# ICDAR 2009 Handwritten Farsi/Arabic Character Recognition Competition

Saeed Mozaffari and Hadi Soltanizadeh
*Electrical and Computer Engineering Department, Semana University, Semnan, Iran.*
*mozaffari@semnan.ac.ir, h_soltanizadeh@kumesh.ac.ir*

## Abstract

*In recent years, the recognition of Farsi and Arabic handwriting is drawing increasing attention. This paper describes the result of the ICDAR 2009 competition for handwritten Farsi/Arabic character recognition. To evaluate the submitted systems, we used large datasets containing both binary and gray-scale images. Many different groups downloaded the training sets; however, finally 4 systems successfully participated in the competition. The systems were tested on two known databases and one unknown dataset. Due to the similarity between some digits and characters in Farsi and Arabic, each recognizer was tested for digit and character sets separately. For benchmarking, only the recognition rates, as the most important characteristic, are considered. Since participants used different software and even operating systems, the relative recognition speed is not compared in this competition.*

**Keywords**: OCR benchmarking, Performance evaluation, Farsi/Arabic languages, large database, isolated digits and characters.

## 1. Introduction

English, Chinese and Kanji handwritten character recognition have long been a focus of study and high recognition rates are reported. But little researches have been done on Farsi and Arabic in the last decade. This is a result of the lack of adequate support in terms of funding, and other utilities such as text database, dictionaries, etc [1].

Fortunately, many research groups around the world focused on Farsi/Arabic document analysis recently and promising results have been reported. However, there are not standard databases in Farsi/Arabic to be considered as a benchmark. Each of research groups implemented their system on set of data gathered by them and different recognition rates were reported. Therefore, it is very difficult to give comparative results for the proposed methods.

Despite many efforts, lack of communication among Farsi/Arabic OCR researchers caused wasteful duplication of efforts. The aim of ICDAR 2009 handwritten Farsi/Arabic character recognition competition is to bring together researchers working on this field. By benchmarking the state of the art Arabic character recognition techniques on large-scale dataset, a comparative result can be obtained. Furthermore, the result of this competition would have widespread benefits to other languages such as Farsi (Persian) and Urdu which have the same characters set.

## 2. A Short review of Farsi and Arabic Handwriting Characteristics

Since the characteristics of Farsi (Arabic) handwriting is different from the Latin one, and some of the readers maybe unfamiliar with these script, a brief description of the important aspects of Farsi/Arabic will be presented in this section. Farsi text is inherently cursive both in handwritten and printed forms and is written horizontally from right to left. Farsi writing is very similar to Arabic in terms of strokes and structure. Therefore, a Farsi word recognizer can also be used for recognition of Arabic words. The only difference between Farsi and Arabic scripts is in the character sets.

Farsi character set, comprises all of the 28 Arabic characters plus four additional ones (marked with the * in Table.2). A Farsi character is written as a single main stroke and in most cases is completed with other complementary strokes such as dot(s), zigzag bars, etc. The complementary strokes might be placed above, below, or in the middle of the main stroke. Some Farsi characters have a unique main stroke (overall shape); however they are distinguished from each other only by the presence/absence, position or number of some secondary strokes. An example of different characters with similar main stroke is shown in Fig. 1. Ambiguous writing of these secondary strokes sometimes causes a word image to be read in many various forms with completely different meanings. These secondary strokes make the recognition of Farsi/Arabic characters more difficult.

In contrast to English, Farsi characters are not divided into upper and lower case categories. Instead, a Farsi character might have several shapes depending on its relative position in a word. The shape of a

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.283

character should be changed if it is located at the beginning of the word, in the middle of the word, at the end of the word, and in isolation. An example is shown in Fig. 2. In this competition only the isolated form of each character is considered. Although Farsi has 32 characters, two of them can be written as two different styles in the isolated form (as shown in Fig.3). So, character sets are expanded into 34 classes in this competition.

Similar to Indian digits, Farsi language has ten digits. However, digits 4 and 6 can be written in two different shapes. Therefore, digits are considered as 12 classes (see Table.1).

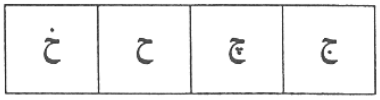


**Figure 1.** An example of different Farsi Characters with a unique overall shape.

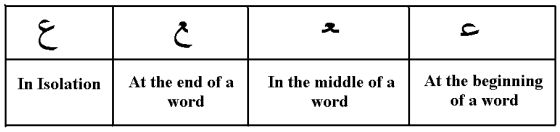



**Figure 2.** An example of different shapes of a Farsi Character.

**Figure 3.** Two characters with different isolated forms.

## 3. Database

Ten to 15 years ago, large databases were developed for handwritten Latin script recognition. For example, CEDAR database was released in 1994 includes images of city names, state names, ZIP codes and alphanumeric characters [2] or NIST database was developed for digit recognition [3]. Similar databases also exist for a few other languages such as Chinese and Indian [4,5,6].

