Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 8, No. 4, November 2017), Pages: 13-21
www.jacr.iausari.ac.ir

Categorization of Persian Detached Handwritten Letters Using Intelligent Combinations of Classifiers

Hossein Sadr[1,2,3], Mojdeh Nazari Solimandarabi[2,3], Mahsa Mirhosseini Moghadam[3]
1) Department of Computer Engineering, Parand Branch, Islamic Azad University, Parand, Iran
2) Young Researchers and Elite Club, Lahijan Branch, Islamic Azad University, Lahijan, Iran
3) Department of Computer Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran
sadr@qiau.ac.ir; mozhdeh_nazary@yahoo.com; mahsa.moghaddam1368@yahoo.com

Received: 2016/06/13; Accepted: 2016/10/23

Abstract

Detecting optical characters is considered as the main responsibility to convert printed documents and manuscripts to digital format. In this article, detecting Persian handwritten letters using the combination of classifiers and features are assessed employing geometric and statistical sections' features. In order to detect each letter, it is divided into two parts; the major and the minor parts. Then, some features are presented for them. Preprocess algorithm prepares the possibility to unify dimension features for multiple words and delivers them to classifier for detection. The hierarchy classification can be obtained by separating the letters. In the following, the optimal answer will be reached by using GA method of different SVM, ML and KNN classifications. Extraction algorithm of required features is proved by using the evaluation of the basis of PCA. Empirical results represent classification of 94.3 and 92 accuracy in simple and multiple parts in 20 times repetition, respectively.

Keywords: *Classifiers' Combination, Optical Character Recognition, Persian Handwritten, Reducing Feature*

1. Introduction

Using Heuristic Classifiers' Combinations will cause the automatic detecting of Persian letters in mailings classification, detecting handwritten texts and etc. In this regard, different computing methods can help people. Detecting letters will be performed in both online and offline moods. Up to now, around 234 million people speak Arabic and this society is known as Modern Standard Arabic (MSA). According to the similarity between Persian, Arabic and Urdu languages, we can help people (more than those mentioned above) by teaching them the computer networks. In this paper, besides reviewing previous and current methods of letters' recognition, an appropriate method for detecting letters is also presented.

2. Persian Letters Detecting Process

The process of detecting letters in continuous and discontinuous modes are shown in Figure1.

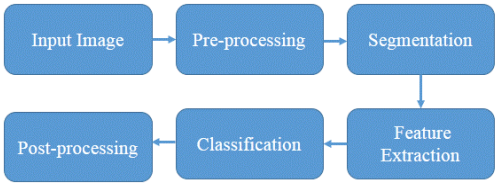


Figure 1. Letters detecting process

2.1. Getting photo of database

Hoda Handwritten numbers collection, the first Persian handwritten numbers collection, includes of 102353 black and white handwritten samples with the resolution of 200 dots per inch. Number of training samples is 6000 samples from each class and the number of experimental samples is 2000 samples from each class. Various methods were applied in this database to detect letters [1]. Tree decision method and genetic algorithm were present with the accuracy of 97.83 to 98.89 [2, 3]. Used classifiers were KNN[2], SVD[4], SVM [5] and the combination of classifiers [6]. Because of having details in letters, getting photo must be without slip and shadow.

2.2. Preprocess

In order to make a binary photo, conventional method was used. Sauvola, Otsu and Shafait methods [7,19] were simulated and the results showed the highest practicality of Shafait method in reducing the process time and in contrast improvement over the two other methods. Detecting gray shades will easily remove water markers and background from gray photos including black texts. On the other hand, detecting and removing water markers from scanned photos is not possible using this method due to black and white points. Thus, they can be removed by finding the letters frameworks [8]. In the following, tilt lines were removed in two steps. First, low-pass filter was used to stick letters and each line was saved separately. Then, using a ratio of length and width, line angle was calculated and rotated. In Figure 2, these lines are shown after angle correction [9].

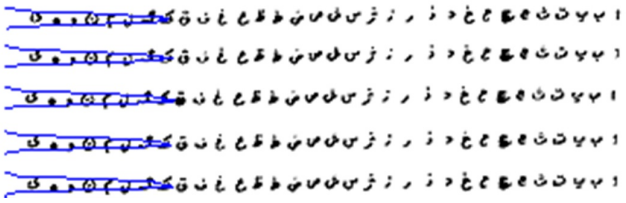


Figure2. Line Segmentation and Baseline Detectionin OCR

2.3. Segmentation

Segmentation characters include three parts. 1) Detecting lines from each other [10]. 2) Detecting words, 3) Detecting second part words.

