A Comprehensive Isolated Farsi/Arabic Character Database for Handwritten OCR Research

Saeed Mozaffari, Karim Faez, Farhad Faradji, Majid Ziaratban, S. Mohamad Golzan

A Comprehensive Isolated Farsi/Arabic Character Database for Handwritten OCR Research

Saeed Mozaffari, Karim Faez, Farhad Faradji, Majid Ziaratban and S.Mohamad Golzan
Pattern Recognition and Image Processing Laboratory
Electrical Engineering Department, Amirkabir University of Technology, Tehran, Iran,15914.
{ s_mozaffari , kfaez, f_faradji,m_ziaratban and smgolzan }@aut.ac.ir

Abstract

This paper presents a new comprehensive database for isolated offline handwritten Farsi/Arabic numbers and characters for use in optical character recognition research. The database is freely available for academic use. So far no such a freely database in Farsi language is available. Grayscale images of 52,380 characters and 17,740 numerals are included. Each image was scanned from Iranian school entrance exam forms during the years 2004-2006 at 300 dpi. The only restriction imposed on the writers is to write each character within a rectangular box. The number of samples in each class of the database is non-uniform corresponding to their real life distributions. Also, for comparison purposes, each dataset has been properly divided into respective training and test sets.

Keywords: OCR, Farsi/Arabic, Comparative database, offline, isolated numbers and characters.

1. Introduction

During the past decade, remarkable progress has been achieved in the filed of handwritten alphanumeric, words and scripts recognition and many practical applications of character recognition systems such as automatic reading of postal addresses, bank checks and forms have been emerged.

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. However, although almost a third of a billion people worldwide, in several different languages, use Farsi and Arabic characters for writing, little research progress, in both on-line and off-line has been achieved towards the automatic recognition of these languages. This is a result of the lack of adequate support in terms of funding, and other utilities such as text database, dictionaries, etc [1].

Some previous works on recognition of isolated characters, words and scripts of Farsi and Arabic language have been proposed, including different features and different classifiers such as structural methods [2,3,4], statistical features [5,6], neural networks [5,6,7] and Support Vector Machines (SVM) [8,9,10].

It is notable that most of the above work was done on isolated characters or assuming that the Farsi (Arabic) handwritten word is already segmented into separated characters before recognition.

Unfortunately, there are not standard databases in Farsi/Arabic to be considered as a benchmark (such as NIST data set for English Digits). 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. Among them, the proposed method in [8] reached the recognition rate of 99.57%. While the recognition rates achieved by others were less than 95% [4,5,6].

To validate the effectiveness of a proposed system for Farsi (Arabic) OCR research, it is necessary to compare it with other approaches. Now, such comparison is possible by implementing the concurrent approaches concurrently and then applying them with the proposed method on the same database. Therefore, in the filed of Farsi (Arabic) OCR, a standard database is needed to facilitate researches.

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 [11] or NIST database was developed for digit recognition [12]. Similar databases also exist for a few other languages such as Chinese and Indian [13,14,15].

Recently, for Arabic language, researchers have prepared some databases for handwritten texts [16], machine-printed documents [17], handwritten words [18] and bank checks [19].

This paper presents IFHCDB (Isolated Farsi Handwritten Character Database), new comprehensive database for isolated offline handwritten Farsi (Arabic) numbers and characters, for the use in optical character recognition research. This database also can be used for recognition of other languages such as Urdu, Kurd and Arabic, which use the same characters in writing.

2. Current Arabic Databases

This section presents an overview of current Arabic databases in more details and makes a comparison of their size, contents and data gathering methods.

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 approximately 5,000 city names, 5,000 state names, 10,000 ZIP codes and 50,000 alphanumeric characters [11]. Recently released databases for Arabic handwritten recognition have similar size and scope.

In 1999 Al-ISRA database was released by a group of researchers at University of British Columbia in Canada. It contains 37,000 Arabic words, 10,000 digits (in two types, "Mashreqi" and "Magharibi"), 2,500 signatures, and 500 free-form Arabic sentences gathered from five hundred randomly selected students at Al-Isra University in Amman, Jordan [20].

