Recognition of Farsi Handwritten Digits Using a Small Feature Set

G. Mirsharif, M. Badami, B. Salehi, and Z. Azimifar

Abstract—Recognition of Farsi/Persian handwritten numeral characters has been the focus of study recently and has many applications such as postal code reading and check processing. One important step in any recognition system is feature extraction. We propose to use a small set including only 20 domain specific features which are simple to understand, easy to implement and extracted in a way similar to how humans discriminate digits. These features are extracted by simply counting the pixels which are confined in different curves of digits. Unlike the universal methods this way of feature extraction is related to the problem. Evaluating the proposed features indicates an achievement of 97% recognition rate on Hoda dataset. This method is scale and shift invariant and no pre-processing is needed.

Index Terms—Persian handwritten digits, feature extraction, domain specific features.

I. INTRODUCTION

Recently, recognition of Persian digits has been the focus of study in Iran. A very important step in any recognition system is feature extraction. Feature extraction is a special form of dimensionality reduction and can decrease time complexity and cost in processing large images. In the past two decades, a number of methods have been proposed for machine recognition of Persian handwritten characters.

Classification of digits is presented in earlier studies using moment features and the Bayesian classifiers. Shahreza et al. [1] Used shadow coding method for recognition of Persian handwritten digits. In this method, a 32-segment mask is overlaid on the digit image and the features are calculated by projecting the image pixels into the segments.

In a study by Said et al. [2] first, the size of each digit image is normalized into 16×20 . Then pixel values of the normalized image is fed into a neural network for classification.

In another paper, presented by Sadri et al. [3] the first and last profiles of the image at both vertical and horizontal orientations are obtained at feature extraction stage. Then, each of these profiles is represented as a one-dimensional signal. The derivative of each of these signals is represented by a vector of size 16, where these vectors constitute the feature vector.

Soltanzadeh and Rahmati [4] presented a novel method for recognition of Persian handwritten digits. In their method the authors used the image profile calculated at multiple orientations as the main features.

In this paper, we propose a new method for Persian handwritten recognition on HODA [5] data set which contains digits in very different sizes and handwritings. Ebrahimpour et al. [6] proposed a new method on this data set. Their technique is based on characterization loci and mixture of experts which utilizes the characterization loci, as the main feature. The recognition rate by this method is 97.52%. The approach taken by Moradi et al. [7] is another technique on HODA dataset which combines two feature extraction methods. These methods are executed in parallel on an FPGA chip and result in 96% accuracy. Their Experiments result of function simulation PC-based system is 97%.

The major problem with universal features is they are usually unrelated to the discriminative features used by humans to recognize the different classes in a particular problem. To address this issue, this paper demonstrates that domain-specific features are superior to universal features in classification accuracy. In addition, our proposed features are easy to understand and simply extracted. We can also implement this method on FPGA due to its simplicity. In addition the number of features is small and they are shift and scale invariant.

Abdelazeem [8] proposed a new method for Arabic dataset similar to the way human recognizes digits. Results show that a carefully chosen feature vector of only 35 features could outperform many universal feature sets of hundreds of features in both recognition accuracy and speed. Arabic and Farsi digits are different in digit 2, 3, 4 and 6. Persian digits like 2 and 3 or 4 and 6 can be confused using this method when they are written fast. Thus, some superior features are needed to improve the recognition rate.

In this paper, we improve some of Abdelazeem's features and remove those features, which are irrelevant for Persian digits and define new features for more accurate classification of Persian digits. Finally 20 features are achieved and are tested on HODA data set that leads to 97% accuracy. It is noticeable that we can achieve this great rate without any pre-processing on primary images.

The rest of this paper is organized as follows: Section II introduces the proposed method including details of data set, feature extraction and classification method. Section III presents our experimental results for HODA dataset and finally the Conclusion appears in section IV.

II. DATA COLLECTION, FEATURE EXTRACTION AND CLASSIFICATION METHOD

A. Data Collection

We used HODA dataset introduced by Khosravi and Kabir in 2007 [5]. They used 12000 registration forms of two types for this dataset. It includes 60000 train and 20000

Manuscript received June 12, 2012; revised July 6, 2012.

The authors are with School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran (e-mail: Mirsharif@cse.shirazu.ac.ir, Mbadami@cse.shirazu.ac.ir, BaharSalehi@cse.shirazu.ac.ir, Azimifar@cse.shirazu.ac.ir)

test digits and the dataset is very big that digits are from various handwritings and sizes. Some of the digits are broken and some are even hard to recognize by human. (See Fig. 1)

Fig. 1. Some examples from HODA dataset: (a) Digit 2 in different handwriting formats. (b) Broken digits. (c) Examples of digits which are hard to recognize.

Fig. 2. (a) Being broken does not affect the recognition rate of digit because the area is still small. (b) The ratio of height of bounding box to width of digit 1. (c) The ratio of height of bounding box to width of bounding box which is same for 1 and 2.

