Recognition of Persian Numeral Fonts by Combining the Entropy Minimized Fuzzifier and Fuzzy Grammar

H. EBRAHIMNEZHAD
PH.D Student of Elec. Eng.

GH. A. MONTAZER*
Assistant Prof. of IT Eng.

N. JAFARI
PH.D Student of Indust. Eng.

School of Engineering, Tarbiat Modarres University
Address: P.O.Box:14115-179, Tehran, IRAN, Tel: +98(21)88011001

Abstract: In this paper, we propose a combined method for recognition of multi-font Persian numeral characters. At first, the binary image of a character is divided into a fixed number of sub-images called boxes. The average vector distance and angle of each box are computed as features. These features have some variations in different fonts of any character. So, we can employ the fuzzy sets to face with recognition problem. To have a best effect of fuzzy measure for any box features we employ an exponential fuzzification involving two extra parameters, which take account of the variations in the fuzzy sets. These parameters are obtained by minimizing the entropy of fuzzy membership function. After defuzzification, the three most probable candidates of numbers are selected. These candidates are post-processed with another fuzzy recognition system which uses the other features of numerals, i.e. the type of primitives. This combined method increases the robustness of recognition.

Key-Words: Persian Numeral Font, Entropy Minimizing, Fuzzy Sets, Fuzzy Grammar

1 Introduction

There are several methods that address the problem of recognition of numerals depending on the type of features extracted. Recognition of handwritten numerals is a well-researched topic and many techniques of recognition of both machine printed and handwritten characters are available. Recognition of multi-font, multi-size of printed Korean characters was attempted using a two-stage classification method [1]. Handwritten numeral character recognition using fuzzy logic was first attempted by Siy and Chen [2, 3]. In [2], the handwritten numeral is decomposed into straight lines, arcs and circles. A more flexible scheme is proposed in [4] where the decomposition is based on the detection of a set of feature points: terminal point, intersection point and bend point. Recently Suresh and Arumugam have applied fuzzy logic for recognition of numeral handwritten characters [5].

Handwritten numeral recognition by combined self-organizing maps (SOM) and fuzzy rules was offered by Z.Chi [6]. In this study, SOM algorithm is used in the learning phase to produce prototype, which are used to determine fuzzy regions and membership functions. The algorithm generates fuzzy rules by learning from the training patterns. In the recognition phase, a fuzzy rule based classifier is used to classify an input pattern. A doubtful pattern is reclassified by a SOM classifier.

The perturbations due to writing habits and instruments are taken into account in the recognition of off-line handwritten numerals by Ha and Bunke [7, 8]. A new approach to separating single touching handwritten digit strings is presented in [9]. Here the image of the connected numerals is normalized, preprocessed and then thinned before feature points are detected. Potential segmentation points are determined based on decision line in the image. The partitioning path is determined precisely and then the numerals are separated before restoration is applied. A genetic framework using contextual knowledge is proposed in [10] for segmentation and recognition of handwritten numeral strings.

The back-propagation neural network is used by Hanmandlu et al for the recognition of handwritten characters. In that, feature extraction is done using three different approaches, namely, ring, sector and hybrid. The features consist of normalized vector distances and angles. The hybrid approach, which combines the ring and sector approaches, was found to yield the best results [7]. The same features are adopted in [6, 11]. In [11] authors follow the same theme of feature extraction as employed in [6]. However, their study is meant for the case when there are a small set of samples. They make a new fuzzification function, which is a modified version of the function used in [6]. The new function contains structural parameters, which are derived so as to maximize the recognition rate and minimize the entropy. In this paper we use the entropy minimizing method with some difference in box size to recognize the Persian multi-numeral fonts. To increase the recognition rate we select three most probable candidates of this classifier and examine them against the other features of unknown character. We show that our combined method increases the robustness of recognition.

2 The Box Method Approach

The normalization of the numerals is essential because of the different types of fonts, which result in several variations in the shapes and sizes. Therefore, to bring about uniformity among the input numerals, all of them should be made of the same size. For this reason, the numerals are fit into a standard size window of 42 × 32. Every measure has to be taken to preserve the exact aspect ratio of the numerals. This size of window is selected due to the fact that usually height of the numeral is almost 1.5 times more than the width. Experiments show that the normalization of thinned numeral causes the skeleton to be discontinuous. Hence, the un-thinned numeral is normalized and later thinned. The result of thinning constitutes a skeleton of the original pattern, one-pixel wide, consisting of a subset of the original black pixels. This thinned image is then considered for extracting features.

