

A Decision Tree Based Method to Classify Persian Handwritten Numerals by Extracting Some Simple Geometrical Features

Hamidreza Alvari, Seyed Mehdi Hazrati Fard, and Bahar Salehi

Abstract—Automatic recognition of handwritten numerals has been widely proposed in various languages. However, some languages such as Persian still need more consideration. In this paper, we proposed a handwritten Persian numerals dataset, which is gathered from people with different range of educational level. Thus, it is more general than other similar Persian datasets. Additionally, a method to classify Persian handwritten numerals is presented, which uses simple geometrical features based on their shapes, and classifies them via a rule-based decision tree classifier. Compared to other similar methods, our proposed method has the advantages of high speed running, employing too few and simple features, and elimination of training phase.

Index Terms— Decision tree, feature extraction, geometrical shape, preprocess.

I. INTRODUCTION

Recent developments in electronic publishing has brought up the need for transformation of paper-based and printed documents to digital formats, hence, text recognition has been highlighted further [1].

In the past three decades, numerous methods have been proposed for machine recognition of handwritten characters, especially for languages such as English, Japanese and Chinese. In contrast, very few studies are reported for the recognition of Persian (Arabic) handwritten digits, and the results are not satisfactory [2]. For instance, in [4]-[7], a number of methods with high recognition rates in English handwritten digits had been proposed. Some recently proposed methods for recognition of Persian numerals are as follows:

In [8], first, each numeral shape is normalized to the size of 35×35 pixels. Then, the gradient vector of the image is calculated in 8 directions, and 5×5 feature values are extracted from each direction. Finally a 200 dimensional feature vector is fed to 6 classifiers to classify this numeral. In [9], shadow coding is used, where features are calculated by projecting pixels on 32 segments mask. In [10], Persian numerals are displayed in the form of 11 segments and then values of features for each segment are calculated and after combination, they are fed into a classifier. In [11], first, each numeral shape is normalized into the size of 16×20 pixels. Then the values of its pixels are fed into a neural network with a dynamic number of hidden neurons.

From all of these methods, the method used in [8] has the

highest recognition rate at the expense of the more complexity in time and also introduced features. It is also important to note that these methods used different datasets.

Generally the performance of a character recognition system greatly, depends on the feature extraction approach and the classification/learning scheme [4]. Thus, different approaches for feature extraction from characters are proposed [12]. For instance, stroke direction feature [13]-[14] and curvature feature [15] are two of these approaches.

In this paper, we proposed some features which are very simple compared to other methods. High speed running and not requiring to the training phase are other advantages of our method.

The rest of paper is organized as follows. First we briefly introduce our approach in section II and then introduce our features. Later, we will introduce our dataset and discuss the performance of our proposed method in section III. Finally, we will end with a conclusions and future work section.

II. PROCESS OF NUMERAL RECOGNITION

The recognition of Persian handwritten numerals is done in a number of steps as shown in Fig. 1. The Preprocess step consists of smoothing and noise reduction, background elimination, edge extraction and normalization steps. In smoothing and noise reduction step, a 5×5 Gaussian low pass filter with standard deviation of 1.4 is applied on the input image. As a result, the image becomes smoother and the noise is reduced. Then, in background elimination step, we eliminate the background of the image according to Otsu method [3].

After that, in edge extraction step, we extract edges using Sobel method and finally in normalization step, we reduce the image size into the size of its containing character. The preprocess step applied on '4' is depicted in Fig. 2.

It is likely that the strong and chromatic noises are not eliminated during smoothing and noise reduction step, therefore we applied another step, named as unification step. In this step, we find rows and columns which have no pixels and eliminate the smaller and wasted part of numeral which is appeared after normalization step. This is done based on image histogram, in the horizontal and vertical directions. The effect of this step on '0' is shown in Fig. 3.

After preprocessing step, we start to extract features. Generally, in this paper, the extraction of features is based on geometrical shapes of numerals and consists of numeral size, number of intersection points of the numeral with hypothetical lines, and distances between these lines, number of jags, direction of tail slope, curvature direction and

continuity status in the upper and lower halves.