Fig. 3. (a) The extra part in digit 3 is indicated by red colour. (b) Pixels which are confined in bends of digit 4. (c) Pixels which are confined from top and right in digit 6 are shown by red colour. (d) Pixels which are confined from top and left in digit 6.

B. Feature Extraction

We explain how to extract each of 20 specific features for every digit. In HODA dataset the extra pixels around each digit are diminished so the digit is confined in a bounding box. This makes features shift invariant. The first 12 features are extracted as was used in Abdelazeem's method. We made change to some of those features to make them appropriate for Persian digits.

For digit 0, we used the area of bounding box because this digit is small comparing to other digits. This feature can discriminate 0 up to 97%. Because of using size feature for digit 0 even if it is broken, this feature can still recognize 0 to a great extent. (See Fig. 2.a)

As it can be seen in Fig. 2.b, for digit 1 we used the ratio of height of bounding box to width of the digit. Width of digit can be estimated by finding the maximum of distances between first and last digit pixels in each row. As shown in Fig. 2.c, if we use width of bounding box instead of width of digit as used in Abdelazeem's method digit 1 and 2 can be confused especially when digit 1 is written with an angle or skewed. For recognizing digit 2 we used the same main feature as was used in Abdelazeem's method for this digit. (Refer to [8])

A great part of error when Arabic features are used to recognize Farsi digits is related to digit 3 because digit 2 in Arabic is written like digit 3 in Persian. To differentiate digit 2 and 3 we count the pixels of extra part in Persian digit 3 as shown in Fig. 3.a.

Persian digit 4 can be written in two formats as seen in Fig. 3.b. Pixels confined in the bend of left profile of digit 4 from top, right and below and pixels confined in top right quarter of image from top, left and below are used to recognize digit 4. We use the same main feature as was used by Abdelazeem for recognition of digit 5. (Refer to [8])

Fig. 3.c illustrates two formats of writing Digit 6 in Persian. They can be recognized by counting pixels which are confined from top and right by digit pixels. Number of pixels confined from top and left also can help recognizing digit 6 in format shown in Fig. 3.d.

To recognize digit 9 from others, especially from one format of digit 6 which is written without bend, we count the pixels confined from 4 main directions in top left quarter of image as shown in Fig. 4.

Fig. 4. Number of red pixels is a good feature for discrimination of digit 9 from 6

In addition to these 6 features, we use 6 extra features which are extracted mainly for digits 2,5,6,7 and 8. These features are named white_2, white_5, white_6, white_7, white_8 and white_depth_7 in Abdelazeem's method. The detail explanation of these extra 6 features can be referring to [8].

Using these 12 main features the accuracy is 94.80%. We also extracted other features in the same way. We add each feature to previous features. If the new feature increases the accuracy it is appended to the main features else it is

discarded. We continue the process until the recognition rate reaches 97%. The 8 extra features which are added in this process are explained in next section.

Fig. 5. (a) Digits 2, 3 and 4 showing the direction to the right. (b) Digits 6 and 9 showing the direction to the left. (c) Digits 2 and 3 with elongated part. (d) The bottom part of digit 4 is shown with red colour.

Fig. 6. (a) Red pixels show the feature white_3 which can lead to confusing digits 2 and 7. To discriminate digit 7 we use right steep as shown by the red arrow. (b) Left steep of digit 8 is shown by red arrow. (c) Red arrows show the curvy shape of digit 3. (d) Red pixels show the bend of digit 4.

Fig. 7. Digit 9 contains more red pixels than digit 6.

Digits like 2, 3, and 4 contain digit 1 plus an extra part which is elongated to the right and digit 6 and 9 have this extra part elongated to the left. Direction of this extra part can be used to distinguish these groups. To obtain the direction, first measure the distance between first and last digit pixels in each column then subtract the index of column with maximum distance from the index of column with minimum distance which is a positive number for first group and negative for second group. The sign of this number shows the direction as shown in Fig. 5.a and 5.b.

Digit 3 is usually elongated to right more than digit 2. The length of this elongated part could help discriminating these two digits. (See Fig. 5.c)

Digits 4 and 6 in Persian are very similar except for the bottom part in digit 4. Counting the pixels of digit in 5 last rows can be useful. (See Fig. 5.d)

When digit 2 is written with a big font size it can be confused with digit 7 using feature named white_3 in

Abdelazeem's method as shown in Figure 6.a by red colour. Right steep of digit 7 is a good feature to discriminate digit 2 from digit 7. To obtain right steep of digit 7, count the number of times column index of last digit pixel in each row is smaller than column index of last digit pixel in previous row.

Left steep is also a useful feature to discriminate digits 5 and 8 from others. (See Figure 6.b)

As illustrated in Figure 6.c, some people write 3 in a curvy shape. Indicating this curve by value one if row index of first digit pixels in each column decrease and then increase with a big steep. This can help discriminating this digit from digit 2.

Counting the digit pixels in top right quarter of image can be used to recognize digit 4 from 2, 3 and 6. (See Figure 6.d)

For the last feature, the digit pixels in top left quarter are counted to distinguish digit 9 from 6 as seen in Figure 7.