3. Feature Extraction

In order to categorize letters, first we have to define the term extracting letters' characteristics. We can use categories to detect Persian handwritten letters. For this purpose, we have to explain features we used and then present techniques to extract these features. The main purpose of this part is to express those features which include Persian

letters and are considered as good norms for category. These features are classified into two groups and in each one a set of special features are investigated. These groups include Template Features (pattern) and Statistical Features.

3.1. Template Features (pattern)

In this section, features related to the writing pattern of each letter will be investigated. These features will be presented without a complex calculation and are based on the way of writing and natural form of letters which contain:

1) Number of sections
2) Position of sections
3) The kind of sections (sections detection)
4) The combination of letters (set of constituent characters' directions).

3.1.1. Number of sections

The majority of letters contain more than one section including point, diagonal line and etc. This feature helps us to separate letters from each other, especially for those letters with similar main body that have different extra sections in their natural forms. For example, by using this feature we can easily distinguish between "س" and its dotted counterpart "ش". In table 1, letters are categorized based on the number of sections.

Table1. Categories of letters based on the number of sections

No.	Number of sections	Letters
1	Single part	ا - ح - د - ر - س- ص - ع - ل- م - و - ه - ي
2	Bipartite	آ - ب - پ - ت - ث - ج - چ - خ - ذ - ز - ژ - ش - ض - ط - غ - ف - ق - ک - گ – ن
3	Three part	پ - ت - ث - چ - ژ - ش - ق - گ – ظ
4	Quartet	پ - ث - چ - ژ – ش

The repetition of letters in multiple sections is for variety of letter's sections writing. For example, we can write the extra section of letter" پ" in three ways: 1) Three separate points, 2) Two points with a point and 3) Three points together.

3.1.2. The position of second section (if it presents)

This feature represents the position of placing extra sections to the main body of written letter. This feature as above one is helpful for separating letters with similar body and different extra sections. For example, placing points above the letter "ث" and below the letter "پ" cause the separation of these two letters from each other. It is worthy to mention that this feature is only usable for at least bipartite letters. Table 2 categorizes letters according to position of extra sections.

Table2. Categorizing letters based on the position of extra sections

The position of extra section	letters
Above	آ-ت-ث-خ-ذ-ز-ژ- ش- ض - ظ - غ - ف - ق - ک - گ - ن
below	ب - پ - ج – چ

3.1.3. Types of sections (Detecting sections)

In Persian, extra sections belong to one of these categories. Such as points, diagonal line, handle and tilde. This feature (section type) make it possible to separate letters from each other and place them in different categories. Therefore, we will investigate these four forms:

Table 3. Categorizing letters based on the type of extra sections

No.	Section type	Letters
1	One-point	ب - ج - خ - ذ - ز - ض - ظ - غ - ف ــ ن
2	Two-points	ت ـ ق
3	Three-points	پ - ث - ژ - چ ـ ش
4	diagonal line	ک ــ گ
5	Handle	ط ـ ظ
6	Tilde	آ
7	No-point	ح- د-ر- س- ص- ع- ل- م- و- ه- ی

1) Points: Using points, we can separate letters from each other. In this way, they are divided into four types; No-point, one-point, two-point and three-point. This method helps us to distinguish significant numbers of letters.

2) Diagonal line: This feature helps us to separate two letters either from the whole letters or from each other. Letters "ک" and "گ" are distinguished in this way by using this method.

3) Handle: This feature like the previous one will cause the separation of two letters "ط" and "ظ" from each other.

4) Tilde: At last, tilde feature will cause the separation of "آ" from other letters. Table 3 shows the letters' category based on their sections.

3.1.4. The combinatory directions of letters (set of constituent letters)

Writing structure of each Persian letter include a set of different directions with unique arrangement. Considering this feature in vertical and horizontal axis, each letter can be placed in different category and each letter can be separated easily from each other. Figure 3 shows letter "ح" with the formed combinatory directions of it in X and Y axes and Figure 4 shows NRL feature extracted from 'د' letter.

