PRESENTING A DECISION-TREE-BASED METHOD FOR RECOGNITION OF PERSIAN HANDWRITTEN NUMBERS

Navid Samimi Behbahan*

Zohreh Mousavinasab*

Abstract

This paper presents an effective method for recognition of Persian handwritten numbers. The method is based on structural features and decision-tree learning technique. The presented method includes three stages: First a preprocessing is made on the number, in this way the entry number has been thinned down while at the same time some noise coming as a result of lack of congruence in numbers thickness is deleted. In the second stage four properties applied at the stage of recognition have been extracted. The third stage comprises of creating a decision genealogy (tree) related to the order of classification. The total dataset used in the experimental stage is the database of Hoda documented upon by researchers. The experiment's results indicate the efficacy on the part of the method suggested for recognizing handwritten Persian numbers.

Key words: Persian handwritten numbers, features extraction, decision tree, thinning

* Sama Technical and Vocational Training School, Islamic Azad University, Omidiyeh Branch, Omidiyeh, Iran.

1. Introduction

Recognizing handwritten numbers is one of the significant problems within the area of light letters cognition, which is of much application. As an example for handwritten numbers recognition systems applications, reference could be made to reading digital data put into forms. The use of a practical handwritten numbers recognition system faces a number of challenges of which the most important would be the necessity for a high level of recognition rate. In the field of Persian language, simply because of great similarity among digits in addition to differences in the way they are written, creating a recognition system with an acceptable degree of exactitude for practical purposes comes across some difficulties. It is for this reason that expanding methods to improve their precision would be of necessity. During the last few years, various works have been done over the issue of Persian and Arabic handwritten numbers and letters' recognition. In a piece of research conducted by Darvish et al, a shape congruence algorithm is made use of to recognize Persian handwritten digits. For each and every sampling point on the shape's connector, description is arrived at by means of the placement distribution of other points' connector(s) [1]. In another piece of research by Parvin et al, another method for improving upon the functionality of recognition system has been put forward. The original idea in the suggested method would be use of classifiers on a binary basis [2]. In yet another research by Alizadeh et al, some methodology based on genetic algorithm to make a neural network grouping using a classifying pick-up method of giving weights based on opinion has been propounded [3]. The research conducted by Shahabi and Rahmati, use has been made of Gabor filter banks which is suitable for the construction of Persian handwritten texts in addition to visual system [4]. Still in another work by Parvin et al, application has been made of categorizing even classifiers to boost this group of classifiers. These can reduce the error rate for more precision in features space [5]. Another research has used Bays' classification moment torque to recognize Persian handwritten digits [6]. Masroori has applied dynamic temporal torsion algorithm to recognize the numbers [7].

What we have done is based on bringing out a series of constructional features from amongst the set of handwritten numbers. These features are the existence of an enclosed space within the digit, branching-out and terminal points, the directionality of semi-circles and the degree of pixel density in various areas. Later on, the genealogical decision-making tree would have been created upon such features to be evaluated.

Mention would be made of what has been said in various sections as this article goes on. In the second part, the focus of attention has been on introducing the dataset applied. In the third part, the preprocessing stage is brought to attention which includes how to drop noise and extract digit skeleton. In the fourth part, we shall talk of the methodology of bringing out features. The fifth part is devoted to classification stage. In the end, will come the results out of the suggested model.

2. Hoda dataset features

Hoda database which is documented upon by researchers is also of good use in this research. The Hoda handwritten numbers set is the first sizable Persian handwritten numbers comprising of 102353 samples of black and white handwritten digits. This set has been made during a Master's degree project concerning the recognition of handwritten forms [8]. The data within this set have been extracted from something around 12000 registration forms of M.Sc. entrance examinations of 2005 in addition to the Associate's examinations in Applied and Science Comprehensive University in the year 2004. The properties in this dataset are as follows:

The sample separablility degree: 200 points per inch

The total number of samples: 102352

Educational samples numbers: 6000 samples of each class

Experimental samples number: 2000 samples in each class

Other samples: 22352

3. Preprocessing

The preprocessing stage is made up of extracting digit skeleton (thinning). In this section, use has been made of the accelerating algorithm in order to thin out and get to the skeleton of each number without creating any erosion or disruption. This algorithm gets at a suitable system of skeletons capable of keeping the original shape of the number in such a manner as not to beget forged data. Figure (1) is an example of the output in this stage. The output of this stage is made only of a continuous skeleton. The aim in this stage would be removing problems resulting from differences in handwritten numbers' thickness.