Recently, for Arabic language, researchers have prepared some databases for handwritten texts [7], machine-printed documents [8], handwritten words [9] ,bank checks [10] and isolated handwritten digits and characters [11][12]. With the help of this new datasets, Farsi/Arabic OCR is going to be mature. In the following a short review of the databases used in this competition is presented.

### 3.1. Hoda database

Hoda database which contains handwritten digits is presented by Khosravi and Kabir in 2007 [13]. Binary images of 102,352 digits were extracted from about 12,000 registration forms of two types, filled by B.Sc. and senior high school students. These forms were scanned at 200 dpi with a high speed scanner. Fig. 4 shows sample forms for data gathering in this database.

Each digit's bounding box was extracted automatically. They deleted abnormal samples by a manual refinement procedure. For better comparison, Hoda database was partitioned into train (60,000 samples) and test (20,000 samples) subsets. The remaining samples (22,353 samples) can be used for verification. Although Hoda database was gathered from educated writers, variety of participants in data collecting process can be regarded as its power point.

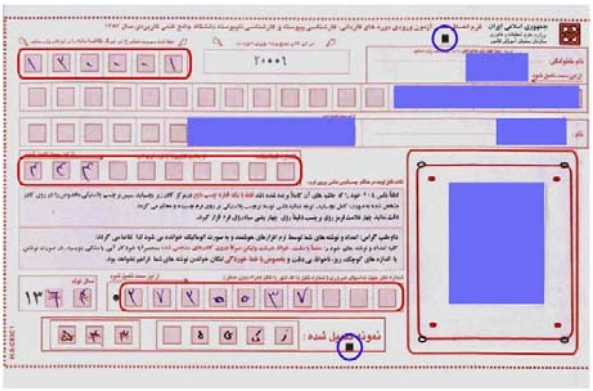


**Figure 4.** Sample form for data collection in Hoda database [13].

### 3.2. Farsi CENPARMI database

In 2006, the Center of Pattern Recognition and Machine Intelligence (CENPARMI) of Concordia University have presented six databases consisting of handwritten Farsi numerical strings, digits, letters, legal amounts and dates [12]. Fig.5 shows a filled form for data collection in Farsi CENPARMI database.

All samples were presented in both binary and gray-scale formats. We only used binary forms of handwritten digits and characters. Digits set is composed of 11,000 training and 5,000 test samples while characters set includes 7,140 training and 3,400 test samples.

The data entry forms were filled by 175 writers selected from different ages, genders, and jobs. The main drawback of Farsi CENPARMI is limited writers.

**Figure 5.** A filled form in Farsi CENPARMI database [12].

### 3.3. Extended IFHCDB database

The IFHCDB database was released by Amirkabir University in 2006 [11]. It includes 52,380 isolated characters and 17,740 numerals gathered from Iranian high school and guidance school entrance exam forms during the years 2004-2006. Fig. 6 shows a form in IFHCDB database.

This overcomes the subject-bias problems of other databases that were scanned in laboratory settings. The data were also scanned at 300 dpi in 8-bit grayscale IFHCDB database is a non-uniform dataset in which the distribution of samples in each class was inspired from a very large set consisting of 10,236,040 samples.

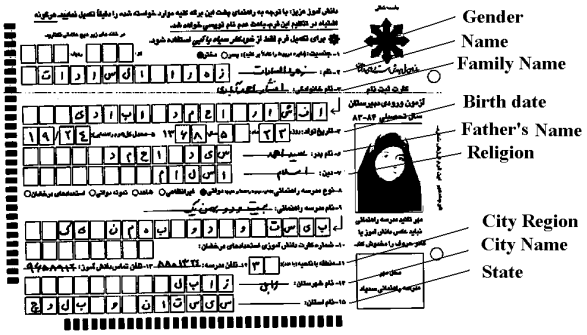



**Figure 6.** A sample form in Farsi CENPARMI database [11].

### 3.4. Training and Test samples distribution

To evaluate Farsi/Arabic handwritten digits and characters recognizers, we combined training and test set samples in Hoda, Farsi CENPARMI, and extended IFHCDB databases. In ICDAR 2009 competition, we used Hoda's training and test sets as a part of our training and test sets respectively. Although Farsi CENPARMI database consists different types of data, we only used its characters set. Likewise Hoda database, training and test datasets are used for training and testing correspondingly.

Fig.7 shows some samples in the database. Table.1 compares the specifications of ICDAR 2009 database and some other datasets.

Tables.1 and 2 show training and test distributions. Unlike characters set, digits have a uniform distribution.