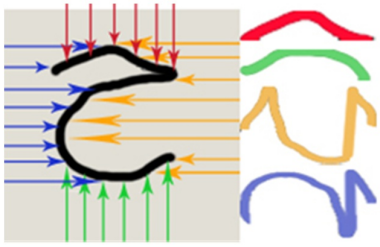


Figure3. The combinatory directions of "ح" letter

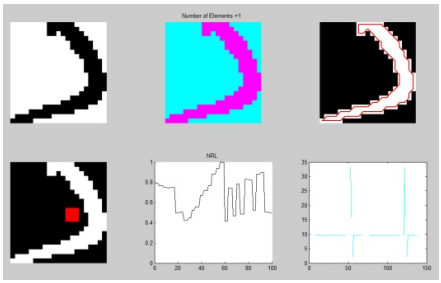


Figure4. Feature NRL from letter 'ر'

3.2. Statistical features

Those features which require complicated computing and are the results of letter's statistical information are called "statistical features". These statistical features are as below:

1. Zoning[11].
2. 17 Geometric properties [12, 19,20].

In this article, the starting point framework will be calculated and then, this point will be stick clockwise to its closest neighboring point. The movement direction from beginning to end is as a signal including 8 numbers. The rotation of the pen is shown in Figure 5.

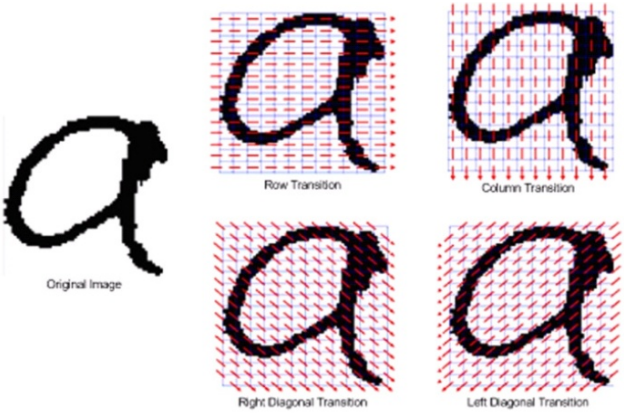


Figure5. Detecting the pen direction in writing a letter[14]

3.3. Feature reduction

In feature extraction, section's features (section's numbers, second section position (if it exists) , 40 Projection signal of the base section, statistical features (zoning and 17 geometric properties) ,NRL features and movement structural signal were calculated [15].

Emerging of high volume information, researchers focus on size and data dimensions' reduction in order to save both the time and the used memory[16].

Moreover, the number of requests for using different programs of system in limited speed and limited memory is increasing. In dimension reduction method, system removes the less important extracted features from the data collection sample. Special samples of these techniques are Principal Component Analysis (PCA) [17], Singular Value Decomposition (SVD), Random Projection (RP) and Mean-variance method.

PCA is a linear transform which has been widely used in data analysis and compression. The following presentation is adapted from [18]. Principal component analysis is based on the statistical representation of a random variable. Suppose we have a random vector population x, where shown in (1):

$$\underset{\sim}{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} \tag{1}$$

$$\begin{aligned} Y_1 &= \underset{\sim}{a}_1'\underset{\sim}{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ Y_2 &= \underset{\sim}{a}_2'\underset{\sim}{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ Y_p &= \underset{\sim}{a}_p'\underset{\sim}{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned} \tag{2}$$

$$\begin{aligned} &Var(Y_i) = \underset{\sim}{a}_i'\underset{\sim}{\Sigma}\underset{\sim}{a}_i,\ i = 1, \ldots, p \\ &Cov(Y_i, Y_k) = \underset{\sim}{a}_i'\underset{\sim}{\Sigma}\underset{\sim}{a}_k,\ i, k = 1, \ldots, p \end{aligned} \tag{3}$$

Y is linear output by (2) and Eigen values is calculated and replaced. In random vector X with covariance matrix S and eigenvalues $l_1 \geq l_2 \geq \cdots \geq l_p \geq 0$, the i^{th} principal component is and Eigen vectors given by (4) and (5).

$$Y_i = \underset{\sim}{e}_i'\underset{\sim}{X} = e_{i1}'X_1 + e_{i2}'X_2 + \cdots + e_{ip}'X_p,\ i = 1, \ldots, p \tag{4}$$

$$\sigma_{11} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i) \tag{5}$$

In base section feature part, three features exist. These features combined with PCA technique and the Bayes classifier components 1 to 3 are given. Besides, direction features will be classified by using SVM classifier in one against more. Results of these classifications will be discussed in combination section. The final decision will be reached based on the second section position.

3.4. Combining classification

Figure 6 presents process overview. This combination is obtained by Genetic Algorithm [13].The proposed method is shown in Figure 6. GA optimized combined structure. Figure 7 demonstrates GUI about OCR. GA optimizes accuracy and time consuming.