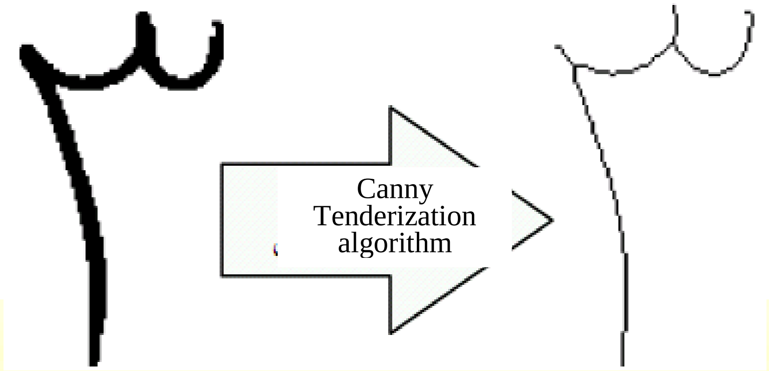


Figure 1: a sample that Tenderized

It is worthy of mention that in a number of digital classification stages, preprocessing has been used, so that the bright thinned numbers' pixels have been omitted from bottom-to-top to such a degree that there might exist more than two bright pixels within a specific line. This causes the dropping of those pixels that have had some negative effects on classification.

4. Features Extraction

One of the most efficacious things to be done for improving the degree of precision significantly in a recognition system of handwritten numbers is using appropriate features to represent numbers. This necessitates a suitable procedure for bringing out features and determining its parameters in an optimal fashion. This part is devoted to investigating into how these properties are extracted and what effective parameters are in handwritten digits. Of great importance is extracting those features capable of setting together the largest number of samples. On the other hand, the larger number of features could lead to a more complex temporal level. The applied features may be categorized into four classes.

4.1. The existence of enclosed space within the digits

This property causes dividing out various samples of digits like zero, five, and nine from among other numbers. Off course, we do know that because of lack of precision in the manner of writing, there is always some noise within the digits. This brings about the non-enclosed space. The method put forward for flexibility against such types of noise would be finding white pixels for each digit to have been surrounded by black pixels on all four sides. In figure (2), samples of such numbers are to be observed.

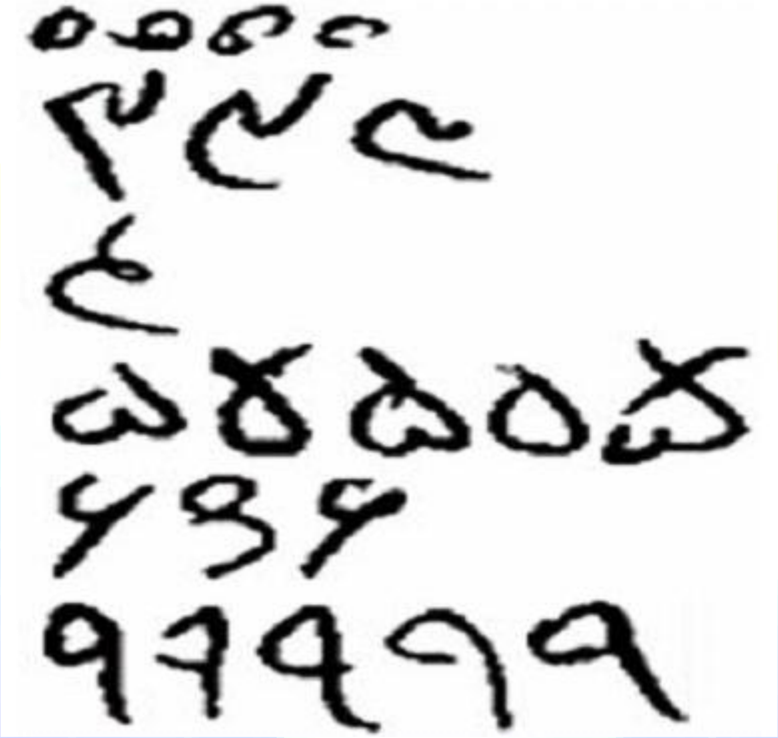


Figure 2: The sample digits that have white pixels surrounded by black

4.2. Terminal and branching-out points

Points of terminus include black pixels after which there could have been no other black pixel. The branching-out points include those on the intersection of three lines [9]. Terminal and branching-out points have been pinpointed in the figure (3). Exactitude in number and placement of such points would inevitably lead to Persian handwritten numbers classification. For instance, digits SEVEN and EIGHT possess two terminal points in the lower or upper half of their images.