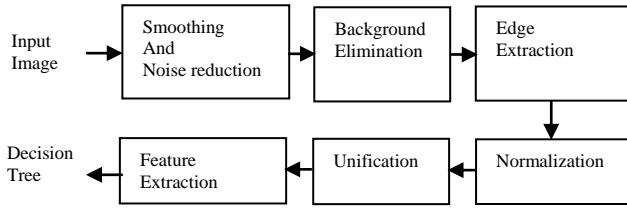


Fig. 1. Total process of numeral recognition



Fig. 2. Preprocess of '4' from left to right



Fig. 3. Preprocess of '0' by considering unification step from left to right

Based on these features, a rule-based decision tree is constructed. In order to classify each new test input numeral, the test sample will traverse the tree from its root to one of its leaves. The general scheme of our proposed decision tree is shown in Fig. 4. The correspondence between Persian and English numerals is shown in Table I:

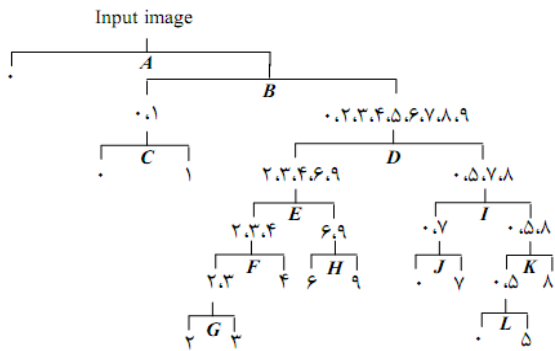


Fig. 4. Decision tree

TABLE I: THE CORRESPONDENCE BETWEEN PERSIAN AND ENGLISH NUMERALS

	0	1	2	3	4	5	6	7	8	9
Persian	0	1	2	3	4	5	6	7	8	9
English	0	1	2	3	4	5	6	7	8	9

In the following subsections, each extracted feature from leftmost subtree to rightmost subtree will be explained. Before, we define two terms used in some features:

Peak: A point which is higher than its left and right neighbors.

Valley: A point which is lower than its left and right neighbors.

A. Classification of '0' from other Numerals

Most samples of '0' have smaller sizes compared to other

numerals and this simple feature can be used to classify them. But if this size is bigger than a predetermined threshold, recognition will be done in two branches other than current branch, as we can see in Fig. 4. In other words, if we have a big '0', in this branch, it will not be classified and must go down further in decision tree in order to be recognized.

B. Classification of '1' from other Numerals

We introduce three simple features to classify '1' from all other numerals here:

Feature 1: At each row of the image, we pass a line and the maximum number of their intersection points with image is considered as a feature. In the samples of '1' there are at most four intersection points, while in other numerals, often there are more. (In the continuity status of pixels, we consider only two intersection points: one as input and another as output)

Feature 2: Even if samples of '1' are very oblique, the width of their normalized boundaries will not exceed from a threshold.

Feature 3: We consider the first point of each column from the top of the image. In the samples of '1', there is exactly one peak, while often in other numerals there are more.

At this point, some '0' numerals are not classified completely and as we go down further in the tree, they will be recognized completely.

C. Classification of '0' from '1'

At each row of the image, we pass a line and the distance between the first and the last intersection points is calculated. The maximum of these distances is considered as the numeral width. This value is less for '1' compared to '0', because naturally, samples of '1' are thinner than samples of '0'.

D. Classification of '2', '3', '4', '6' and '9' from '0', '5', '7' and '8'

In the predetermined fraction (e.g. half) of lower half of the image, we consider one horizontal line. In the second category numerals, the distance between the first and the last intersection points of this line with the numeral will be more than this distance in the first category numerals.

E. Classification of '2', '3' and '4' from '6' and '9'

Feature 1: As it is illustrated in Fig. 5 for three forms of '6' numeral, tails of the first two forms are in '/' direction while the third form's tail and also '2', '3' and '4' tails are in the opposite direction. This is a simple and good feature to recognize some samples of '6'.

Feature 2: First, we find the tail's upmost pixel. Most of the weights of '6' and '9' are in the left side of this point, while in '2', '3' and '4' are in the opposite side.

F. Classification of '4' from '2' and '3'

Feature 1: In Persian, there are two general forms of '4'. The first form is shown in the leftmost image in the last row of Fig. 5 and two samples of the second form are shown in Fig. 6. It is clear that the number of pixels in the lower half of the second form is more than the first form and also '2' and '3' numerals, due to the curvature existence in the second form.

Feature 2: We eliminate the tail and again adjust the boundaries of image. If in '2' or '3', we move from upmost right pixel and top to down, we will always proceed to left side of the image, but in the first form of '4' this is vice versa.

As a result, '2' and '3' will be recognized from the first form of '4'.