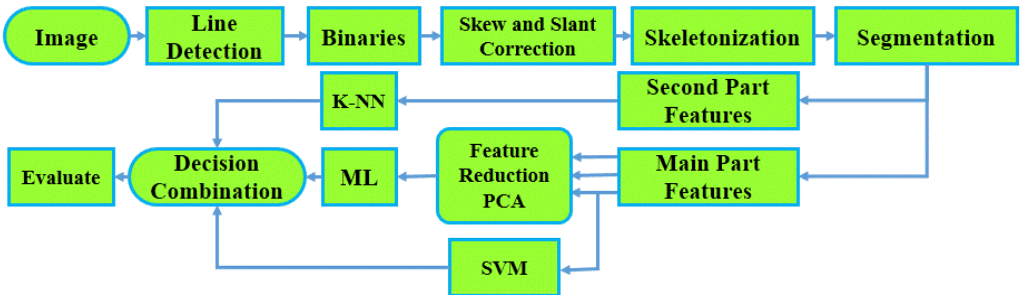


Figure6. The proposal flowchart

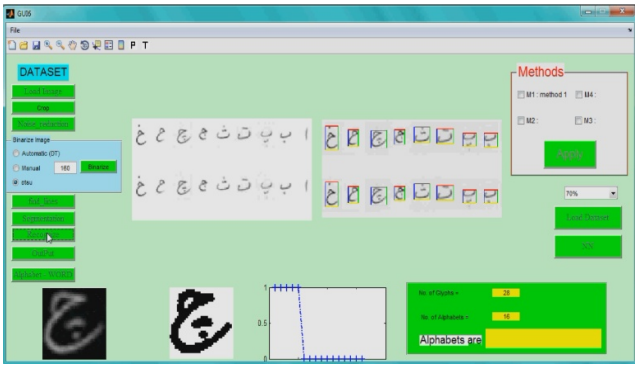


Figure 7. The GUI proposal program

Single letters in Persian language consist of 12 groups. The simulation is repeated 20 times and Simulation results of single letter are shown in Figure 8. Multiple part letters in Persian language consist of 20 groups. The Simulation results of single letter are shown in Figure 9.

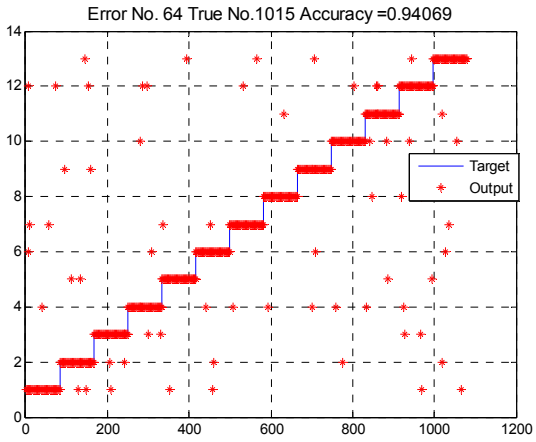


Figure8. Classification result in method of combining classifiers in single letters

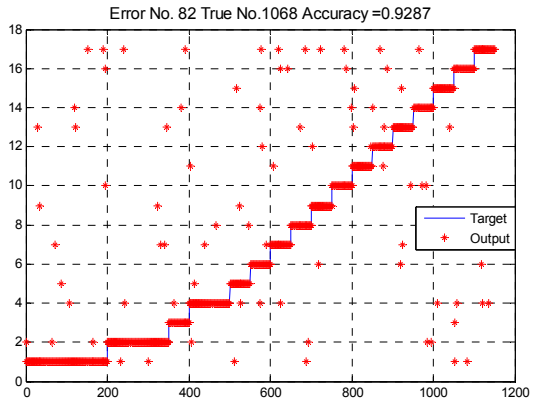


Figure9. Classification result in method of combining classifiers in multiple letters

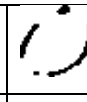
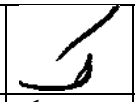
4. Error detection

Although the proposed system is reliable, it is confronted with some errors in some cases. In this section errors and their categorization is taken into consideration.

4.1. Main part confusing

This error appears in the main detection part such as some error samples. This part has many kind. Some of errors has shown in table 4.

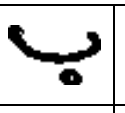
Table4. Error in detextion main part

					sample
ک	س	ن	ک	ا	Target label
ی	ص	ت	ت	ک	Goal label

4.2. Point confusing

This error appear between ‘ چ‘ , ‘ج’ / ’پ’,’ب’ / ’ز’,’ژ’ . Table 5 shows some error samples.

