A Fast Algorithm for Persian Handwritten Number Recognition with Computational Geometry Techniques

Keivan Borna[1], Vahid Haji Hashemi[2]

[1] **Faculty of Mathematics and Computer Science, Kharazmi University, Tehran, Iran**
borna@khu.ac.ir

[2] **Faculty of Engineering, Kharazmi University, Tehran, Iran**
Hajihashemi.vahid@yahoo.com

Abstract

This paper aims to improve the feature extraction of Persian handwritten number recognition systems. In this paper, we introduced nine new features for detection and recognition of Persian handwritten digits using the technique of finding the smallest enclosing disc in computational geometry. All these features are based on the geometry form of numbers and are much better than the features in terms of accuracy such as gradient which are mentioned as the strongest feature in the literature. In fact our defined features are highly resistant to resize and rotation operations due to the circular form of Persian digits. In terms of feature extraction performance, our newly defined features are tested on Hoda database and the calculated results confirmed the efficiency of these features. Due to the improvement of recognition rate and acceptable speed, our defined features are better than other common features in digit recognition.

Keywords: *Persian handwritten numbers, OCR, feature extraction, computational geometry, smallest enclosing disk, KNN.*

1. Introduction

Recognition of handwritten documents including letters and numbers and converting them to typed text has always been one of the favorite topics in image processing. In the field of the English letters and contents, many different methods have been implemented and commercial versions of them can be seen in the form of OCR software which are usually provided with equipment's such as scanners. Obviously, the accuracy of the numbers and letters recognition in English language is very high as many previous works admit this. These methods usually have very small errors. On the other hand many works on Persian and Arabic letters and numbers detection have been done [11]. Obviously work on Persian digits is very difficult due to the similarity of Persian handwritten digits. For example, the small difference between configurations of numbers 1, 2, 3 (in Persian) and also the variety of handwritings. This research aims to provide simple and highly accurate features in recognition of Persian digits which are explained explicitly after we give a literature review at the end of next section.

The structure of this article is as follows: Section 2 provides a literature review on the background. In Section 3, our algorithm is presented and the extracted features are used to train a classifier (KNN). The implementation and test results on the Hoda binary image database are provided in Section 4.

2. Literature Review

Several works have already been done in Recognition of handwritten documents. Basically they could be divided into three categories in general form:

2.1 Methods based on geometric features

These methods use geometry of figures to extract features, the statistical invariants, histograms, directional properties and even fractals can be used to identify digits [2]. For example, in [3] the authors used correspondence between figures and digits. First, they extracted the features of a set of handwritten digits and then for a new digit, its rate compared with previous digits, was the recognition digit criteria. The disadvantage of this feature extraction method was its complexity. Selection a proper correspondence algorithm was also a challenge. In [4] gradient and histogram functions, that represent energy in pixels and different points of the image, are used. This idea is better than previous methods in terms of the feature extraction rate and the accuracy. Superiority of the gradient method is its ability in extension to Latin numbers [7].

2.2 Structural features such as Fourier and wavelet descriptors

These methods use the image analysis in other spaces such as frequency (Fourier) or spaces other than the main space of the image. In this case, first the image must be mapped to the desired spaces, which are of course computationally time taking, and then the characteristics of each digit in this new space will be compared with the input digits. Superiority of these methods is their robustness against noise, rotation and scaling of handwritten digits. The disadvantages of these methods are computational load in the new space and the complexity of defined features. For example, [8] has been introduced some examples of these methods.

2.3 Using fuzzy classifiers, fuzzy rules and neural network

In this category, more focus is on the classifier selection or conversions that can directly analyze and identify digits without pre-processing. For example, [1] used multilayer neural networks, [5] used SVM, fuzzy, PCA and linear discriminator simultaneously.

In this paper, in order to increase the performance, several new spatial features are proposed for detecting Persian handwritten digits, which are very tangible, understandable, and have high accuracy. For this purpose, we used the ``smallest enclosing discs'' technique from computational geometry that due to the curved nature of Persian digits, will naturally increase the accuracy of our approach. In addition, ability to detect rotation and lack of size sensitivity is the characteristics of our proposed method.

3. Our methodology for feature extraction

After receiving the input image, a binary image, in the pre-processing stage, the skeleton of image, with one pixel thickness, is extracted using MATLAB software. The basic assumption is that there is no possible error in the skeleton extraction process and the handwritten digits skeleton is correctly extracted. Figure 1 shows a sample digit of the Hoda database with its skeleton.