"Indian digits" also called "Mashreqi" are the numeric digits normally used in Arabic writing in the Middle Eastern countries, as opposed to "Arabic numerals" ("Magharibi") used in Latin scripts at the North of Africa.

Alma'adeed et al. presented the AHDB, a database of samples from 100 different writers, including words used for numbers and in bank checks in 2002 [16]. It also contains the most popular words in Arabic writing and free handwriting pages on any topic of the writer's choosing.

In 2003, Center for Pattern Recognition and Machine Intelligence (CENPARMI) in Montreal developed databases of images from 3,000 checks provided by a banking corporation. These databases are subwords, Indian digits, legal amounts (numeric amounts written in words), and courtesy amounts (numeric amounts written with Indian digits) [19]. The subwords database contains 29,498 samples, the Indian digits database 15,175, and the legal and courtesy databases 2,499 each.

The IFN/ENIT dataset was created by the Institute of Communications Technology (IFN) at the University of Braunschweig in Germany and the Ecole Nationale d'Ing'enieurs de Tunis (ENIT) in Tunisia. It consists of 26,459 images of the 937 names of cities and towns in Tunisia, written by 411 different writers. The images are partitioned into four sets so that researchers can use and discuss training and testing data in this context [18].

3. Farsi Handwriting Characteristics

Since the characteristics of Farsi (Arabic) handwriting is different from the Latin one and some of the readers maybe unfamiliar with Farsi script, a brief description of the important aspects of Farsi script will be presented. 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, shown in Fig. 1, comprises all of the 28 Arabic characters plus four additional ones (marked with the * in Fig. 1).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. 2. Ambiguous writing of these secondary strokes sometimes causes a word image to be read in many various forms with completely different meanings.

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 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. 3.

In Farsi language, there are ten digits (similar to Indian digits) that are shown in Fig.4. Digits (4 and 6) can be written in two different shapes.

4. Data Collection

The database discussed here is a subset of a large database with 10,236,040 images of Farsi/Arabic isolated handwritten alphanumeric gathered as part of a research project sponsored by Iran National Information and Communication Technology (ICT).

Characters and numerals were extracted from digital images of Iranian high school and guidance school entrance exam forms (Fig.5) during the years 2004-2006. Each form consists of different fields such as Name, Family, Father Name, Religion, City Name, Average, Birth date and City part. The only restriction imposed on the writers is to write each character within a rectangular box in the appropriate field.

Each form was scanned at 300 dpi and stored as a grayscale image. A software has been developed to identify and extract each entity by detecting horizontal and vertical thick lines. After finding each character bounding box, the image is stored as a 77×95 BMP image. In some cases a character may touches or crosses the horizontal or vertical lines of the bounding box. Therefore two types of errors may happen. In the major error case, some character's dots or some complementary strokes of it were omitted and the result was not distinguishable. But in the minor error case, usually the last part of the character was missed.

An evaluation process has been done on 133,529 randomly selected images of the database, 0.42% and 0.08% error were observed for major and minor error cases respectively.

5. Data Storing

The IFHCDB database is divided into explicit training and testing sets (70% training and 30% testing) to facilitate the sharing of results among researchers as well as performance comparisons.

16	15	14*	13	12	11	10	9	8	7*	6	5	4	3*	2	1
ش	س	ژ	ز	ر	ذ	د	خ	ح	چ	ج	ث	ت	پ	ب	ا
32	31	30	29	28	27	26*	25	24	23	22	21	20	19	18	17
ی	ه	و	ن	م	ل	گ	ک	ق	ف	غ	ع	ظ	ط	ض	ص

Figure 1. Isolated Farsi character set.

Each image was saved with a name indicates its belonging set (training or testing), its class and its sampling number in that class. The first two characters indicate if this sample belongs to train (tr) or test set (ts) followed by 3 digits for class number. The last 5 digits in the image name show sample number. For example the 145th training sample of the 25th class is saved as tr025000145.bmp.