Feature extraction is the crucial phase in numeral identification as each numeral is unique in its own way, thus distinguishing itself from other numerals. Hence, it is very important to extract features in such a way that the recognition of different numerals becomes easier on the basis of the individual features of each numeral. For extracting the features, we have used the Box-Method approach proposed in [6, 11]. The major advantage of this approach stems from its robustness to small variations, easy to implement and relatively high recognition rate. In this approach, the binary array of 30 × 20 is fit into horizontal and vertical grid lines of 6 × 4. Thus as shown in Fig.1, 24 boxes of equal size are derived in which the complete binary image is enclosed. Each box will have the size 5×5, so that, the portions of a numeral will lie in some of these boxes. There could be boxes that are empty. However, all boxes are considered for analysis in a sequential order. The choice of number of boxes is arrived at by experimentation. By considering the bottom left corner as the absolute origin (0,0), the coordinate distances (Vector Distances) of all the pixels of a box are calculated. By averaging the total of all distances with the number of pixels in a box, a Normalized Vector Distance for each box is obtained. Therefore, 24 vector distances corresponding to 24 boxes will constitute a feature set. However, for empty boxes, the value will be zero. The Vector Distance is calculated for each pixel at coordinates (i, j) with respect to the origin, using the formula:

$$d(i,j)=\sqrt{i^2+j^2} \tag{1}$$

Similarly, for all pixels in a box, the vector distances are computed and normalized with the total number of pixels available in that box using the formula:

$$d_{av}(b)=\frac{1}{n_b}\sum_{k=1}^{n_b} d_b(i_k, j_k) \tag{2}$$

Where, n_b is number of pixels in a box and b is the number of boxes. The angle features are not considered here for recognition, as they are not as much effective as the distance features.

3 Feature Identification System

The extracted features are used to identify multi-font numerals. Therefore, a database of numerals of different fonts is required to extract the features, which form the knowledge base. We have considered here a standard database of multi-font numerals consisting of 24 different fonts. Only nine samples of each font are taken into account, as there is not much difference between the individual samples. Thus we have 216 samples in all (Table 1). The features extracted from the size normalized and thinned binary array, form the input for recognition process. The identification system is designed so as to optimize the structural parameters in the fuzzification function, in order to enhance the recognition rate.

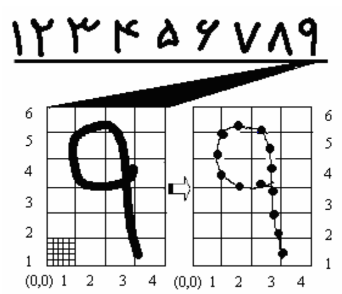

Fig.1 Numeral Persian'9' enclosed in 6×4 grid

Table 1 Database of Persian multi-font numerals

Font	Numerals	Font	Numerals
Ramsar	123456789	Zar	123456789
BElham	123456789	Shiraz	123456789
Badr	123456789	Compset	123456789
Siavash	123456789	Mitra	123456789
Koodak	123456789	Arial	123456789
Nasim	123456789	Majidsh	123456789
Nazanin	123456789	Davat	123456789
BFerdosi	123456789	Farnaz	123456789
Naskh	123456789	Yagut	123456789
Negar	123456789	Traffic	123456789
Traditional Arabic	123456789	Tabassom	123456789
Lotus	123456789	Kamran	**123456789**

In order to recognize the unknown numeral set using fuzzy logic, an exponential variant of fuzzy membership function is selected. The fuzzy membership function is proposed in [6, 11] based on the normalized vector distance. The application of fuzzy logic to the recognition of numerals was proposed by Hanmandlu et al and was extensively used in [6]. The concept of fuzzy sets is as follows:

If there are 'N_f' possible features for each numeral and if there are 'N_s' such samples, then a particular feature from each of the samples forms a fuzzy cluster. Thus, for each particular numeral, the resultant matrix thus formed is $N_s \times N_f$. The reference numeral data set is obtained from training samples. The mean and variances are computed for each of the 24 clusters and taken as Knowledge Base (KB). The procedure is repeated for all reference numerals. Therefore, there are 24 mean and variances corresponding to 24 features for each of the 9 numerals. Given a very large number of samples, by choosing the fuzzification function, the membership function of each feature value in the cluster can be determined. However, we need to compute the membership functions for the features of unknown numerals and not the reference numerals. The unknown numeral features are matched with all reference numeral features available in the form of KB. It is possible to compute the membership functions by associating the features of the input numeral with the clusters. The KB consists of means m_i and variance σ_i for each of the 24 clusters.