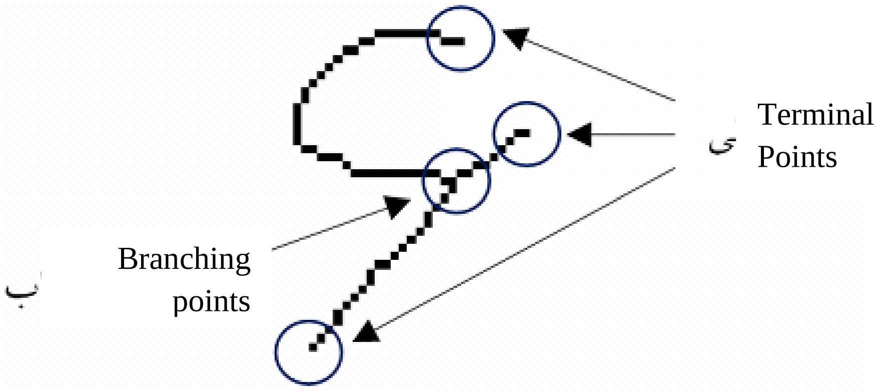


Figure 3: Endpoints and branching number six

4.3. The directionality of the semi-circles

Digits like two, three, four, and six have semi-circles with differing directionalities (figure 4 shows the semi-circles' directionality for the digit THREE). It would be of great significance to recognize the compassing directions on the part of these semi-circles. The intended direction for each class's numbers is similar in various writings. For example, in the digit THREE, there exist three directions upward while it would be to the right for FOUR. As a result, the digit pattern of THREE has white pixels ending in black pixels to the right, to the left, and to the terminus (surrounded by black pixels). Therefore, the digit FOUR has semi-circle to the right.

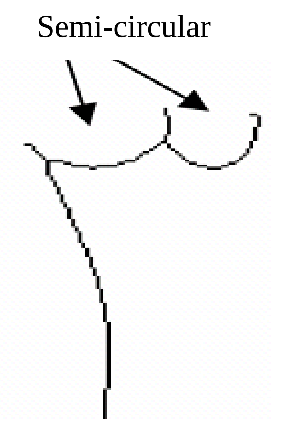


Figure 4: number three has half-circles upward

4.4. Pixel existence density in the lower, upper, right, left half

It would be necessary – in recognizing this point – that a digit is divided into two halves of upper or lower, right or left for making distinct the degree of bright pixels compaction in each area [10]. For instance, in view of figure FIVE, density in the upper half of digits TWO & SIX are more than those of their lower halves (nearly double).

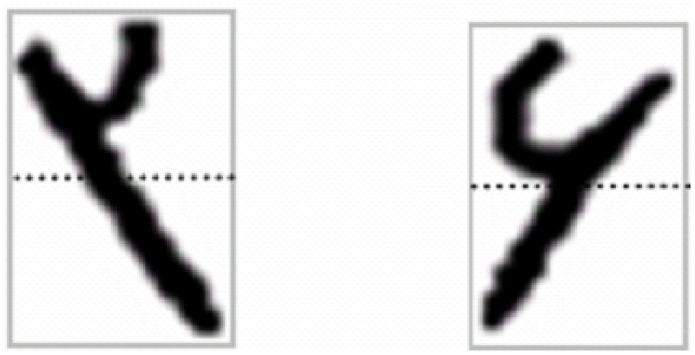


Figure 5: Black pixel density in the upper half

5. Classification Stage

In order to better classify the genealogical decision-making tree, some pick-up has been made of various handwritten numbers (figure 5). These numbers have been recognized in various parts of the decision-making tree.