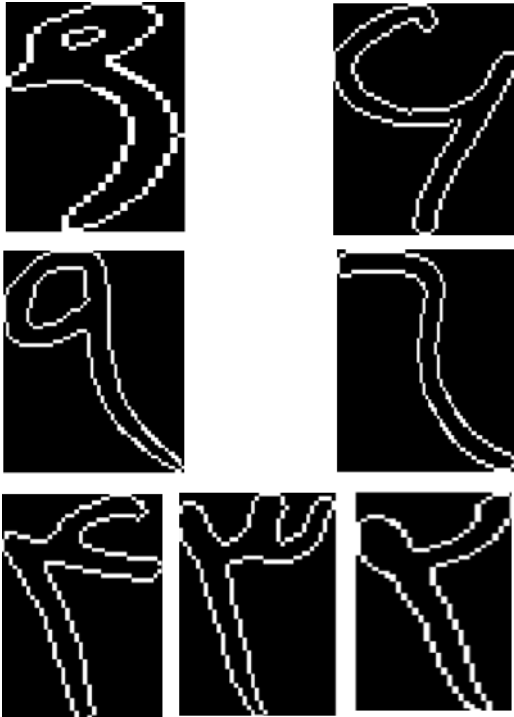


Fig. 5. In the first row, from right to left, the first and second forms of '6' are shown which their tails are in '/' direction. The third form of '6' and also a form of '9' are shown in the second row. In the third row, the first forms of '2', '3' and '4' are depicted from right to left. All numerals tails in the second and third rows, are in '\ ' direction.

G. Classification of '2' from '3'



Fig. 6. Two samples of the second form of '4'



Fig. 7. The second forms of '2' and '3' from left to right

Here, we first eliminate the tail of the numeral. Considering ascending and descending directions of upmost pixels in the upper half and from left to right, in the first form of '3' shown in the middle image in last row of Fig. 5, there are two valleys while in the first form of '2' there is one. Note that in the second form of '3' shown in Fig. 7, there is only one valley. However the distance of the leftmost peak from the left boundary in '3' will exceed from a threshold, so this form of '3' will also be recognized from two forms of '2'.

H. Classification of '6' from '9'

Feature 1: As shown in Fig. 5, tails of the first and second forms of '6' are in '/' direction, while '9' and also the third

form of '6' have a tail in opposite direction. Based on this feature, the first and second forms of '6' are recognized from '9' and the third form. However some samples of '9' are also classified as '6' and will be recognized further from the first and second forms of '6' using next feature.

Feature 2: We consider two upmost pixels in the upper half of the image. Due to continuity status in head of '9', the difference between columns number of these points in '9' is less than the first and second forms of '6'.

Feature 3: After eliminating the tail, we calculate maximum vertical distance between head pixels in '9' and the third form of '6'. Since '9' has circular head, this distance will be far larger compared to the third form of '6'.

I. Classification of '7' from '5' and '8'

In the upper and lower halves of the image, we consider two horizontal lines based on some predetermined thresholds. The distance between the first and the last intersection points of these two lines are found. If this distance in the upper half is more than the distance in the lower half, the numeral is recognized as '7', otherwise as '5' and '8'. Note that some samples of '0' are the same as '5', '7' and '8'. Therefore, they exist in two current branches of decision tree.

J. Classification of '0' from '7'

The second feature which was based on continuity status of head of the numeral and was introduced to classify '6' from '9' is also applicable here. Since '7' has discontinuity in its head, '0' is classified from it using this simple feature.

K. Classification of '8' from '0' and '5'



Fig. 8. The first, second and third forms of '5' from right to left

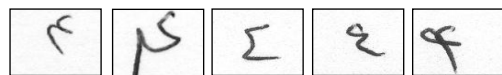


Fig. 9. Various forms of '4' in Persian

From the lowest row of the image and based on some fraction of it (e.g. half) as a threshold, we consider a hypothetical vertical line from bottom to up. The distance from starting point of this line to the first intersection point with numeral, is a good feature in recognizing '8' from '0' and '5'. It is noticeable that, in '8' the first intersection point of this line with numeral is near its peak, while in '0' or '5' is at their bottoms. Therefore in '8', this distance is far larger than in '0' and '5'.

L. Classification of '0' from '5'

Some samples of '0' are not still recognizable from '5'. In Fig. 8, three forms of '5' are illustrated.