For an unknown input numeral x, the 24 features are extracted using the box method. The membership function is chosen as:

$$\mu_{xi} = \exp\left(\frac{-(x_i - m_i)^2}{\sigma_i^2}\right) \tag{3}$$

Where, x_i is the i^{th} feature of the unknown numeral. If all x_i 's are close to m_i's which represent the known statistics of a reference character, then unknown numeral is identified with this known numeral because all the membership functions are close 1 and hence the average membership function is almost 1 as explained below:

Let, $m_i\ (r),\ \sigma_i\ (r)$ belong to the r^{th} reference numeral with $r = 1 \ldots 9$. We then calculate the average membership as:

$$\mu_{av}(r) = \frac{1}{N_f}\sum_{i=1}^{N_f}\exp\left(-\frac{(x_i - m_i(r))^2}{\sigma_i^2(r)}\right) \tag{4}$$

Then $x \in r$ if $\mu_{av}(r)$ is the maximum for $r = 1,...,9$.

4 Modified Fuzzification Functions

The fuzzification function (3) does not perform well when we have a limited database such as the case of multi-font numerals, shown in Table1. This is because some of the fuzzy sets have a very small variance and others have a large variance. So, we use a scheme to separate the fuzzification functions on the basis of the value of variance. In doing so, our objective is to get larger values for the membership functions. Accordingly, we split up the equation (3) into the following two exponential functions:

$$\mu_{xi}(r) = \exp\left(\frac{-|x_i - m_i(r)|}{\sigma_i(r)^2}\right) \quad for \quad \sigma_i^2(r) \geq 1 \tag{5}$$

$$\mu_{xi}(r) = \exp\left(-(x_i - m_i(r))^2 . \sigma_i(r)^2\right) \quad for \ \sigma_i^2(r) < 1 \tag{6}$$

The bifurcation of the membership function into two different functions depending on the variance is compelled by the fact that some variances are too small and some are too large. This has spurred the choice of a new membership function involving structural parameters s and t are given by:

$$\mu_{xi}(r) = \exp\left(-\frac{(1 - S(r)) + S(r)^2 . |x_i - m_i(r)|}{(1 + t(r)) + t(r)^2 \sigma_i(r)^2}\right) \tag{7}$$

The new variance and the new mean are the functions of the variance and mean of original clusters. Thus the structural parameters s, t model the variations in the mean and variance over all 24 clusters (boxes). For $s=1$ and $t=-1$ equation (7) will reduce to equation (5). We now present the estimations of these parameters using an optimization approach. The values of s and t are chosen so that the average membership function defined in (8) is maximized. The average membership function of any unknown numeral must be close to 1, and then its features are fit into the statistics of a known numeral.

$$\mu_{av}(r) = \frac{1}{N_f} \sum_{i=1}^{N_f} \mu_{xi}(r) \quad (8)$$

Hence, for each numeral character we find the different *s* and *t* parameters. In order to reduce the uncertainty in *s* and *t*, we define the entropy of each numeral character *E(r)* as:

$$E(r) = \frac{1}{N_f Ln2} \times \sum_{i=1}^{N_f} -\left[\mu_{xi}(r)\ln\mu_{xi}(r) + (1-\mu_{xi}(r))\ln(1-\mu_{xi}(r))\right] \quad (9)$$

Since the maximum value of $\mu_{av}(r)$ is 1, instead of maximizing the equation function (8) we can also minimize the function *J(r)* in equation (10).

$$J(r) = (1 - \mu_{av}(r))^2 \quad (10)$$

The objective function to be minimized is therefore chosen as:

$$G(r) = E(r) + \lambda(r) \cdot J(r) \quad (11)$$

Now, in order to find the parameters *s* and *t*, we use the gradient descent technique:

$$s^{new}(r) = s^{old}(r) - \varepsilon_1 \cdot \frac{\partial G}{\partial s} \quad (12)$$

$$t^{new}(r) = t^{old}(r) - \varepsilon_2 \cdot \frac{\partial G}{\partial t} \quad (13)$$

$$\lambda^{new}(r) = \lambda^{old}(r) - \varepsilon_3 \cdot \frac{\partial G}{\partial \lambda} \quad (14)$$

The learning factors ε_1, ε_2, ε_3 are chosen as 0.01. The initial values of λ, *s* and *t* are taken as -1, 1 and 2 respectively. With the choice of learning factors, the converged values of *s* and *t* are obtained as given in Table 2. After estimating *s* and *t* parameters for each numeral (r =1,…,9), we use equations (7) and (8) to calculate the average membership function of each numeral for unknown character and select the three most probable candidates. These candidates are post processed to find the unknown character.