خ	ح	چ	ج

Figure 2. An example of different Farsi Characters with a unique overall shape.

ع	ع	ـعـ	ـع
In Isolation	At the end of a word	In the middle of a word	At the beginning of a word

Figure 3. An example of different shapes of a Farsi Character.

0	1	2	3	4		5	6		7	8	9
۰	۱	۲	۳	٤	۴	٥	٦	۶	۷	۸	۹

Figure 4. Digits in Farsi and English.

6. Database Size

The IFHCDB database includes 52,380 isolated characters and 17,740 numerals gathered from real-life documents. Fig.6 and Fig.7 show some digit and character samples in the database respectively.

In a real-life, numbers of samples in each class are not the same. For example frequency of character " آ " is more than character " ث ".

IFHCDB database is a non-uniform data set in which the distribution of samples in each class is the same as ICT report based on 10,236,040 samples.

The distribution of samples in 32 character and 10 digit classes are given in Table.1 and Table.2 respectively.

Figure 5. An entrance exam form for data collection.

7. Meta Information

Detailed information for each image file was stored in a TXT file with 10 attributes that cover all kinds of useful meta data. The attributes are: the file name "*Name*", the unique identifier of the image file "*Charname*", the assigned class of the image "*Class*" (a number between 1 and 47), the writer's gender "*Gender*" (male="پ", female="د"), the city of the form writer "*City*" , the field to which the character belongs "*Extracted Field*" (Name, Family, Father's Name, Religion, State, City Name, Average, Birth Date, City Part), the index of the character/digit in the word "*CharNumber*" (characters are counted from right and digits from left), the type of scanner "*Scanner*", the form writer's index "*CharWriterIndex*", the state of membership in train or test set "*TrainOrTest*".

Fig.8 shows TXT ground truth file for an image in the database.

Figure 6. Handwritten Farsi digit samples.

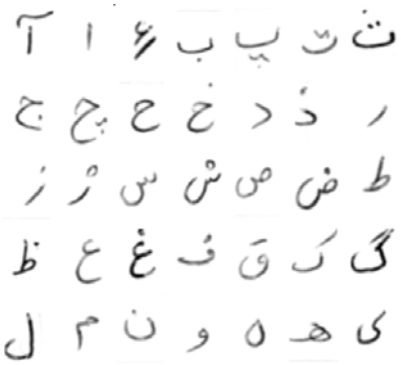

Figure 7. Handwritten Farsi character samples.

8. Conclusion

This paper presents a new comprehensive database for isolated offline handwritten Farsi/Arabic numbers and characters for use in optical character recognition research. IFHCDB consists of 52,380 gray scale images of handwritten characters and 17,740 numerals.

Since the dataset has been properly divided into respective training and test sets, it can be used for comparison purposes.

The non-uniform IFHCDB database includes real-life data collected from Iranian high school and guidance school entrance exam forms during the years 2004-2006. 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. This allows for experimentation with preprocessing and grayscale recognition techniques.

This database is freely available for academic use at http://ele.aut.ac.ir/ imageproc/downloads/IFHCDB.html.

Table.1 Distribution of characters in IFHCDB

Character		Number	Percent
Alef	آ	70	0.134
	ا	10000	19.091
	ء	70	0.134
Be	ب	1050	2.005
Pe *	پ	210	0.401
Te	ت	1420	2.711
Se	ث	40	0.076
Jim	ج	560	1.069
Che *	چ	50	0.095
He	ح	1230	2.348
Khe	خ	360	0.687
Dal	د	2660	5.078
Zal	ذ	40	0.076
Re	ر	4400	8.400
Ze	ز	1115	2.129
Zhe *	ژ	50	0.095
Sin	س	3400	6.491
Shin	ش	790	1.508
Sad	ص	470	0.897
Zad	ض	360	0.687
Ta	ط	200	0.382
Za	ظ	50	0.095
Ayn	ع	870	1.661
Ghayn	غ	140	0.267
Fe	ف	915	1.747
Ghaf	ق	470	0.897
Kaf	ک	640	1.222
Gaf *	گ	260	0.496
Lam	ل	3000	5.727
Mim	م	4970	9.488
Noon	ن	3830	7.312
Vav	و	1360	2.596
He	ه	2530	4.830
	ه ـه	280	0.535
Ye	ی	4520	8.629
Total		52380	100

