial dictionary that can be used in a word recognition system to eliminate the search space is yielded. The mean of each cluster is used as its entry in the pictorial dictionary.

A method for recognition of Persian characters in machine printed documents has been developed based on the morphological hit/miss transformation [15]. Morphological operators, especially Hit/Miss operator, have been used to descript each subword, and by using a template matching approach it has been tried to classify generated description.

Ben Amor and Ben Amara presented a hybrid technique based on both neural networks and the hidden Markov models for classification in a multifont Arabic OCR system [16]. This hybrid approach has been tested by a method based on the Hough transform for features extraction. Overall recognition rate for the HMM/MLP approach using the Hough transform on a set of 85.000 samples of characters corresponding to 5 different fonts in Arabic writing was 97.36%.

Menhaj and Adab have proposed a new segmentation and recognition method for multifont, multisize Farsi/Arabic texts using multilayer feed forward neural networks [17]. The OCR recognition method has used MLP type neural networks with Fourier descriptors.

Aghbari and Brook proposed a novel holistic technique for classifying and retrieving historical Arabic handwritten manuscripts (HAH manuscripts) [18]. First, the HAH manuscript's image is segmented into words, and then each word is segmented into its connected parts. Second, several structural and statistical features which are devised for Arabic texts are extracted from these connected parts and combined to represent a word with one consolidated feature vector. Finally, a neural network is used to learn and classify the input vectors into word classes. These classes are then utilized to retrieve HAH manuscripts. This technique is robust to different styles and font sizes.

A word-level recognition system for machine-printed Arabic text has been reported in [19]. The Arabic recognition system has computed a vector of image-morphological features on a query word image. In the recognition stage, this vector has been matched against a precomputed database of vectors from a lexicon of Arabic words. Vectors from the database with the highest match score are returned as hypotheses for the unknown image.

In the proposed AHTR (Arabic handwritten text recognition) system [20], smoothing and segmentation processes have been used to segment the subwords of text into characters. The discrete wavelet transform (DWT) is invoked to extract efficient features of each isolated character which results from the segmentation stage. A 24-dimensional feature vector representing the relative energy and zero crossing was extracted from the output signal of the DWT. The recognition of each character is continued with classifying a set of characters using these features.

In [21], five various methods have been used to extract the Arabic word image features. In one method, two-dimensional discrete wavelet transform (DWT) is used to extract the features, as it is well acknowledged that DWT coefficients can provide a powerful insight into an image's frequency and spatial characteristics. During the implementation stage, each word image is decomposed by 4 levels of wavelet transform. Wavelet features used in this system are extracted from the $LL4$ subband that represents the low-frequency components in level 4.

The wavelet transform is a tool that has been applied in many disciplines, including image processing [22, 23]. Due to the multiresolution property, it decomposes the signal into different frequency scales. For a given image, the wavelet packet transform produces a low-frequency subband image reflecting its basic shape and three subband images that contain the high-frequency components of the image at horizontal, vertical, and diagonal directions. These components

can be used to construct the feature vector in the recognition systems.

In the recent years, wavelet analysis has been successfully applied in the field of pattern recognition. Wavelet descriptors of a character, as a set, can be used to replace the combination of many conventional features used in character recognition systems. They serve as excellent features for character recognition, are reasonably insensitive to intraclass variation, and exhibit good interclass separation.

There are no works based on holistic recognition of printed Farsi subwords by their wavelet packet features which are font invariant and size invariant. In this paper, we are going to present a novel shape descriptor for the recognition of the printed Persian subwords independent of font and size variations that build a dictionary. Here, at first, we normalize subword size. Afterwards wavelet packet transform is performed to get proper features. In the recognition process, we use the dot feature vectors to post-process the printed Persian subwords. The integration of these methods provides a high accuracy and reliable recognition method.

The remainder of this paper are organized as follows. In Section 2 Farsi writing characteristics are presented. The dataset that we use for building a dictionary is introduced in Section 3. Wavelet packet transform theory is summarized in Section 4. Then a brief overview of feature extraction architecture and dictionary building is explained in Section 5. Section 6 presents the recognition systems using pictorial dictionary and the postprocessing stage. The paper concludes in Section 7.

2. Some Characteristic of Farsi Script

In this section, we will briefly describe some of the main characteristics of Persian script to point out the main difficulties which an OCR system should overcome. Some of these characteristics greatly complicate recognition. The most notable feature of Farsi writing is its cursiveness. Compared to other features, the cursiveness of Farsi/Arabic words poses the most difficult problem for recognition algorithms [1]; this is why segmentation is a crucial step for many Farsi/Arabic character recognition systems. Many recognition errors are attributed to the segmentation phase. In this work, we describe the design and implementation of a Farsi subword recognition system. To recognize a subword, the system does not segment it into characters in advance; rather, it recognizes the input subword using wavelet packet features as holistic shape information.

Words are separated by long spaces, and each word consists of one or more isolated segments each of them is called subword. Subwords are separated by short spaces, and each subword includes one or more characters [6]. Figure 1 shows a sample Persian script where "a" represents the space between two different words and "b" represents the short space between subwords.

Some characteristics of Farsi script are as follows.

(i) Unlike English, Farsi texts, are written from right to left [24–26].

(ii) Farsi scripts include 32 characters, and each character can appear in four different shapes/forms depending on the position of the word (beginning form, middle form, isolated form, and end form) [24–26]. Table 1 shows Farsi characters in their different forms.

(iii) The Farsi characters of a word are connected along a baseline [1, 26]. A baseline is the line with the highest density of black pixels. The existence of the baseline calls for different segmentation methods from those used in other unconnected scripts.