Table 2 Estimation of structural parameters

Persian Numeral character	t	s
۱	1.9980	0.9898
۲	2.0058	1.1812
۳	1.9960	1.1112
۴	1.9964	0.9788
۵	1.6791	0.9177
۶	2.0494	1.1670
۷	2.0000	1.0181
۸	1.9290	0.8727
۹	1.9869	1.0376

In the post processing step we employ synthetic approach that is based on the analysis of the skeleton of numeral. In a skeleton image, information of end points, junction points and branches make it possible to identify numerals. There are three basic elements of numeral skeleton characters: straight line, curvature with different orientation and loop. A fuzzy curve is defined with two parameters: curvature and orientation. Each parameter has been defined by a fuzzy set. An information vector *FC [C,O]* is used to descript the fuzzy curve, where *C* and *O* represent the curvature and the orientation, respectively. The orientation is defined as the angle between the real axis and the perpendicular to the line formed by the two end points of the curvature (see Fig.2). The range of the orientation parameter is [0, 360] for arc segments and [-90, 90] for line segments. To provide a compact representation of all the possible orientations of arc segments, four fuzzy sets can be used: left, right, up, and down orientations. For line segments we use other fuzzy sets i.e. Horizontal, Negative slop, Vertical and Positive slop. The Orientation membership functions are shown in Fig.2 , where triangular membership functions are used. Since the variability of numerals, there is no perfect straight line, arc and loop. So the idea of fuzzy line, fuzzy arc and fuzzy loop will be used to characterize shapes in numerals. Here we use the measurement *C* to define the membership function of degree of curvature [12]; *C=D/L*, where *D* is the distance between its two nodes and *L* is the curvature length (see Fig.2).

If the orientation changes monotonically, it means the point is still in the integrated curve until it is violated. A threshold is added in case of local small changes will end the integrated curve search prematurely. If a junction point is met, the algorithm will select the nearest curve. After separating the elements of numeral skeleton, orientation and curvature of each element is calculated. Then this information is compared with the structure of three candidate numerals which obtained in the earlier recognition system and one of them is selected as

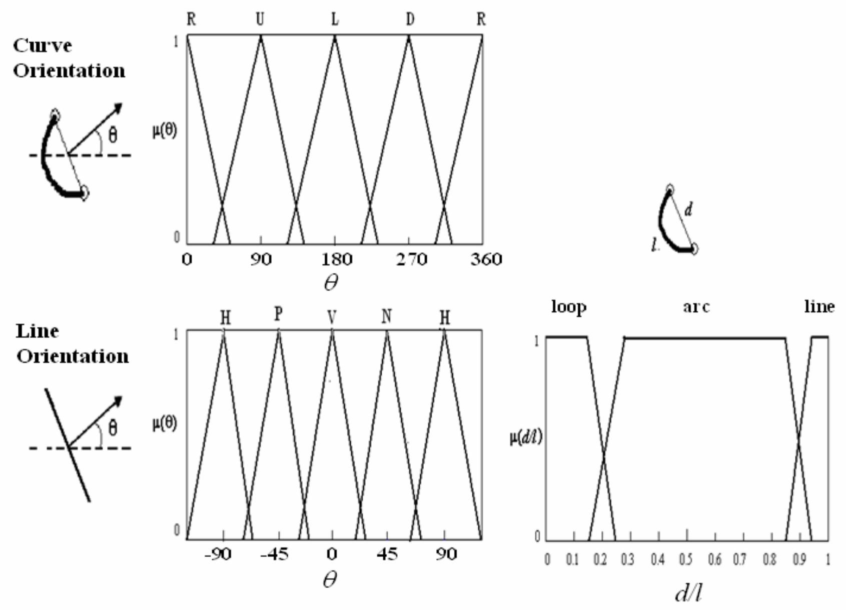

Fig.2 Orientation and curvature membership function

unknown numeral. In Fig.3 the structure of Persian numerals is illustrated.