Feature 1: To classify the first form of '5', from the second and third forms and also '0', we consider vertical lines from left to right and from the lowest row of the image to the upper rows. The row numbers of the first intersection points of these lines with numeral are found. Similar to what was

done in section G to classify '2' from '3', there is one peak in the first form of '5', while there is no peak in '0' and also in the second and third forms of '5'.

Feature 2 and 3: In the most samples of '0', the width and the height sizes are less than some thresholds while in samples of '5', they are bigger. Based on these features, extremely small samples of '0' and extremely big samples of '5' are recognized.

Feature 4: As shown in Fig. 8. the second and third forms of '5' have uncus in their upmost section. This is not the case in '0'. For example in the second form, there is a valley at the top of the numeral.

Feature 5: In the upper halves of some samples of '5' (e.g. the third form), there is a horizontal valley when we move from the middle of the image to the top of it.

III. RESULTS

In order to study Persian handwritten numerals, we gathered a large dataset of handwritten Persian digits. Gray images of 13,620 digits were extracted from 62 forms which we designed for this research project. The forms were filled by 658 males and 704 females random candidates. Among 1362 total candidates, 91% of them were right-handed and the other 9% were left-handed. Candidates education levels varied from primary level (under diploma) to doctorate and their ages were in the range of 10 to 79. This unique property of our dataset resulted in a very diverse handwritten digits dataset which is shown in Fig. 9 for '4'.

Other previously presented datasets such as [16] are only gathered from educated people (at least diploma) whose handwritings are very different from low educated people. The forms were scanned at 300 dpi, and then segmented with the aid of a predefined mask. No registration or any types of image enhancement were used on raw images. Some samples of our dataset are shown in Fig. 10.

We evaluated this approach on a subset of our dataset and got the accuracy of 99.36% totally. In Fig. 11, the recognition rate corresponding to each numeral is depicted.

Additionally, in order to show the generality of our proposed method, we applied it on another previously presented dataset [17] and we got 98.67% accuracy. It is important to note that our contribution in this paper is first in constructing a good preprocess step and second in introducing too few and simple features.

IV. CONCLUSIONS AND FUTURE WORK

In this study, we proposed a novel Persian numeral recognition method. The method uses some geometrical and simple features, which are employed in a rule-based decision tree and consists of numeral size, number of intersection points of the numeral with hypothetical lines and distances between these lines, number of jags, direction of tail slope, curvature direction and continuity status in the upper and lower halves.

Compared to the similar recognition methods, our method does not need to the training phase and employs too few features, which results in a low execution time.

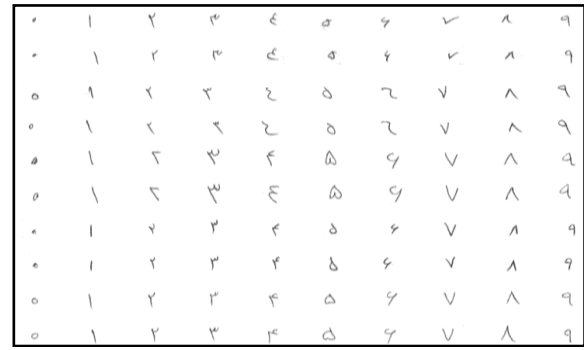


Fig. 10. Ten samples of each Persian numeral from left to right columns

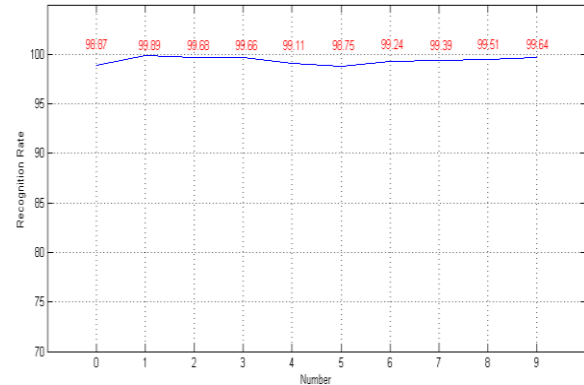


Fig. 11. Recognition rates for each number

ACKNOWLEDGMENT

Authors would like to thank Dr. Zohreh Azimifar for her useful advices on this study. We are also very thankful to our friends, Dr. Jamileh Erfani and Borhan Kazimipour, for their helps and useful comments in gathering data.