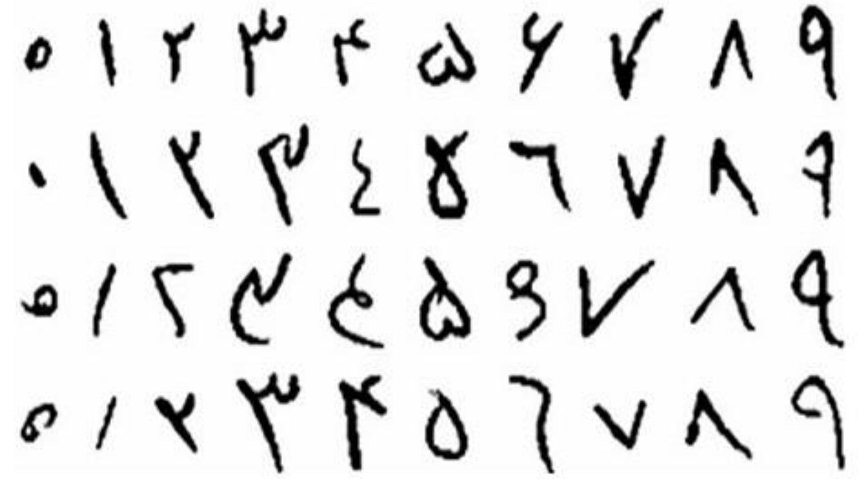
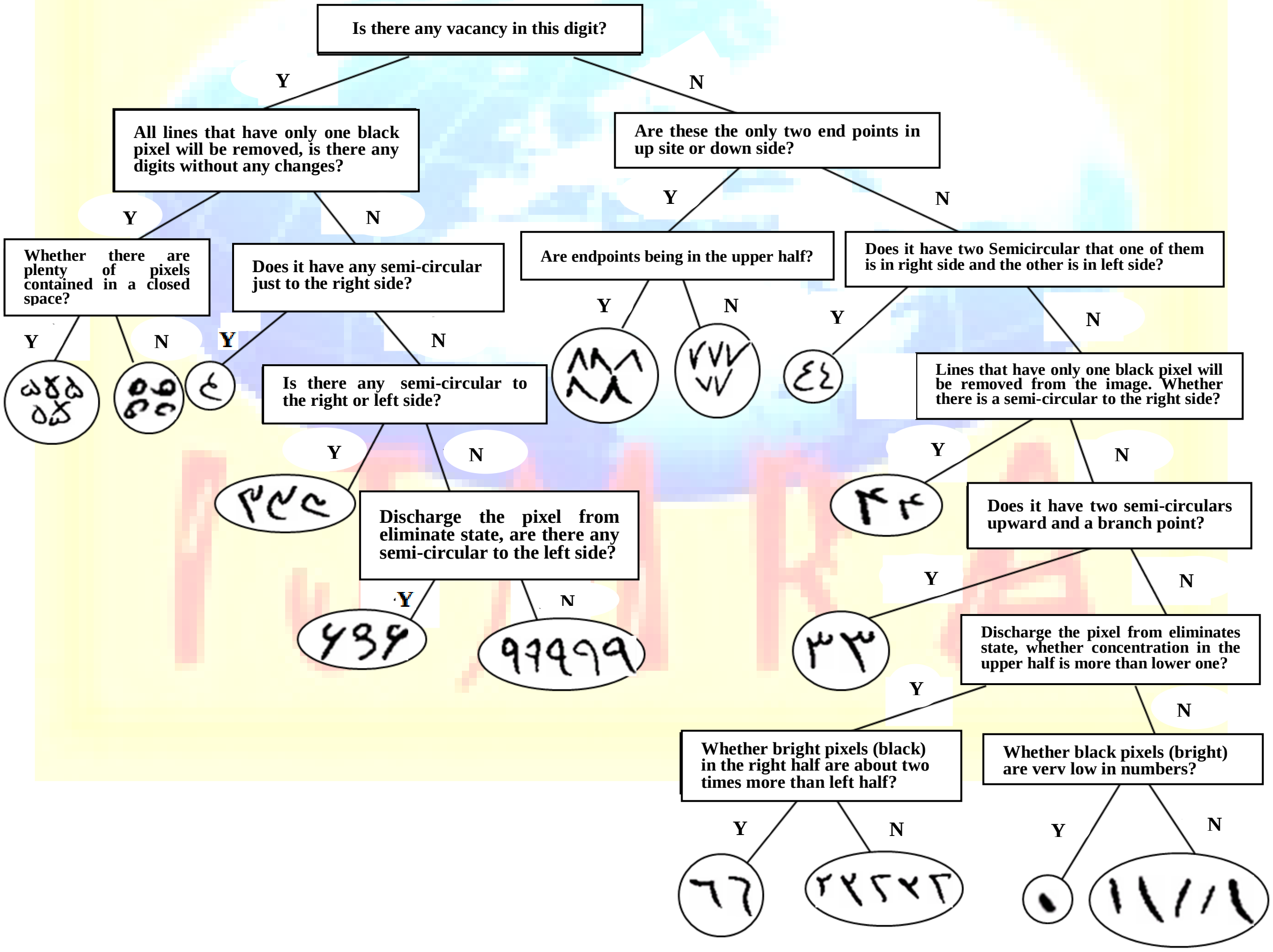


Figure 6: A variety of Persian handwritten digits

The extracted features as described above have been applied in the classification stage. The method for categorizing using the decision-making tree to recognize Persian handwritten digits in this article has been shown in figure (6).