Table 5. Some of letter confuse in point classification

						Sample
ب	پ	ج	چ	ز	ژ	Target label
پ	ب	چ	ج	ژ	ز	Goal label

5. Conclusion

In this article, the combined method by the accuracy of 94 in single-part and 92 in multi-parts for recognizing Persian handwritten letters is presented. The proposed method is based on the body identification using geometric, directional and statistical features of SVM classifier. Therefore, the accuracy of subdivision features classification was increased.

Using GA, classifiers' combination directions were proposed based on the simulation, time-consuming and accuracy results. But the remarkable point is a result of distinctive groups' interference reduction degree. Source codes are available for reading and classifying by C++, Matlab software.

References

[1] H. Khosravi and E. Kabir, "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties," *Pattern Recognition Letters,* vol. 28, pp. 1133-1141, 2016.

[2] H. Parvin, *et al.*, "Divide & conquer classification and optimization by genetic algorithm," in *Convergence and Hybrid Information Technology, 2008. ICCIT'08. Third International Conference on*, 2015, pp. 858-863.

[3] H. Parvin, *et al.*, "A scalable method for improving the performance of classifiers in multiclass applications by pairwise classifiers and GA," in *Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on*, 2008, pp. 137-142.

[4] H. Salimi and D. Giveki, "Farsi/Arabic handwritten digit recognition based on ensemble of SVD classifiers and reliable multi-phase PSO combination rule," *International Journal on Document Analysis and Recognition (IJDAR),* vol. 16, pp. 371-386, 2013.

[5] A. Alaei, *et al.*, "Using modified contour features and SVM based classifier for the recognition of Persian/Arabic handwritten numerals," in *Advances in Pattern Recognition, 2009. ICAPR'09. Seventh International Conference on*, 2009, pp. 391-394.

[6] H. Parvin, *et al.*, "A new approach to improve the vote-based classifier selection," in *Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on*, 2008, pp. 91-95.

[7] K. Ntirogiannis, *et al.*, "A combined approach for the binarization of handwritten document images," *Pattern Recognition Letters,* vol. 35, pp. 3-15, 2014.

[8] W. Abu-Ain, *et al.*, "Skeletonization Algorithm for Binary Images," *Procedia Technology,* vol. 11, pp. 704-709, 2013.

[9] M. Feldbach and K. D. Tonnies, "Line detection and segmentation in historical church registers," in *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, 2001, pp. 743-747.

[10] L. Likforman-Sulem, *et al.*, "Text line segmentation of historical documents: a survey," *International Journal of Document Analysis and Recognition (IJDAR),* vol. 9, pp. 123-138, 2007.

[11] O. P. Sharma, *et al.*, "Recent trends and tools for feature extraction in OCR technology," *International Journal of Soft Computing and Engineering (IJSCE),* vol. 2, 2013.

[12] V. J. Dongre and V. H. Mankar, "Devnagari Handwritten Numeral Recognition using Geometric Features and Statistical combination Classifier," *arXiv preprint arXiv:1310.5619,* 2013.

[13] V. J. Dongre and V. H. Mankar, "A Review of Research on Devnagari Character Recognition," *International Journal of Computer Applications* vol. 12, pp. 8-15, 2010.

[14] T. Saba, *et al.*, "Improved statistical features for cursive character recognition," *Int J Innov Comput Inf Control IJICIC,* vol. 7, pp. 5211-5224, 2011.

[15] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, 1999, pp. 1150-1157.

[16] M. A. Shayegan and S. Aghabozorgi, "A new dataset size reduction approach for PCA-based classification in OCR application," *Mathematical Problems in Engineering,* vol. 2014, 2014.

[17] P. K. Sharma, *et al.*, "Radon Transform and PCA based feature extraction to design an Assamese Character Recognition system," in *Emerging Trends and Applications in Computer Science (NCETACS), 2012 3rd National Conference on*, 2012, pp. 46-51.

[18] I. Jolliffe, *Principal component analysis*: Wiley Online Library, 2005.

[19] AH Jadidinejad, H Sadr, *Improving weak queries using local cluster analysis as a preliminary framework*: Indian Journal of Science and Technology 8 (15), 2015.

[20] MN Soleimandarabi, SA Mirroshandel, A Novel Approach for Computing Semantic Relatedness of Geographic Terms: Indian Journal of Science and Technology 8 (27), 2015.