Standard Databases for Recognition of Handwritten Digits, Numerical Strings, Legal Amounts, Letters and Dates in Farsi Language

Farshid Solimanpour, Javad Sadri, Ching Y. Suen

Standard Databases for Recognition of Handwritten Digits, Numerical Strings, Legal Amounts, Letters and Dates in Farsi Language

Farshid Solimanpour[1] *Javad Sadri*[2] *Ching Y. Suen*

CENPARMI (Center for Pattern Recognition and Machine Intelligence), Computer Science Department, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, Quebec, Canada, H3G 1M8, Tel: (514)-848-2424-Ext:7950, Fax: (514)-848-2830

Emails:{f_solima, j_sadri, suen}@cs.concordia.ca

Abstract

This paper describes an important step towards the standardization of the research on Optical Character Recognition (OCR) in Farsi language. It describes formations of novel and standard handwritten databases including isolated digits, letters, numerical strings, Legal amounts (used for cheques), and dates. Despite conventional research and an Internet search, no publicly accessible Farsi database was found. Hence, it was decided that it would be a worthwhile academic effort to create several Farsi databases that could stand on their own merit functioning as useful tools for OCR researchers. Also, in order to show the potential uses of our new databases we also conducted some experiments on the recognition of handwritten isolated Farsi digits.

Keywords: Farsi OCR, Farsi Handwritten Databases, Arabic Handwritten Databases, Indian Digits Database.

1. Introduction

An essential part of the development and evaluation of every offline character recognition technique is the comparison of the results by using the same standard database as other researchers [1]. There are many examples of widely used databases in the field of handwriting recognition such as NIST [2], CEDAR [3], CENPARMI [4], UNIPEN [5], CENPARMI Arabic Cheques [6], ETL9 (Japan) [7], and PE92 (Korea) [8]. But to the best of our knowledge, no standard database for the Farsi language is available.

The Farsi language is spoken by more than 110 million people, mainly in Iran, Afghanistan, Tajikistan, and partly in some other countries. There are also other languages which use the same alphabets and digits or subsets of them such as: Arabic, Urdu, and Pashto.

In Farsi, words, sentences and dates are written from right to left, but numbers are written from left to right which match the style of writing numbers in the English language. Farsi has 32 letters in the alphabet and is a cursive language, which means within one word, letters can be connected. Due to connectivity, the shape of Farsi letters may change significantly depending on their positions in a word, the identity of neighboring letters, the font, or the way that writer connects successive letters. Considering these facts, it is crucial to have standard databases in order to improve research on Farsi handwritten recognition.

In this paper, we will describe the details of formation of the following databases: Farsi isolated digits, numerical strings, isolated letters, legal amounts, Farsi dates (called Hijri Shamsi); and a small set of English digits (written by Farsi native speakers). In order to show the usefulness of our database, we also report the results of some of our experiments on the recognition of isolated handwritten Farsi digits taken from this database.

The rest of this paper is organized as follows: Section 2 describes our steps towards collecting the data. In Section 3, data extraction methods are covered, which include the pre-processing of the images. Section 4 details our experiments on the recognition of Farsi isolated digits. In Section 5, we discuss the output of our work and compare it with some other works. Finally in Section 6 we present some concluding remarks and suggestions for future research.

2. Data Collection

Two data entry forms were designed for our data collection process. The first form contained Farsi numerical strings, isolated letters, the date, and English digits. The Farsi digits database was formed by segmenting the numerical strings in this form. The second form was completely dedicated to cursive legal amounts. In order to automate the process of cutting the fields out of the scanned forms, two types of anchoring marks were added to the forms: the form identifiers, and the edge identifiers. The form identifiers consisted of 8 squares such that each one can have two states: empty or blackened. Therefore, they could represent 255 binary numbers and could serve as identity of 255 different forms. In our case, for the form 1, squares 1, 5, and 8; and for the form 2, squares 2, 4, and 7 were blackened. By detecting these squares our program could automatically identify the form it was working on. Edge identifier marks consisted of four squares located at each corner of the form, and detecting them enabled the program to correctly determine the coordinates of the region that contained the actual data. Two samples of the data entry forms are shown in Figure 1 and Figure 2.

[1 & 2] Authors have the same contribution

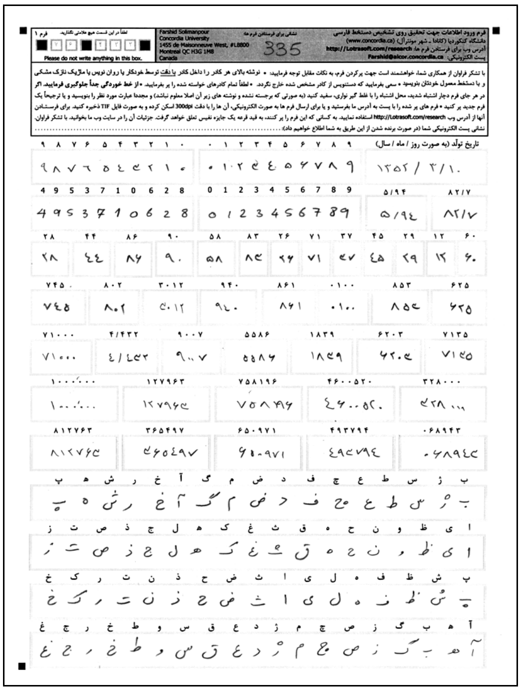

Figure 1. Sample of a filled form 1.

The data entry forms were filled by 175 writers selected from different ages, genders, and jobs; and among those, 105 writer were randomly assigned to our training set, 50 writer to the testing set, and 20 writer to the verifying set. We ensured that the data in each set was completely genuine and that there would be no relation between sets. Our final work includes these databases: numerical strings, isolated digits, Farsi letters, cursive legal amounts, and a small set of English isolated digits. In the following subsection we give details on each database.