REFERENCES

- [1] A. Broumandnia and J. Shanbehzadeh, "Fast Zernike wavelet moments for 590 Farsi character recognition," *Image and Vision Computing*, vol. 25, 2007, pp. 717-726.
- [2] H. Soltanzadeh and M. Rahmati, "Recognition of Persian handwritten digits using image profiles of multiple orientations," *Pattern Recognition Lett*, vol. 25, no. 14, 2004, pp. 1569-1576.
- [3] N. Otsu, "A threshold selection method from gray-level histogram," *IEEE Trans. Syst. Man Cybernet*, vol. 9, 1979, pp. 62-66.
- [4] C. L. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: benchmarking of state-of-the-art techniques," *Pattern Recognition*, 2003, vol. 36, no. 10, pp. 2271-2285.
- [5] M. Shi, Y. Fujisawa, T. Wakabayashi, and F. Kimura, "Handwritten numeral recognition using gradient and curvature of gray scale image," *Pattern Recognition*, vol. 35, 2002, pp. 2051-2059.
- [6] L. N. Teow and K. F. Loe, "Robust vision-based features and classification schemes for off-line handwritten digit recognition," *Pattern Recognition*, vol. 35, 2002, pp. 2355-2364.
- [7] K. Cheung, D. Yeung, and R. T. Chin, "A Bayesian framework for deformable pattern recognition with application to handwritten character recognition," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 29, no. 12, 1998, pp. 1382-1388.
- [8] C. L. Liu and C. Y. Suen, "A new benchmark on the recognition of handwritten Bangla and Farsi numeral characters," *Pattern Recognition*, vol. 42, 2009, pp. 3287-3295.
- [9] M. H. S. Shahreza, K. Faez, and A. Khotanzad, "recognition of handwritten Persian/Arabic numerals by shadow coding and an edited probabilistic neural network," in: *Proc. of the International Conference on Image Processing*, vol. 3, 1995, pp. 436-439.
- [10] H. M. M. Hosseini and A. Bouzerdoum, "A combined method for Persian and Arabic handwritten digit recognition," in *Proc. of the Australian New Zealand Conference on Intelligent Information Systems*, 1996, pp. 80-83.

- [11] F. N. Said, R. A. Yacoub, and C. Y. Suen, "Recognition of English and Arabic numerals using a dynamic number of hidden neurons," in: *Proc. of the Fifth International Conference on Document Analysis and Recognition*, 1999, pp. 237–240.
- [12] O. D. Trier, A. K. Jain, and T. Taxt, "Feature extraction methods for character recognition—a survey," *Pattern Recognition*, vol. 29, no. 4, 1996, pp. 641–662.
- [13] M. Yasuda and H. Fujisawa, "An improvement of correlation method for character recognition," *Trans. IEICE Japan* vol. 62, no. 3, 1979, pp. 217–224.
- [14] Y. Yamashita, K. Higuchi, Y. Yamada, and Y. Haga, "Classification of tailprinted Kanji characters by the structured segment matching method," *Pattern Recognition Lett*, vol. 1, 1983, pp. 475–479.
- [15] M. Shi, Y. Fujisawa, T. Wakabayashi, and F. Kimura, "Handwritten numeral recognition using gradient and curvature of gray scale image," *Pattern Recognition*, vol. 35, no. 10, 2002, pp. 2051–2059.
- [16] H. Khosravi and E. Kabir, "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties," *Pattern Recognit Lett*, vol. 28, no. 10, 2007, pp. 1133–1141.
- [17] Geometrical Shape. Online]. Available: http://www.ocr.majlesi.info/farsi/index.php?option=com_content&view=article&id=4&Itemid=11



Seyed Mehdi Hazrati Fard received his B.Sc. degree from Shiraz Azad University, Shiraz, Iran in 2007 and is currently a last year M.Sc. student at Shiraz University, Shiraz, Iran. He is currently an instructor in the Department of Computer science and IT, Zand and Pishtazan institutes of higher education, Shiraz, Iran. His research interests are in the fields of image processing, machine learning and data mining



Bahar Salehi received her B.Sc. and M.Sc. degrees in computer science and engineering from Shiraz University, Shiraz, Iran, in 2008 and 2011, respectively. She is currently a lecturer in the Department of Computer Science, Payam-e-Noor University, Shiraz branch. Her research interests include image processing, vision, machine learning, genetic algorithms and natural language processing.



Hamidreza Alvani received his B.Sc. degree from Shahid Chamran University of Ahvaz, Ahvaz, Iran. He is currently a last year M.Sc. student at Shiraz University, Shiraz, Iran. His research interests include image processing, machine learning, data mining, graph mining, social network analysis, game theory and multi-agent systems.