Fig. 1 A sample digit of the Hoda database with its skeleton

Based on the present points in the skeleton, the circle with smallest radius covering all the points, ``smallest enclosing disk'', is calculated and its radius is determined. Figure 2 shows this circle:

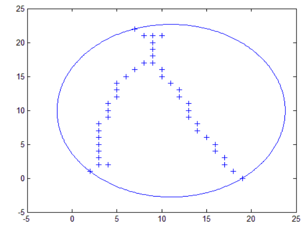

Fig. 2 The minimum radius circle around a Persina digit

To calculate the minimum circle, an algorithm from [9] is used. It ought to be mentioned that the fast and efficient computation of the smallest enclosing circle is one of the basic problems in computational geometry that has many applications. Time complexity of these algorithms are O(n) and relatively efficient codes are presented to obtain them in [10] . A Relatively efficient pseudo-code for finding the smallest enclosing circle is given below:

Algorithm 1: Find the smallest enclosing circle

- Input: A set P of N points in the plane
- Output: Smallest enclosing circle of P
- Take a random order of points of P.
- Let D2 be the smallest circle through P1 and P2.
- For i from 3 to N, repeat:
- If Pi is a member of the circle by now, then the circle will not change and the index will be incremented.
- Otherwise, find a new circle, name it Di, which has Pi on its boundary and other points are inside or on it.
- Finally, the output is Dn.

After calculating the center and radius of the smallest enclosing disk, to specify handwritten digits from the skeleton number, the following 9 features can be defined:

1. The average distance of the skeleton points from the center of the circle divided by the radius: This feature in a perfectly circular shape like 0 are close to one and in the rod-shaped digits like 1 are close to half. It is obvious that by this division, the handwritten will be normalized and will have no effect on the calculation.

2. The variance of distance of skeleton points from the center divided by the mean value: In zero, all point's distances from the center are almost constant and therefore the variance is close to

ACSIJ Advances in Computer Science: an International Journal, Vol. 3, Issue 3, No.9 , May 2014
ISSN : 2322-5157
www.ACSIJ.org

zero. As long as the distances from the center change more, this number would also be higher. Table 1 presents the mean features for each 1000 random samples of handwritten digits. Very large difference of characteristics 0 and 5 with other figures is obvious.

3. The average of absolute value of the distance of the vertical diameter to the radius: For numbers 9, 6, 4, 3, 2 and 1 which has a vertical base, this number should be relatively low particularly in 1 and very high for numbers of 5, 7, 8, 0. For 0, this number is approximately 0.5.

4. The average of absolute value of distance of the horizontal diameter: Digits 0 and 1 have the largest value. Digits 2, 3 and 4 may give this number close to but less than 0.5 due to the points in the bottom quarter. However, this feature is not distinguishing according to Table 1, 2.

5. The ratio of number of points in the right hand side of the vertical diameter to all points of the image

6. The ratio of the number of points above the horizontal diameter to all points: This is a very good criterion for determining the digits difference. For example, in digits 0 and 1 it must be equal to 0.5, while in digits 2 and 3, the points above the horizontal diameter is over than the points below the horizontal diameter and this feature will be more than 0.5. In addition, this feature in digit 3 is more than digit 2 certainly.

7. The number of points in the lower half plane that their distance to the center is changed: In other words, if two points of the image skeleton in the bottom half plane has the same height, only one of them is considered. In the symmetrical structures such as 7, 8 and 0, the points in each row, or the points with the same height, is exactly two. Thus the ratio of the number of points with different heights in the bottom half plane to the total number of points in the bottom half plane is 0.5. Digits 7, 8 and 0 have this value equal to 0.5 whereas in digits 4, 3, 2 and 1 this value is 1. Note that this feature is not sensitive to the arc.

8. The number of points in the upper half plane that their distance to the center is changed: This feature can be a very good feature to differentiate between the numbers 1, 2, 3, 4 and 9. It is obvious that Feature 7 can separate digits 7, 8, 5 and create discrimination between the digits 0, 1, 2, 3, 6 and 9. By using Features 7 and 8, the digits can separate to some classes with using binary classifiers such as SVM or ADABOOST. Then features 1-6 will be used for detecting the digits with high accuracy.

9. The ratio of Feature 7 to 8: An additional feature shows the symmetry in the top and bottoms half of the image that has a value close to 1 in the digits 0, 1, 7 and 8.

Table 1: Our 9 proposed features computed for digits 0 to 4

Numbers	0	1	2	3	4
Feature 1	0.71	0.54	0.60	0.64	0.63
Feature2	2.14	35.56	26.25	19.39	23.11
Feature3	0.47	0.12	0.26	0.34	0.33
Feature 4	0.45	0.52	0.51	0.48	0.47
Feature 5	0.51	0.48	0.33	0.32	0.30
Feature 6	0.51	0.51	0.64	0.70	0.67
Feature 7	0.40	0.95	0.96	0.95	0.87
Feature 8	0.41	0.96	0.54	0.38	0.42
Feature 9	1.01	1.00	1.82	2.58	2.20

Table 2: Our 9 proposed features computed for digits 5 to 9

Numbers	5	6	7	8	9
Feature 1	0.70	0.58	0.63	0.62	0.61
Feature2	13.94	34.89	21.15	20.94	36.45
Feature3	0.43	0.29	0.39	0.39	0.29
Feature 4	0.47	0.47	0.42	0.41	0.49
Feature 5	0.49	0.49	0.44	0.41	0.44
Feature 6	0.40	0.67	0.45	0.54	0.71
Feature 7	0.29	0.91	0.56	0.51	0.91
Feature 8	0.51	0.45	0.55	0.57	0.38
Feature 9	0.57	2.05	1.04	0.91	2.43

4. Implementation and results

This section firstly presents the pseudo-code of feature extraction and the use of the classifiers. Then our results are presented.