(iv) Many Farsi characters have dots, which are positioned above or below the letter body. Dots can be single, double, or triple [26]. Ten of them have one dot ((be) ب, (noon) ن, (za) ظ, (zal) ذ, (fe) ف, (ghayn) غ, (khe) خ, (jim) ج, (ze) ز, (zad) ض); 3 of them have two dots ((ye) ي, (te) ت, (ghaf) ق); 5 of them have three dots ((shin) ش, (che) چ, (pe) پ, (je) ژ, (se) ث); and 14 of them do not have dots ((mim) م, ک (kaf), ل (lam), (ayn) ع).

(v) Different Farsi letters can have the same body and differ in the number and position of dots identifying them, like ((se) ث, (be) ب, (te) ت, and (pe) پ) [14, 24–26].

(vi) Each word, machine-printed or handwritten, may consist of several separated subwords. A subword is either a single character or a set of connected characters [24, 25]. Although seven Farsi characters out of 32 do not connect to their left neighbors or the next letter [14, 24, 25], the others join to the neighboring characters to make a word or subword. These seven letters are ((alef) ا, (dal) د, (zal) ذ, (re) ر, (ze) ز, (je) ژ, and (vav) و). To form a subword, the letters are connected to each other [14]. For example, (samar) ثمر is a subword formed by three letters: (se) ث, (mim) م, and (re) ر. One or more subwords could form a word.

(vii) The neighboring letters in words or subwords may overlap vertically depending on their shapes [1, 14, 24, 25]. Some characteristics of Farsi script are shown in Figure 2 [24, 25, 27]. Table 2 outlines a comparison of the Farsi and Latin Scripts.

3. The Data Set

The used database in this work is prepared from Persian subwords which are printed by a laser printer in four fonts: Nazanin, Zar, Lotus, and Mitra and three sizes: 12, 14, and 16 and scanned in 400 dpi. Overall, our dataset includes 87804 subwords. Thus, in this work 87804 subwords of 4 fonts and 3 sizes are selected for feature extraction to build pictorial dictionary. Some samples of subword images with various

Table 1: Four different shapes of letters in the Persian alphabet [28].

Isolated	Initial	Medial	Final	Roman	Name	Isolated	Initial	Medial	Final	Roman	Name
ا	ا	ـا	ـا	á	alef	ص	صـ	ـصـ	ـص	ṣ	sád
ب	بـ	ـبـ	ـب	b	be	ض	ضـ	ـضـ	ـض	ḍ	zád
پ	پـ	ـپـ	ـپ	p	pe	ط	طـ	ـطـ	ـط	ṭ	tá
ت	تـ	ـتـ	ـت	t	te	ظ	ظـ	ـظـ	ـظ	ẓ	zá
ث	ثـ	ـثـ	ـث	th	se	ع	عـ	ـعـ	ـع	‘	ayn
ج	جـ	ـجـ	ـج	j	jim	غ	غـ	ـغـ	ـغ	gh	ghayn
چ	چـ	ـچـ	ـچ	ch	che	ف	فـ	ـفـ	ـف	f	fe
ح	حـ	ـحـ	ـح	h	he	ق	قـ	ـقـ	ـق	q	qáf
خ	خـ	ـخـ	ـخ	kh	khe	ک	کـ	ـکـ	ـک	k	káf
د	د	ـد	ـد	d	dál	گ	گـ	ـگـ	ـگ	g	gáf
ذ	ذ	ـذ	ـذ	dh	zál	ل	لـ	ـلـ	ـل	l	lám
ر	ر	ـر	ـر	r	re	م	مـ	ـمـ	ـم	m	mim
ز	ز	ـز	ـز	z	ze	ن	نـ	ـنـ	ـن	n	nún
ژ	ژ	ـژ	ـژ	zh	zhe	و	و	ـو	ـو	v/ú	váv
س	سـ	ـسـ	ـس	s	sin	ه	هـ	ـهـ	ـه	h	he
ش	شـ	ـشـ	ـش	sh	shin	ی	یـ	ـیـ	ـي	y/i	ye

Table 2: Ambiguous patterns in printed words.

Characteristics	Farsi	Latin
Justification	R-to-L	L-to-R
Cursive	Yes	No
Diacritics	Yes	No
Number of vowels	3	5
Letters shapes	1–4	2
Number of letters	32	26

Figure 1: Sample of Persian script and virtual baseline [6].

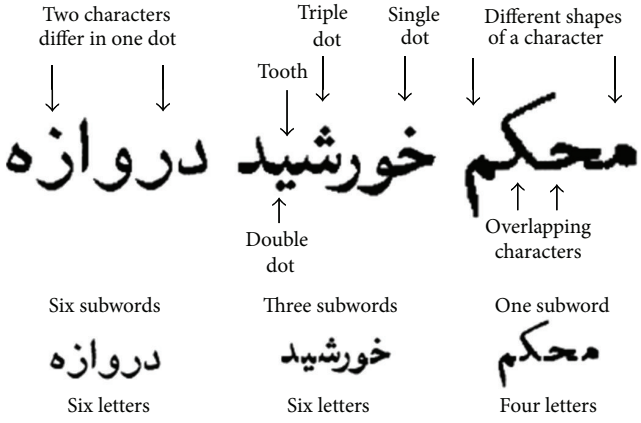

Figure 2: Some features of Farsi script [24].

fonts and sizes in the database are shown in Figure 3. Table 3 shows a sample of subwords that have 12 different shapes of subword “(tasmim) تصمیم”. The reason for using these fonts is their popularity in Farsi magazines, newspapers, books, and official documents.