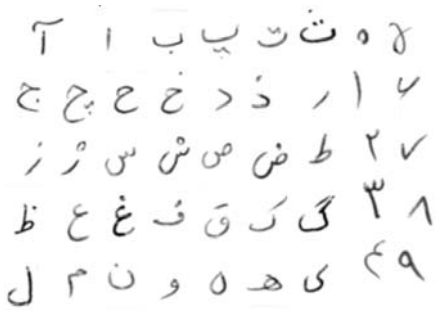


**Figure 7.** Some samples in Farsi/Arabic ICDAR 2009 database.

**Table 1.** Number of training and test digit samples.

| Digit | | Number of Training Samples | Number of Test Samples |
|---|---|---|---|
| Zero | ۰ | 2500 | 4000 |
| One | ۱ | 2500 | 4000 |
| Two | ۲ | 2500 | 4000 |
| Three | ۳ | 2500 | 4000 |
| Four | ۴ | 2500 | 4000 |
| | ٤ | 2500 | 4000 |
| Five | ۵ | 2500 | 4000 |
| Six | ۶ | 2500 | 4000 |
| | ٦ | 2500 | 4000 |
| Seven | ۷ | 2500 | 4000 |
| Eight | ۸ | 2500 | 4000 |
| Nine | ۹ | 2500 | 4000 |
| Total | | 30000 | 48000 |

## 4. Submitted Systems

This section gives a brief description of the submitted systems to the competition. Each system description has been provided by the system's authors

and summarized by the competition organizers. The descriptions vary in length due to the level of detail in the information provided by the system's developers.

### 4.1. CEMATER-JU

Nibaran Das from department of Computer Science and Engineering, Jadavpur University, Kolkata, India, presented a system based on longest-run and Convex-hull features and Multilayer Perceptron (MLP) neural network.

Within a rectangular image region of a character, a longest run feature are computed in four directions (row wise, column wise and along the directions of two major diagonals). The longest run feature is computed by considering the sum of the lengths of the longest bars that fit consecutive black pixels along each direction. Then, each of the longest run feature values is normalized by the entire image's size. Different features are extracted based on different bays attributes of the convex hull of the character patterns. Then these patterns are divided into four sub-images with respect to the centroid of the convex hull boundary. From each such sub-image some other bays features are also calculated.

### 4.2. MDLSTM

Alex Graves from Technical University of Munchen, Germany, developed a system based on a hierarchy of multidimensional recurrent neural networks. It uses the multidimensional Long Short-Term Memory network architecture. This system works directly on raw input data with no preprocessing and feature extraction. This system can be used for online and offline handwriting data recognition.

For this competition, the system uses sequence classification approach. Sequence classification is a simple extension of the standard softmax/ cross-entropy output layer used for multilayer Perceptron classification. With this layer the network outputs are summed over the entire data sequence, and the softmax function is used to normalize the summations, yielding a probability distribution over the possible classes. The training is then carried out with cross-entropy and back propagation as usual.

### 4.3. REGIM

The REGIM system developed by Mohamed Ben Halima and Adel M. Alimi from Research Group on Intelligent Machines (REGIM), University of Sfax in Tunisia. In the preprocessing step, binarization, noise reduction, thinning, and size normalization steps were performed. Both statistical and structural features were used with K-nearest neighbor classifier for Farsi/Arabic handwritten character recognition.

### 4.4. ECA

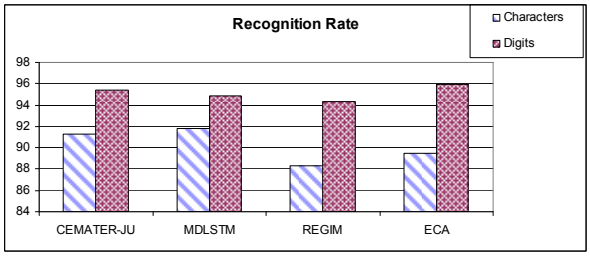
Mohamad nahvi and kourosh kiani from Electrical and Computer Department in Semnan University submitted a system based on gradient operator in 16 directions and DCT. They used a three layers multilayer Perceptron neural network as a recognizer.