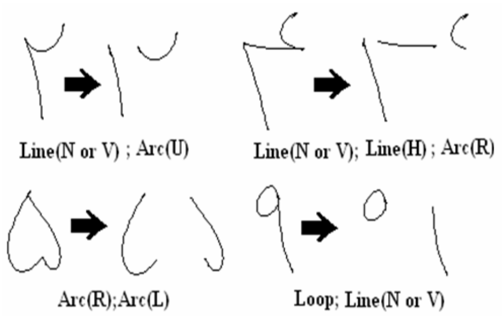

Fig.3 Fuzzy grammar to explain the structure of Persian numeral characters

5 Results

The various types of fonts introduce large variations in the numerals. In this work, we have used 24 Persian numeral fonts. The structural parameters are given in Table 2. We have obtained a recognition rate of 75.9% with the function (4) and 95.5% with the functions (5) and (6). But with the modified membership function (7), the recognition rate has risen to 97.5% as shown in Table 3. Fault is occurred for numerals 2, 4 and 7. The proposed combined method and post processing gives very good result with 100% recognition rate and there is no fault in this method.

6 Conclusion

In this paper, we have devised a combined fuzzy method for the recognition of unconstrained, isolated, multi-font machine printed characters, especially, numerals. The box method has been used for obtaining the distance features. The mean and variances of these features are adapted using two structural parameters, which model their variations across the different fonts and styles of numerals. A recognition rate of 100% was achieved with the combined fuzzy method. Suggested future work may include testing of the algorithm with more numeral data, consisting of many other fonts, which have not been considered in the present study.

Table 3 Recognition rates of Persian numerals

Persian Numeral character	Recognition Rate (%)			
	Fuzzifier function Eq.(4)	Fuzzifier function Eq.(5,6)	Fuzzifier function Eq.(8)	Combined Method
۱	91	96	100	100
۲	74	91	91	100
۳	65	100	100	100
۴	91	91	91	100
۵	100	100	100	100
۶	96	91	100	100
۷	87	91	96	100
۸	83	100	100	100
۹	78	100	100	100
Overall	**75.9**	**95.5**	**97.5**	**100**

References

[1] J.S.Lee, O.J.Kwon and S.Y.Bang, Highly accurate recognition of printed Korean characters through an improved two stage classification method, *Pattern Recognition,* Vol. 32, No. 12, 1999, pp. 1935-1945.

[2] P.Siy and C.S.Chen, Fuzzy Logic for handwritten numeral character recognition, *IEEE Trans. Systems, klan and Cybernetics,* Vol. 4, 1974, pp. 570-575.

[3] A.Malaviya1, L.Peters, Fuzzy handwriting description language: FOHDEL, *Pattern Recognition*, Vol. 33, 2000, pp. 119-131

[4] G.Baptista and K.M.Kulkarni, A high accuracy algorithm for the recognition of hand written numerals, *Pattern Recognition,* Vol. 21, 1998, pp. 287-291.

[5] R.M.Suresh and S.Arumugam, Fuzzy technique based recognition of handwritten characters, *Image and Vision Computing*, Vol. 25, 2007, pp. 230–239.

[6] M.Hanmandlu, K.R.M.Mohan, S.Chakraborty, S.Goyal and D.R.Choudhury, Unconstrained handwritten character recognition based on fuzzy logic, *Pattern Recognition,* Vol. 36, No. 3, 2003, pp. 603-623.

[7] M.Hanmandlu, O.V.Ramana Murthy, Fuzzy model based recognition of handwritten numerals, *Pattern Recognition*, ARTICLE IN PRESS, accepted 21 August 2006

[8] T.M.Ha and H.Bunke, Off-Line handwritten numeral recognition by perturbation method, *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. 19, No. 5, 1997, pp. 535-539.

[9] A.Elnagara and R.Alhajj, Segmentation of connected handwritten numeral strings, *Pattern Recognition,* Vol. 36, 2003, pp. 625 – 634

[10] J.Sadri, C.Y.Suen, T.D.Bui, A genetic framework using contextual knowledge for segmentation and recognition of handwritten numeral strings, *Pattern Recognition,* Vol.40, 2007, pp. 898 – 919

[11] M.Hanmandlu, M.H.M.Yusof, and V.K.Madasu, Fuzzy Based Approach to the Recognition of Multi-Font Numerals, *2nd National Conference on Document Analysis and Recognition (NCDAR-2003)*, July 2003.

[12] X.Fang and AT.Alouani, Unconstrained handwritten numeral recognition using fuzzy rule-based classifier, *IEEE 2002*.