4. Wavelet Packet Transform Theory

The wavelet packet method is a generalization of wavelet decomposition [29–32] that offers a richer signal analysis than wavelets. In the two-dimensional orthogonal wavelet decomposition procedures, the generic step splits the approximation coefficients into four parts [30, 31, 33, 34]. After splitting, we obtain a subimage of smooth (low pass) coefficients and three subimages corresponding to detail coefficients along horizontal, vertical, and diagonal directions. This is followed by splitting the low-pass subimage, while the successive details are never reanalyzed. So, there is $n + 1$ possible way to decompose or encode the image for an n-level decomposition [31]. In the corresponding wavelet packet situation, each detail coefficient subimage is also

Table 3: "(tasmim) تصمیم" subword shapes printed in four fonts and three sizes.

		Font			
		Zar	Lotus	Mitra	Nazanin
Size	12	تصمیم	تصمیم	تصمیم	تصمیم
	14	تصمیم	تصمیم	تصمیم	تصمیم
	16	تصمیم	تصمیم	تصمیم	تصمیم

Zar	Mitra	Lotus	Nazanin
پز	گلز	گلنا	پستی
چغا	تفتیش	مخفی	یبر
ستد	مخلصا	فیلمسا	بهینه

Figure 3: Sample of printed Farsi subwords.

decomposed into four parts using the same approach as in low-pass subimage splitting [30–33]. So, there are 4^n different ways to encode the image at decomposition level n which provide a better tool for image analysis [31, 33]. The decomposition structure is given in Figure 4(a). The filter used here is symlet 8, and only two levels of decomposition are performed resulting in 16 subimages (subbands).

The standard 2D-DWPT can be described by a pair of quadrature mirror filters (QMF) **H** and **G** [31, 35]. The filter **H** is a low-pass filter with a finite impulse response denoted by $h(n)$. And the high-pass **G** with a finite impulse response is defined by

$$g(n) = (-1)^n h(1-n), \quad \forall n. \tag{1}$$

The low-pass filter is assumed to satisfy the following conditions for orthonormal representation:

$$\begin{aligned} \sum_n h(n)\, h(n+2j) &= 0 \quad \forall j \neq 0, \\ \sum_n h(n)^2 &= 1, \\ \sum_n h(n)\, g(n+2j) &= 0, \quad \forall j. \end{aligned} \tag{2}$$

The 2D discrete wavelet packet transform (2DDWPT) for an $N \times M$ discrete image **A** up to level $p+1$ ($p \leq \min(\log_2 N, \log_2 M)$) is recursively defined in terms of the coefficients at level p as follows [32]:

$$\begin{aligned} C^{p+1}_{4k,(i,j)} &= \sum_m \sum_n h(m)\, h(n)\, C^p_{k,(m+2i,n+2j)}, \\ C^{p+1}_{4k+1,(i,j)} &= \sum_m \sum_n h(m)\, g(n)\, C^p_{k,(m+2i,n+2j)}, \\ C^{p}_{4k+2,(i,j)} &= \sum_m \sum_n g(m)\, h(n)\, C^p_{k,(m+2i,n+2j)}, \\ C^{p+1}_{4k+3,(i,j)} &= \sum_m \sum_n g(m)\, g(n)\, C^p_{k,(m+2i,n+2j)}, \end{aligned} \tag{3}$$

where $C^0_{0,(i,j)}$ is the image **A**. At each step, the image C^p_k is decomposed into four quarter-size images C^{p+1}_{4k}, C^{p+1}_{4k+1}, C^{p+1}_{4k+2}, and C^{p+1}_{4k+3}. The low-pass filter h and the high-pass filter g are differently corresponding to the different wavelet basis functions. Wavelet packet decomposition has been successfully applied to image analysis [32]. It gives another view of the data, which can be useful to detect features. This transform represents relatively recent mathematical developments, and it has not found any applications in printed Persian OCR systems. In this paper, we will evaluate the use of the more general WPT. The subword images are decomposed through the wavelet packet decomposition using the Symlet 8 with basis function up to level two to get the 20 subband images. This results in a total of 20 subbands, four subbands at the first level and sixteen subbands (four for each subband) in the next level as shown in Figure 4(b). Figure 5 shows a two-dimensional example of a subword image for the wavelet packet decomposition with depth 2.

In this paper we propose a method for feature extraction from printed Persian subwords based on a two-dimensional wavelet packet transform.

After extracting all subwords features and building a pictorial dictionary, we use this feature extraction method in a recognition scheme. The proposed recognition scheme consists of two stages: a feature extraction stage for extracting multiresolution features and a postprocessing stage that uses dot features of subwords.

5. Building Pictorial Dictionary Using Wavelet Packet Transform

In this work, we use wavelet packet transform with symlet 8 as basis function to decompose subword image into following subbands and extract features from subband [2 0] of level two. These features are used as subword shape descriptors in a pictorial dictionary and are used in recognition system.

In this paper, we want to extract the feature vector for the subword image that is font invariant and size invariant. This feature vector for one subword with four fonts, Zar, Mitra, Lotus, and Nazanin, and in three sizes 12, 14 and 16 is invariant. The obtained feature vectors are used in dictionary building.