**Table 2.** Number of training and test character samples.

| Character | | Number of Training Samples | Number of Test Samples |
|---|---|---|---|
| Alef | آ | 1260 | 1209 |
| | ا | 9183 | 9634 |
| Be | ب | 2765 | 2935 |
| Pe * | پ | 1471 | 1393 |
| Te | ت | 3580 | 3775 |
| Se | ث | 1335 | 1186 |
| Jim | ج | 2065 | 2270 |
| Che * | چ | 1284 | 1272 |
| He | ح | 2800 | 2830 |
| Khe | خ | 1715 | 1947 |
| Dal | د | 4800 | 4930 |
| Zal | ذ | 1199 | 1203 |
| Re | ر | 7188 | 6925 |
| Ze | ز | 2956 | 3040 |
| Zhe * | ژ | 1295 | 1004 |
| Sin | س | 5749 | 5735 |
| Shin | ش | 2089 | 1931 |
| Sad | ص | 1734 | 1861 |
| Zad | ض | 1679 | 1605 |
| Ta | ط | 1467 | 1339 |
| Za | ظ | 1230 | 1138 |
| Ayn | ع | 2345 | 2445 |
| Ghayn | غ | 1451 | 1692 |
| Fe | ف | 2100 | 2235 |
| Ghaf | ق | 1609 | 1605 |
| Kaf | ک | 2222 | 2375 |
| Gaf * | گ | 1470 | 1535 |
| Lam | ل | 5760 | 5910 |
| Mim | م | 7735 | 7630 |
| Noon | ن | 6755 | 7091 |
| Vav | و | 3080 | 3180 |
| He | ه | 4235 | 4160 |
| | ھ | 1645 | 1767 |
| Ye | ی | 6930 | 7205 |
| Total | | 106181 | 107992 |

## 5. Results

We evaluated the performance of 4 different Farsi/Arabic handwritten character recognition systems based on their recognition rates. Fig.9 shows the average performance of the participating systems on the test set. It can be seen that all recognition systems have very close recognition rates. The overall character's recognition rate is less than digits one. This is mostly due to the larger number of classes (34 classes of characters Vs. 12 classes of digits). Similarity between characters, especially those with secondary strokes, is another reason for performance degradation.




## 6. Conclusion

The aim of ICDAR 2009 handwritten Farsi/Arabic character recognition competition is to evaluate different systems. The competition results show a remarkable progress in Farsi/Arabic recognition systems. However, the obtained results are still less than English character recognition systems. System ECA was the winner of the digits recognition part with %95.9 recognition rate. In the character recognition competition, system MDLSTM reached %91.85 accuracy and outperforms the other systems.

For the future work we will consider other criteria for performance evaluation such as recognition time and reject ratio.

### References


[1] Amin,A, " Off-line Arabic characters Recognition: The State Of the Art", *Pattern Recognition* 31(5), 517-530, 1998.

[2] J.J.Hull, " A database for Handwritten Text Rcognition Research*", IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol.16, pp.550-554,1994.

[3] Y.LeCun,L.Bottou,Y.Bengio, and .Haffiner,"Gradient based learning applied to document recognition", Proceedings of IEEE, Vol.86(11),pp.2278-2324,1998.

[4] T.Saito, H.Yamada and K.Yamomoto. " One database ELT9 of handprinted characters in JIS Chinese characters and its analysis (in Japanese) " Transaction of IECEJ, Vol.J.68-D(4),pp.757-764,1985.

[5] U.Bhattacharya and B.B.Chaudhuri," Databases for Research on Recognition of Handwritten Characters of Indian Scripts" Internation Conference of Document Analysis and Recognition, pp.789-793,2005.

[6] S.Mihov,K.U.Schulz and et al ," A Corpus for Comparative Evaluation of OCR Software and Postcorrection Techniques" Proceeding of International Conference of Document Analysis and Recognition, pp.162-166, 2005.

[7] S.Al-Ma'adeed,D.Ellimam and C.A Higgins," A data Base for Arabic Handwritten Text Recognition Research", Proceeding of Eighth International Workshop on Frontiers in Handwriting Recognition,pp.485-489,2002.

[8] http://documents.cfar.umd.edu/resources/database/ERIM_Arabic_DB.html

[9] M.Pechwitz and V.Märgner,"HMM Based Approach for Handwritten Arabic Word Reognition using the IFN/ENIT-Database," Proceeding of International Conference of Document Analysis and Recognition ,pp.890-894,2003.

[10] Y.Al-Ohali,M.Cheriet,C.Suen," Databases for recognition of handwritten Arabic checks", Pattern Recognition, Vol.36,pp.111-121,2003.

[11] S.Mozaffari,K.Faez,F.Faradji,M.Ziaratban,S.M.Golzan,Acomprehensive isolated Farsi/Arabic character database for handwritten OCR research, in: Proceedings of the 10th International Workshop on Frontiers of Handwriting Recognition, LaBaule,France,2006,pp.385–389.

[12] F.Solimanpour, J.Sadri, C.Y.Suen, Standard databases for recognition of handwritten digits, numerical strings, legalamounts, letters and dates in Farsi language, in: Proceedings of the 10th International Workshop on Frontiers of Handwriting Recognition, ,LaBaule, France, 2006,pp.3–7.

[13] H.Khosravi and E.Kabir, Introducing a very large dataset of handwritten Farsi digits and a study on their varieties, Pattern Recognition Letters 28 (2007) 1133–1141.