C. Classification Method

To classify the digits into 10 classes, we examined different classifiers like neural network (NN) and support vector machine (SVM). We could achieve the best result with one versus one SVM on the 10000 train and 10000 test samples which were chosen randomly. We used a validation set to choose kernel type and parameters of the classifier.

III. EXPERIMENTAL RESULT

In this section, we evaluate the performance of our 20 proposed features for HODA dataset. Some of these features as explained in Section II part B are based on features used in [8]. Abdelazeem introduced 35 domain specific features for recognition of Arabic handwritten digits with the reported recognition rate of 99%. The accuracy for classification of Persian handwritten digits using these 35 features was 85%.

This reduction in accuracy is due to different ways of writing digits 2, 3, 4 and 6 in Persian. In addition, HODA dataset contains many digits with large variety of handwritings. We select 12 best features in [8] and improve some of these 12 features to make them appropriate for recognition of Persian digits. Using these main 12 features, the accuracy was increased to 94.80 %. We extract other features and choose 8 of them using feature selection. These 8 more features are extracted with a focus on digits 2, 3, 4 and 6 which are written differently in Persian. Using the total 20 features we achieved 97% accuracy on HODA dataset. Recognition rate for each group of features are reported in Table I.

Finally to evaluate the recognition rate in each class, we calculated the confusion matrix which is shown in TABLE II. As seen in this table, digits 1 and 3 have the highest and lowest accuracy, respectively. The only feature extracted particularly to recognize digit 1 is the ratio of height to width of the digit. Other features return value zero or close to zero for this digit, thus digit 1 can be recognized easily from others. Digit 3 is very similar to digit 2 and if written fast and carelessly, it can be confused with digits 4 and 2 using our proposed features.

A few numbers of methods have been proposed on HODA dataset until now. One of them is [6] which use characterization loci as features. Replacing multilayer perceptrons (MLP) by mixture of MLP experts as classifier improves the accuracy up to 2.5% in this method. The final reported recognition rate is 97.51%. Another method is [7] which combines two feature extraction approaches. These approaches are executed in parallel on a FPGA chip and lead to 96% accuracy. Their Experimental result of function simulation PC-based system is 97%. Comparing to these methods, our proposed approach uses a small feature set which can be obtained by simply counting the pixels. In addition, using a simple SVM as classifier and with no pre-processing we achieve 97% recognition rate on HODA dataset.

IV. CONCLUSION AND FUTURE WORK

We proposed 20 domain specific features which are easy to extract and can decrease complexity of recognition system. Using these 20 features which are extracted in a way similar to how human can discriminate digits, without any pre-processing step, we achieved 97% recognition rate on HODA dataset with 99% recognition rate on some of digits. Recognition rate can be improved by extracting better features or adding other features to this feature set.

In addition, we can exploit hybrid methods in which we train the classifier using these features to divide the digits into similar groups and then classify each group using a different feature set [9]. There are also many broken digits in HODA data set for which using morphological operators

to bridge the gaps can help improving the classification accuracy [10].

REFERENCES

- [1] M. H. S. Shahreza, K. Faezand, and A. Khotanzad, "Recognition of handwritten Persian/Arabic numerals by shadow coding and an edited probabilistic neural network," in *Proceeding International Conference on Image Processing*, 3, pp. 436–439, 1995.
- [2] F. N. Said, R. A. Yacoub, and C. Y. Suen, "Recognition of English and Arabic numerals using a dynamic number of hidden neurons," in *Proceeding International Conference on Document Analysis and Recognition*, pp. 237–240, 1999.
- [3] J. Sadri, S. Izadi, F. Solimanpour, Y. C. Suen, and T. D. Bui, "State of the art in Farsi script recognition," *International Conference on Information Science and Signal Processing and its Application*, 2007.
- [4] H. Soltanzadeh and M. Rahmati, "Recognition of Persian handwritten digits using image profiles of multiple orientations," *Pattern Recognition Letters*, 2004.
- [5] R. Khosravi and E. Kabir, "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties," *Pattern Recognition Letters*, pp. 1133–1141, 2007.
- [6] R. Ebrahimpour, M. R. Moradian, A. Esmkhani, and F. Jafarlou, "Recognition of Persian handwritten digits using characterization loci and mixture of experts," *International Journal of Digital Content Technology and its Applications*, vol. 3, 2009.
- [7] M. Moradi, M. A. Pourmina, and F. Razzazi, "A new method of FPGA Implementation of Farsi handwritten digit recognition," *European Journal of Scientific Research*, vol. 3, pp. 309–315, 2010.
- [8] S. Abdelazeem, "A novel domain-specific feature extraction scheme for Arabic handwritten digits recognition," *International Conference on Machine Learning Applications*, 2009.
- [9] Z. Ping and C. Lihui, "A novel feature extraction and hybrid tree classification for handwritten numeral recognition," *Pattern Recognition Letter*, 2002.
- [10] D. Yu and H. Yanon, "Reconstruction of broken handwritten digits based structural morphological features," *Pattern Recognition Letters*, pp. 235–254, 2001.