From Section 5.1 to Section 5.3, we illustrate the preprocessing and feature extraction stages to generate 7317 entries

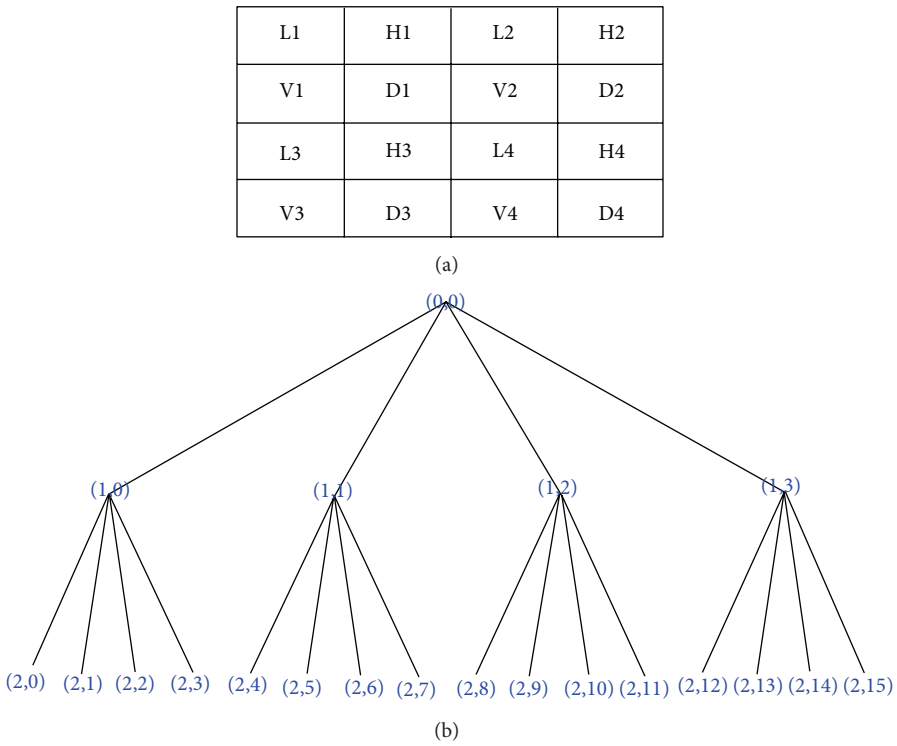

Figure 4: (a) Decomposition structure of wavelet packet transform at level two. (b) Wavelet packet decomposition tree up to level two.

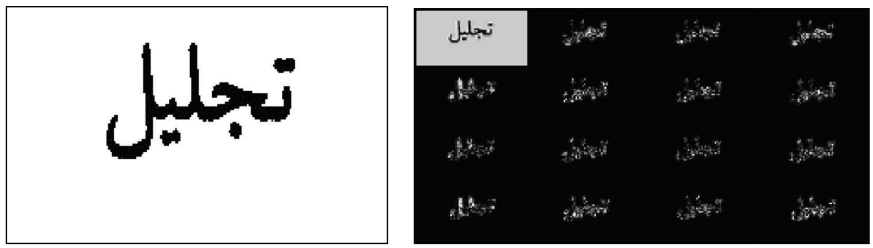

Figure 5: A wavelet packet decomposition: (a) the original image, (b) wavelet packet decomposition subbands in level 2.

that represent 7317 Persian subwords in one pictorial dictionary. Figure 6 shows the flow chart of feature extraction. To obtain the invariant features, preprocessing and feature extraction stages are done 12 times for 12 different shapes of the same subword. Then 12 feature vectors are entered to the averaging stage.

5.1. Preprocessing and Size Normalization. Normalization methods aim to remove the variations of the fonts and obtain standardized data [8].

5.1.1. Size Normalization. It is used to adjust the subword size to a certain standard. Since the sizes of Persian subwords greatly vary, size normalization is often used to scale subwords to a fixed size and to center the subword before feature extraction. Size normalization is done as a preprocessing step in Persian subword recognition systems. Before size normalization, the subword image is cropped with the minimum bounding box.

In the present work, normalized size for each subword is considered to be 64×64 pixels; then, normalized subword is entered to feature extraction stage.

5.2. Feature Extraction. Based on the preprocessed input images, the feature extraction is performed [36]. The obtained subword image from normalization stage is decomposed using wavelet packet transform (WPT) with basis function Symlet 8 up to level two. We use subband [2 0] of level two in 1×729 vector as feature vector. The feature vector is compressed by PCA, reducing the number of features from 729 to 100.

5.3. Computing the Average of 12 Feature Vectors Obtained from Subband [2 0]. The average computing stage of proposed algorithm is shown in Figure 7. In this method, the

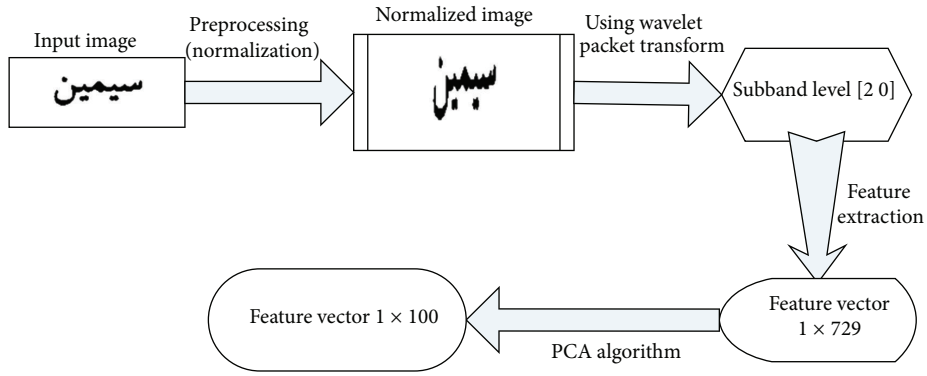

Figure 6: Feature extraction algorithm flow chart.