In each stage, the feature to go under investigation has been brought into an oblong. Each sub-tree comprises of a bunch of classes. The classifying features extant in the decision-making tree have been opted for in various stages in such a way as to bring forward the largest number of classifications against the handwritten variety and the noise present within the digits.


Is there any vacancy in this digit?
Y
N
All lines that have only one black pixel will be removed, is there any digits without any changes?
Are these the only two end points in up site or down side?
Y
N
Y
N
Whether there are plenty of pixels contained in a closed space?
Does it have any semi-circular just to the right side?
Are endpoints being in the upper half?
Does it have two Semicircular that one of them is in right side and the other is in left side?
Y
N
Y
N
Y
N
Y
N
Is there any semi-circular to the right or left side?
Lines that have only one black pixel will be removed from the image. Whether there is a semi-circular to the right side?
Y
N
Y
N
Discharge the pixel from eliminate state, are there any semi-circular to the left side?
Does it have two semi-circulars upward and a branch point?
Y
N
Y
N
Discharge the pixel from eliminates state, whether concentration in the upper half is more than lower one?
Y
N
Whether bright pixels (black) in the right half are about two times more than left half?
Whether black pixels (bright) are verv low in numbers?
Y
N
Y
N

6. Conclusion

The algorithm was implemented using MATLAB software. Detection at 100 samples was selected randomly from each class. The results can be seen in Table 1.

Table 1: The proposed system detects the handwritten digits

digit Number	The number of diagnosed samples	The number of undiagnosed samples	Percentage of correct diagnosis
0	88	12	88%
1	97	3	97%
2	92	8	92%
3	86	14	86%
4	90	10	90%
5	89	11	89%
6	88	12	88%
7	100	0	100%
8	99	1	99%
9	92	8	92%
total	921	79	92.1%

A new method based on a new feature for the detection of Farsi handwritten digits was presented. Although there are some challenges to some digits, but overall detection rate is acceptable. The proposed method by combining several algorithms can be used in applications. Strengths of the proposed method can be as low time complexity.

References

[1] A.Darvish, E.Kabir and H.Khosravi, “figure **Adaptation usage in detection of Persian handwritten digits**”, Eleventh Annual Conference of Computer Society of Iran, pp. 285-296, 2005.

[2] H.Parvin, H.Alizadeh, M.Moshki, B.Minaei-Bidgoli, N.Mozayani, “***Divide & Conquer*** **Classification** ***and Optimization by Genetic Algorithm***”, Third 2008 International Conferences on Convergence and Hybrid Information Technology.

[3] H.Alizadeh, H.Parvin, B.Minaei-Bidgoli, “***A New Approach to Improve the Vote-Based*** **Classifier** ***Selection***”, 2008 Fourth International Conference on Networked Computing and Advanced Information Management.

[4] F.Shahabi, M.Rahmati, “***A New Method for Writer Identification of*** **Handwritten** ***Farsi Documents***”, 2009 10th International Conference on Document Analysis and Recognition.

[5] H.Parvin, H.Alizadeh, B.Minaei-Bidgoli, M.Analoui, “***A Scalable Method for Improving the Performance of Classifiers in Multiclass Applications by Pairwise Classifiers and GA***”, 2009 Fourth International Conference on Networked Computing and Advanced Information Management.

[6] R.Azmi and E.Kabir, “**Recognition of Persian handwritten digits**”, Second Iranian Conference on Electrical Engineering, pp. 285-295, 1994.

[7] K.Masrouri and V.Pour-mohseni, “**Handwritten digits recognition using DTW algorithm**”, Fourth International Conference Computer Society of Iran, pp. 1-7, 1998.

[8] H.Khosravi, “**Recognition Persian handwritten digits and letters from nationwide exam registration**”, Master's thesis in electronics field, tarbiyat modares university, 2005.

[9] M.Teshne-lab, D.Afiuni and M.R.Hasan-zadeh, “**Fingerprint identification using intelligent systems**”, Thirteenth Conference of Iran Computer Engineering

[10] Ahmad T.Al-Taani, Saeed Al-Haj, “**Recognition of On-line Arabic Handwritten Characters Using Structural Features**”, 2010 Journal of Pattern Recognition Research pp. 23-27.