Table.2 Distribution of characters in IFHCDB

Digit		Number	Percent
Zero	۰	1440	8.117
One	۱	5000	28.185
Two	۲	3560	20.068
Three	۳	854	4.814
Four	۴	576	3.247
	٤	75	0.423
Five	۵	600	3.382
Six	۶	460	2.593
	٦	105	0.592
Seven	۷	2380	13.416
Eight	۸	500	2.818
Nine	۹	2190	12.345
Total		17740	100

<z: row
Name=" tr02500145"
Charname="01-F000101001001065974-City1"
Class="01"
Gender="د"
City="ارالک"
ExtractedField="CityName"
CharNumber="1"
Scanner="HP 7450"
CharWriterIndex="RF000101001"
TrainOrTest="Train" />

Figure 8. Ground truth data.

Acknowledgements: The authors wish to acknowledge the support of Iran National Information and Communication Technology (ICT) for helping of generation this database.

References

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

[2] Abuhiba,I.S.I , Mahmoud,S.A and Green,R.G" Recognition of handwritten cursive Arabic characters". *IEEE Trans on Pattern Analysis and Machine Intelligence*.Vol.16,No.6, June 1994,PP.664-672.

[3] Dehghan,M. and Faez,K :Farsi Handwritten Character Recognition with moment invariants, Proceedings of 13th International Conference on Digital Signal Processing, Vol. 2,1997, pp.507-510.

[4] Mozaffari,S. , Faez,K. and Ziaratban,M." A Hybrid Structural/Statistical Classifier for Handwritten Farsi/Arabic Numeral Recognition" *MVA2005 IAPR Conference on Machine VIsion Applications*, May 16-18,2005 Japan .pp 218-211.

[5] Mowlaei,A. , Faez,K. and Haghighat,A.T "Feature Extraction with Wavelet Transform for Recognition of Isolated Handwritten Farsi/Arabic Characters and Numerals" , *Proceedings of 14th International Conference on Digital Signal Processing*, Vol. 2, July 2002, pp. 923-926.

[6] Mozaffari,S. , Faez,K. and Rashidy Kanan,H" Recognition of Isolated Handwritten Farsi/Arabic Alphanumeric Using Fractal Codes" *IEEE Proceeding of Southwest Symposium on Image Analysis and Interpretation.(SSIAI)* , 2004,pp.104-108.

[7] Razavi,M. and Kabir,E. " On-line isolated Farsi character recognition using neural network". *The 3rd Conference on Machine Vision, Image Processing and Applications (MVIP)*,Tehran,Iran, 2005.PP 83-89.

[8] Soltanzadeh,H. and Rahmati,M " Recognition of Persian handwritten digits using image profiles of multiple orientations". *Pattern Recognition Letters* 25(2004) pp. 1569-1576.

[9] Mowlaei,A. and Faez,K " Recognition of Isolated Handwritten Persian/Arabic Characters and Numerals Using Support Vector Machines", *IEEE Int. Workshop on Neural Networks for Signal Processing (NNSP)*, 2003, P.P. 547-554.

[10] Mozaffari,S. , Faez,K. and Rashidy Kanan,H " Feature Comparison between Fractal codes and Wavelet Transform in Handwritten Alphanumeric Recognition Using SVM Classifier". *International Conference on Patter Recognition (ICPR),* Cambridge UK, August 2004, Volume 2, pp.331-334.

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] Nawwaf Kharma, Maher Ahmed and Rabab Ward," A New Comperehensive Database of Hand-written Arabic Words, Numbers, and Signatures used for OCR Testing".