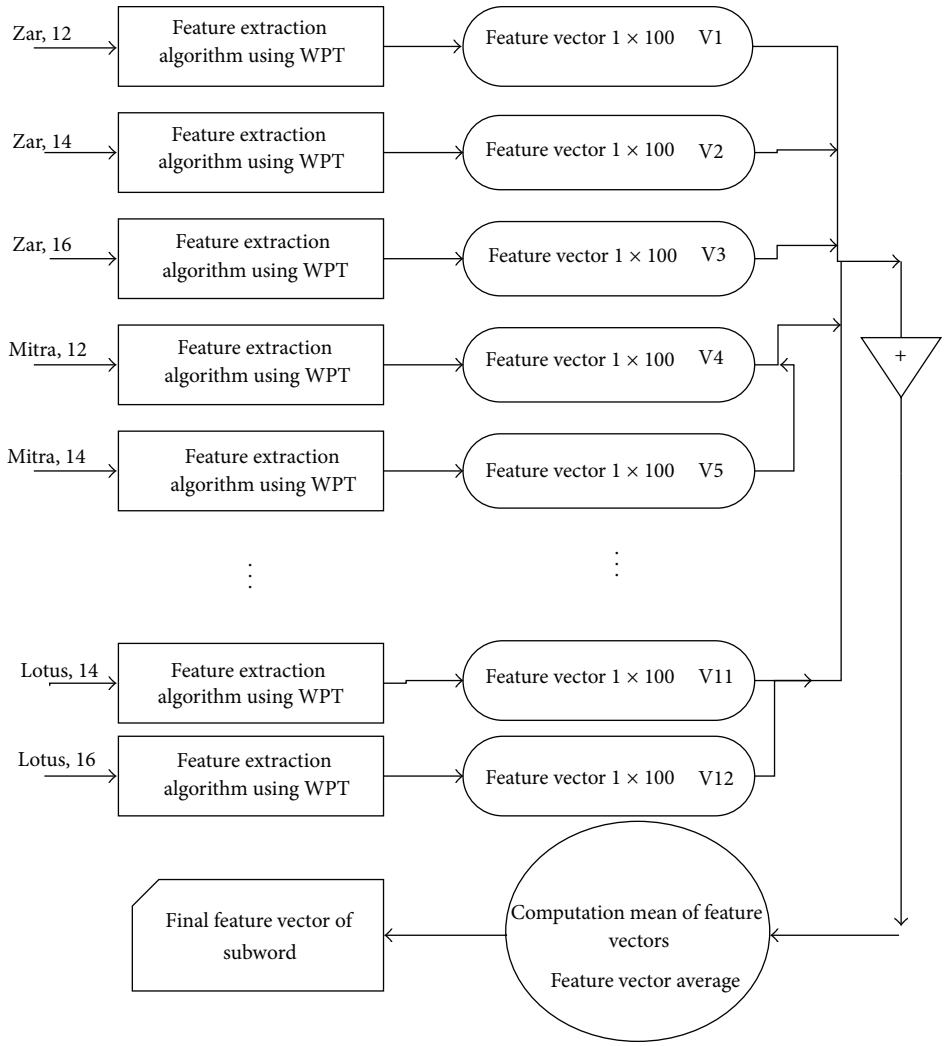

Figure 7: The average computing stage of proposed algorithm.

outputs of feature extraction stage for 12 different shapes of subwords are entered to the averaging stage. Their average is computed in a 1×100 vector as the subword feature that is invariant in font and size. The pictorial dictionary is built using the feature vectors of all subwords.

The dataset consists of 12 groups with the four fonts, including Mitra, Zar, Nazanin, and Lotus, in three sizes 12, 14, and 16 of Persian subwords. Each group has 7317 subwords, with special fonts and sizes. Dictionary is built once and does not need any iteration or learning algorithm.

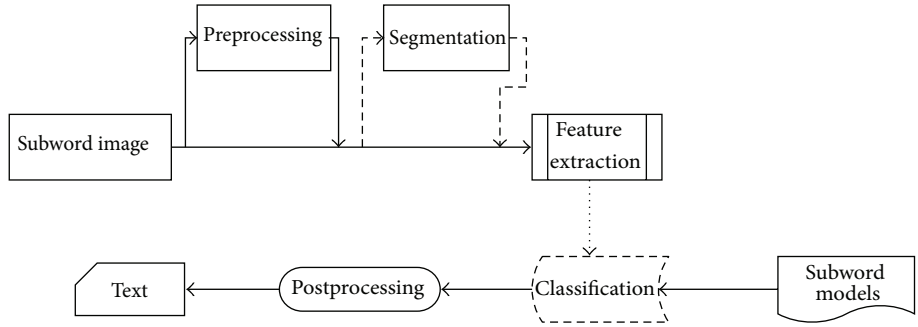

Figure 8: Typical components of OCR system.