4.1 Pseudo- code of our feature extraction algorithm

Algorithm 2: Our feature extraction algorithm

- Input: A binary image of the figure skeleton
- Output: The image corresponding to the digit

1. Using Algorithm 1, find the smallest enclosing circle to cover all the points of the skeleton. Store the center and radius of the obtained circle.
2. Calculate the distance of the all points to the center of the circle in the first step.
3. Calculate the feature 1, 2 based on the obtained distances and the radius of the circle in the first step.
4. Calculate the remaining features based on the skeleton points location on the circle in the first step.
5. Send calculated values to the classified parts.

4.2 Classifying

There are many algorithms using fuzzy logic and neural networks techniques which are known for classifiers. According to the suitable definition and relatively linear nature of the features in this paper, the KNN algorithm is used. For raising the accuracy, we can use the similar algorithm such as multiclass SVM neural network or the combination of the above works with fuzzy logic.

4.3 Implementation of the Hoda database

The database [6] contains 60,000 training samples and 20,000 test samples that are binary and can be converted into standard MATLAB software with an attachment file. After implementation, the pre-processing incorrectly extracted skeletons are corrected. Then by applying the features to the training and testing categories, the following results are obtained.

Table 3: Correction percent

Training sample	Testing sample
100%	99.4%

By removing features 4 and 9 from the features table and reapplying the system, Table 3 changed as follows.

Table 4: Correction percent after removing from the features

Training sample	Testing sample
100%	97.2%

Comparing with similar methods, the proposed method has 100% accuracy in the training stage and increasing at least 1% efficiency in the testing stage. The results based on the latest and best methods before [5] are given in Table 5 for comparison.

Table 5: Correction percent after removing from the features

Training sample	Testing sample
97.3%	96%

Discussion and conclusion

In this paper, new features for detection and recognition of Persian handwritten digits are introduced based on the computational geometry insights. These features are all based on the geometry of digits and work much better, in terms of accuracy, than to other features. In terms of computational complexity, the only time consuming part is computing the minimum circle which covers all points which is of O(n) order and the algorithm is considered highly efficient due to the small size of image points. In terms of efficiency, the feature extraction speed in a core i5 CPU 2.66 GHZ, 480 M, for 10,000 different numbers is taken 28.5 seconds which is equivalent to 1.99 seconds in a typical 700-words page and so this time is considered very fast. Regarding the improvement of recognition rates, based on the above comparisons, the introduced features in digit recognition are much better than the common features.

References

[1] M.Haidarazadzadeh, N.Bayatti , M. Azimifar, "Handwritten number recognition using multi-layer neural

network", first conference in IT and computer engineering, 2011.

[2] M.Nahv, "Improvement of feature extraction and classification of Persian handwritten numbers in large databases", MSc. Thesis, 2010.

[3] A.R.Darwish, E.Kabir, H.Khosravi, "Image matching usage in Handwritten number recognition", Art and architects magazine, Winter 2005, pp. 37-47

[4] H.Khosravi, E.Kabir," Two simple and efficient features in persian handwritten number recognition" ,4th conference of machine vision and image processing 2006.

[5] M.Salehpour, "Rotation and scale invariant Persian handwritten number recognition using fuzzy SVM classifiers based on K-means clustering" 1st conference in Persian handwritten processing, ICPLP 2012.

[6] H.Khosravi, " Handwritten numbers and letters recognition in the entrance exams registration forms", MSc. Thesis, 2005.

[7] O.D. Trier, A.K. Jain and T. Taxt, "Feature extraction methods for character recognition – a survey". Pattern Recognition, 1996. 29(4), pp. 641-662.

[8] C.L. Liu, et al., "Handwritten digit recognition: benchmarking of state-of-the-art techniques". Pattern Recognition, 2003. 36(10), pp. 2271-2285.

[9] Hoda binary images Computational Geometry in C (Second Edition), Cambridge University Press, 1998.

[10] A. Amin; Off-Line Arabic Character Recognition: The State of the Art, Pattern Recognition; Vol. 31; No. 5; 1998, pp. 517-530.

Dr. Keivan Borna He is an Assistant Professor in the Department of Computer Science at Faculty of Mathematics and Computer Science at Kharazmi University of Tehran since 2008. He completed his Ph.D. in September 2008 from the Department of Mathematics, Statistics and Computer Science of University of Tehran in Computational Commutative Algebra where he previously received his M.Sc. in 2004. He was a visiting scholar in ``Dipartemento di Matematicha, Universita' di Genova- Italia" and ``Department of Mathematik and Informatik at Essen University, Germany'', from Sep. 2007 to Apr. 2008 during his graduate work. He is interested in studying and researching in interdisciplinary topics like ``Computer Algebra'', ``Cryptography'', ``Approximation Algorithms'' and ``Computational Geometry'', and managed it to write several papers in such areas. He is the author of the ``Advanced Programming in JAVA'' (in Persian) and is the member of ``Elite National Foundation of Iran''.

Vahid Haji Hashemi He is currently a master student of Computer Engineering at Faculty of Engineering at Kharazmi University of Tehran.