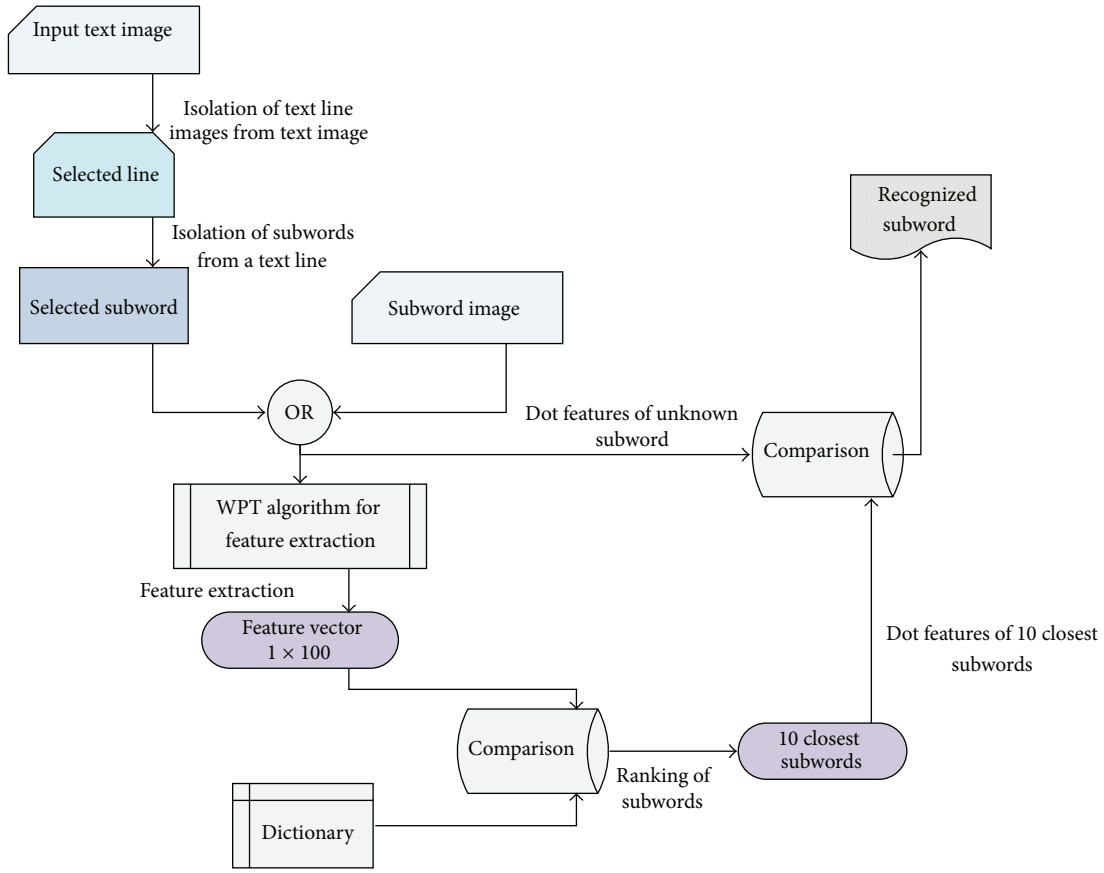

Figure 9: Block diagram of the proposed algorithm.

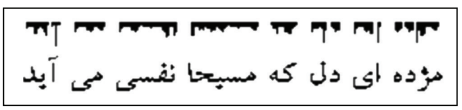

Figure 10: Vertical histogram [37].

Our dictionary contains only 7317 entries that represent 7317 Persian subwords. These entries are decomposition vectors driven from subwords (12 groups with different sizes and fonts × 7317 subwords of the available database).

The dictionary is built as follows. First, apply the decomposition algorithm to every single image in the available database. Then, for any subword, take the average of each group consisting of the same subword with special fonts and sizes. By now there are 7317 decomposition vectors, and one vector for each group consists of 12 different shapes for fixed subword. Therefore dictionary of subword features is generated.

6. The Process of Subwords and Text Recognition

OCR is the process of converting a raster image representation of a document into a format that a computer can process [36, 37]. Printed Persian subwords recognition, in its most

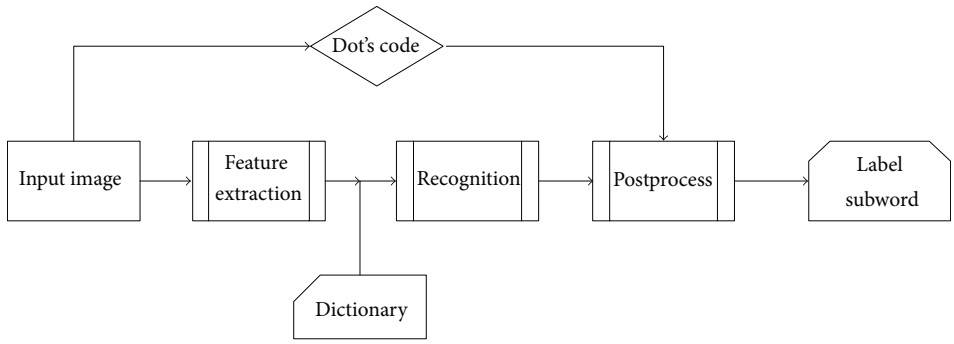

Figure 11: Recognition subsystem is combined with dot recognition modules and postprocessing blocks to recognize subwords.

general form, is the transcription in printed document and machine written text [38]. The process starts at the document level and goes through the following steps: (i) extraction of lines from the document, (ii) extraction of subwords from the line, (iii) holistic recognition of subwords, (iv) a post-processing step for enhancing the recognition process.

The process of recognizing Persian text can be broadly broken down into 5 stages after image acquisition [4, 36]: preprocessing, segmentation, feature extraction, classification, and postprocessing as shown in Figure 8.

The main concept of the proposed algorithm is based on the property that the wavelet packet compressed image is a decomposition vector which can uniquely represent the input image to be correctly reconstructed later at a decompression stage. This property can be effectively used to recognize the subword's image. The subband [2 0] of level two of the wavelet packet transform has been selected as subword feature because it supports enough approximations about the subword image information required to generate a unique vector and minimize errors on detection(recognition) stage. Subband [2 0] coefficients formed a 27×27 matrix from which a 1×729 vector is formed. Then the obtained vector is compressed by PCA, reducing the number of features to 100 to produce the subword feature vector. Figure 9 shows the recognition scheme flow chart.

6.1. Preprocessing for Recognition Process. In preprocessing stage, we have divided lines and then set apart subwords [37]. There is a free space between each two words as well, and we used this feature to segment words. We use vertical histogram of lines to separate words (as depicted in Figure 10). Connected component labeling is used to separate subwords.

The subword is normalized to 64×64 pixels and entered to feature extraction stage.

6.2. Feature Extraction. Block diagram of proposed feature extraction system for a subword is presented in Figure 6. Features are extracted from the shape of the subword. To create the feature vector, the subword image is normalized to 64×64 pixels. After the subword's image is scanned in the system, the proposed algorithm will produce a decomposition vector. This vector is uniquely representing input image.

6.3. Recognition. Then Euclidian distance as correlation between feature vector of unknown subword and each vector in dictionary is computed to measure similarity of two vectors. The ranking of similar subwords is produced. Then 10 closest subword vectors with maximum correlation value are chosen as entries to the next stage.

6.4. Postprocessing. Post-processing stage in the recognition system is shown in Figure 11. In this stage, location and number of dots in subwords are extracted as second features. These features are useful in distinguishing letters and subwords that are only different in the number and the location of their dots with respect to the main body of subword [39]. Therefore, unknown subword dots are recognized using a back propagation neural network. First, we obtain the letter's skeleton and then identify the number of isolated parts of the letters by utilizing connected component analysis and location of dots [39]. After identifying all dots in subword, a code is assigned to it. Then, in the recognition system, this code is compared with 10 recognized closest subwords from the previous stage and subword with the same code is obtained as correct recognized subword.

In the experiment to evaluate the represented method, a set of 2000 subwords is used. For multifont and different-size printed Persian subwords, the average recognition rate of 97.9% was measured for this algorithm.

7. Conclusion

In this paper, a feature extraction system for complete set of printed Persian subwords employing wavelet packet descriptors is presented. We used Symlet 8 basis function for wavelet packet transform to obtain the subword's image features. We apply the wavelet packet of level two of subband [2 0] since higher level of subbands did not give better results, yet caused more complexity. This algorithm was used to built the pictorial dictionary for printed Persian subwords with four fonts and three sizes. The obtained dictionary is used to recognize the subwords in different text documents. The

capabilities of this wavelet operator in detecting patterns with particular properties in the image are used appropriately to accomplish different essential tasks in a pattern recognition process such as holistic Persian subword recognition. For multifont and different-size printed Persian subwords, the average recognition rate of 97.9% was measured for this algorithm.

Acknowledgment

This work is supported by the East Azarbaijan Telecommunication company, Tabriz, Iran.

References

[1] B. Albadr and R. M. Haralick, "A segmentation-free approach to text recognition with application to arabic text," *International Journal on Document Analysis and Recognition*, vol. 1, no. 3, pp. 147–166, 1998.

[2] A. Amin, "Off-line arabic character recognition: the state of the art," *Pattern Recognition*, vol. 31, no. 5, pp. 517–530, 1998.

[3] T. S. El-Sheikh and R. M. Guindi, "Computer recognition of arabic cursive scripts," *Pattern Recognition*, vol. 21, no. 4, pp. 293–302, 1988.

[4] A. A. Aburas and S. M. A. Reheil, "Off-line omni-style handwriting arabic character recognition system based on wavelet compression," *Ariser*, vol. 3, no. 4, pp. 123–135, 2007.

[5] H. Khosravi and E. Kabir, "Farsi font recognition based on sobel-roberts features," *Pattern Recognition Letters*, vol. 31, no. 1, pp. 75–82, 2010.

[6] H. Pirsiavash, R. Mehran, and F. Razzazi, "A robust free size OCR for omni-font persian/arabic printed document using combined MLP/SVM," in *Proceedings of the 18th Iberoamerican Congress on Pattern Recognition (CIARP '05)*, vol. 3773, pp. 601–610, Havana, Cuba, 2005.

[7] G. Y. Chen, T. D. Bui, and A. Krzyzak, "Image denoising with neighbour dependency and customized wavelet and threshold," *Pattern Recognition*, vol. 38, no. 1, pp. 115–124, 2005.

[8] M. S. M. El-Mahallawy, *A Large scale HMM-based omni front-written OCR system for cursive scripts [Ph.D. thesis]*, Department of Computer Engineering, Faculty of Engineering, Cairo University, Cairo, Egypt, April 2008.

[9] M. S. M. Khorsheed, *Automatic recognition of words in Arabic manuscripts [Ph.D. thesis]*, Churchill College, Univesity of Cambridge, Cambridge, UK, 2000.

[10] A. Ebrahimi and E. Kabir, "Printed presiansub-word images clustering using loci characteristic features and *k*-means algorithm," *Journal of Faculty of Engineering*, vol. 33, no. 1, pp. 1–12, 2006.

[11] R. Azmi, E. Kabir, and K. Badie, "An algorithm for clustering and recognition of omnifont farsi sub-words," *International Journal of Engineering Science*, vol. 12, no. 1, pp. 39–49, 2001.

[12] A. Azmi and E. Kabir, "Design of three pictorial dictionaries for recognition of printed Farsi sub-words recognition," *Amirkabir Journal of Science and Technology*, vol. 15, no. 59, pp. 29–43, 2004.

[13] A. Ebrahimi and E. Kabir, "A two step method for the recognition of printed sub-words," *Iranian Journal of Electrical and Computer Engineering*, vol. 2, no. 2, pp. 57–62, 2005.

[14] A. Ebrahimi and E. Kabir, "A pictorial dictionary for printed farsi subwords," *Pattern Recognition Letters*, vol. 29, no. 5, pp. 656–663, 2008.

[15] M. S. Jelodar, M. J. Fadaeieslam, N. Mozayani, and M. Fazeli, "A Persian OCR system using morphological operators," in *Proceedings of the 2nd World Enformatika Conference (WEC '05)*, pp. 137–140, Istanbul, Turkey, February 2005.

[16] N. Ben Amor and N. E. Ben Amara, "Multifontarabic characters recognition using hough transform and HMM/ANN classification," *Journal of Multimedia*, vol. 1, no. 2, pp. 50–54, 2006.

[17] M. B. Menhaj and M. Adab, "Simultaneous segmentation and recognition of farsi/latin printed texts with MLP," in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN '02)*, pp. 1534–1539, May 2002.

[18] Z. A. Aghbari and S. Brook, "HAH manuscripts: a holistic paradigm for classifying and retrieving historical Arabic handwritten documents," *Expert Systems with Applications*, vol. 36, no. 8, pp. 10942–10951, 2009.

[19] E. J. Erlandson, J. M. Trenkle, and R. C. Vogt, "Word-level recognition of multifont arabic text using a feature vector matching approach," in *Document Recognition III*, vol. 2660 of *Proceedings of SPIE*, pp. 63–70, March 1996.

[20] I. A. Jannoud, "Automatic Arabic hand written text recognition system," *American Journal of Applied Sciences*, vol. 4, no. 11, pp. 859–866, 2007.

[21] J. H. AlKhateeb, J. Jiang, J. Ren, F. Khelifi, and S. S. Ipson, "Multiclass classification of unconstrained handwritten Arabic words using machine learning approaches," *The Open Signal Processing Journal*, vol. 2, pp. 21–28, 2009.

[22] I. Daubechies, *Ten Lectures on Wavelets*, Prentice Hall, New Jersey, NJ, USA, 1998.

[23] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[24] M. Omidyeganeh, R. Azmi, K. Nayebi, and A. Javadtalab, "A new method to improve multi font Farsi/arabic character segmentation results: using extra classes of some character combinations," in *Proceedings of the 13th International Conference on Multimedia Modeling (MMM '07)*, T.-J. Cham, Ed., vol. 4351 of *Lecture Notes in Computer Science*, part 1, pp. 670–679, Springer, 2007.

[25] M. Omidyeganeh, K. Nayeb, R. Azmi, and A. Javadtalab, "A new segmentation technique for multi font farsi/arabic texts," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 2, pp. 757–760, March 2005.

[26] M. Sarfraz, S. Nawaz, and A. Al-Khuraidly, "Offline arabic text recognition system," in *Proceedings of the International Conference on Geometric Modeling and Graphics (GMAG '03)*, pp. 30–35, 2003.

[27] R. Azmi and E. Kabir, "A new segmentation technique for ominifont farsi text," *Pattern Recognition Letters*, vol. 22, no. 2, pp. 97–104, 2001.

[28] S. Izadi, J. Sadri, F. Solimanpour, and C. Y. Suen, "A review on persian script and recognition techniques," in *Proceedings of the Conference on Arabic and Chinese Handwriting Recognition (SACH '06)*, D. S. Doermann and S. Jaeger, Eds., vol. 4768 of *Lecture Notes in Computer Science*, pp. 22–35, Springer, 2008.

[29] R. Coifman, Y. Meyer, and V. Wickerhauser, "Wavelet analysis and signal processing," in *Wavelets and Their Applications*, M.

B. Ruskai, Ed., pp. 453–470, Jones and Bartlett, Boston, Mass, USA, 1992.

[30] G. Raju and K. Revathy, "Wavepackets in the recognition of isolated handwritten characters," in *Proceedings of the World Congress on Engineering (WCE '07)*, vol. 1, pp. 635–638, London, UK, 2007.

[31] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[32] K. Huang and S. Aviyente, "Information-theoretic wavelet packet subband selection for texture classification," *Signal Processing*, vol. 86, no. 7, pp. 1410–1420, 2006.

[33] M. Feil and A. Uhl, "2-D Wavelet packet decomposition on multicomputers," in *Proceedings of the 8th Euromicro Workshop on Parallel and Distributed Processing*, pp. 351–356, Rhodos, Greece, January, 2000.

[34] G. K. Kharate, A. A. Ghatol, and P.P. Rege, "Image compression using wavelet packet tree," *ICGST International Journal on Graphics*, vol. 5, no. 7, pp. 37–40, 2005.

[35] S. Mallat, *A Wavelet Tour of Signals Processing*, Academic Press, New York, NY, USA, 1999.

[36] M. S. Khorsheed, "Off-line Arabic character recognition—a review," *Pattern Analysis and Applications*, vol. 5, no. 1, pp. 31–45, 2002.

[37] P. Wunsch and A. F. Laine, "Wavelet descriptors for multiresolution recognition of handprinted characters," *Pattern Recognition*, vol. 28, no. 8, pp. 1237–1249, 1995.

[38] M. SalmaniJelodar, M.J. Fadaeieslam, N. Mozayani, and M. Fazeli, "A persian OCR system using morphological operators," in *Proceedings of the 2nd World Enformatika Conference (WEC '05)*, vol. 4, pp. 137–140, Istanbul, Turkey, February 2005.

[39] J. Shanbehzadeh, H. Pezashki, and A. Sarrafzadeh, "Features extraction from persian hand written letters," in *Proceedings of the Image and Vision Computing New Zealand (IVCNZ '07)*, pp. 35–40, Hamilton, New Zealand, 2007.