Figure 2. Sample of a filled form 2.

2.1. Farsi numerical strings database

Each participant wrote 42 numerical strings in form 1 which were used to form our database of Farsi numerical strings. In Farsi, the normal height of the numeral "0" is approximately one fifth of other characters, and is written differently every time either because of its location in a numerical string or because of its repetition in a numerical string. To cover all forms, we had to repeat it more times than other numerals. In our databases, we have samples of the numeral zero being at the beginning, middle or end of a numeral string as well as when it is repeated two, three or six times in a string. In Figure 3, samples of two different writing styles of repeated zeros can be viewed.

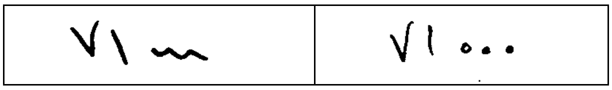

Figure 3. Different styles of writing zeros in the numerical string: 71000.

Table 1. Statistics of numerical strings database.

Classes	Writers	Training Set	Verifying Set	Testing Set
42	**175**	**4410**	**840**	**2100**

2.2. Farsi isolated digits database

A simple segmentation algorithm was developed for separating the digits in the numerical strings and to create the Farsi digits database. When designing the data entry form for the numerical strings, throughout all the strings, digits 1 to 9 were repeated 15 times, digit 0 was repeated 30 times, and the decimal point was repeated 3 times. This way we could control number of isolated digits that we could extract from the numerical strings. Samples of Farsi isolated digits are shown in Figure 4 and statistics of this database are included in Table 2.

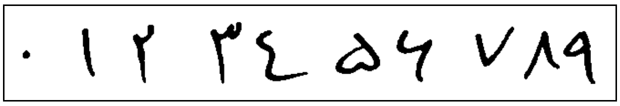

Figure 4. Samples of Farsi isolated digits.

Because separating all the digits was not possible, writers did not equally participate in the database for each digit. Therefore, some of the digits written by those writers that had the most participation were randomly removed from the database in order to normalize the participation. The algorithm is shown in Figure 5. Note that every time a digit is removed the most participating writer changes. This procedure was executed for each digit. Table 2 shows the final statistics for this database.

Table 2. Statistics of the isolated digits database.

Classes	Writers	Training Set	Verifying Set	Testing Set
10	175	11000	2000	5000

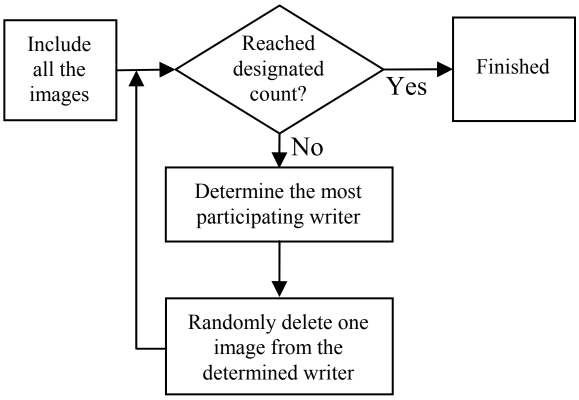

Figure 5. Algorithm of normalizing the participation.

2.3. Farsi isolated letters database

Although Farsi consists of 32 letters, yet when filling data entry forms out people use two different styles for the letter "ه" (pronounced: Heh) and "ا" (pronounced: Alef) and samples of those styles are shown in Figure 6 and Figure 7. With these styles, the number of isolated letters that we included in the form 1 reached 34.

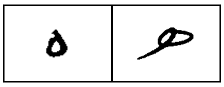

Figure 6. Two styles of writing the letter "ه".

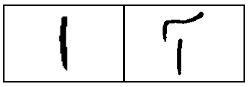

Figure 7. Two styles of writing the letter "ا".

Each writer wrote the isolated letters included in the form 1, two times. The statistics of this database are included in Table 3.

Table 3. Statistics of Farsi isolated letters database.

Classes	Writers	Training Set	Verifying Set	Testing Set
34	175	7140	1360	3400

2.4. Farsi legal amounts database

Two types of data were included in our second data entry form. The first type consisted of 41 words that are normally used for writing the legal amount on bank cheques plus four additional words consist of 2 currency units and the words "Over" and "Equal to" (in Farsi). The second type consisted of four worded number strings where three of those were pre-determined fields and one was a free field. In the free field, writers could write a worded number of their own. When including these images in the database, the free field was labeled manually. A sample of a worded number can be seen in Figure 8. Table 4 shows statistics of this database.

Table 4. Statistics of cursive worded number database.

Writers = 175	Classes	Training Set	Verifying Set	Testing Set
Fields	48	5040	960	2400
Free Field	175	105	20	50
Total	218	5145	980	2450

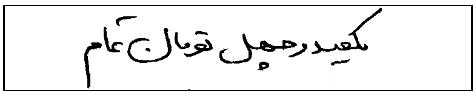

Figure 8. Example of a cursive worded number which reads: "One Hundred and Fourty Toumans Over".

2.5. Farsi dates database

Countries that have Farsi language speakers use a type of date called "Hijri Shamsi". The format of writing the date in Farsi is like this: **year/month/day**. A sample of a date is shown in Figure 9. The statistics of this database are also included in Table 5.

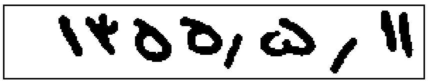

Figure 9. Example of a Farsi date.

Table 5. Statistics of the Farsi dates database

Classes	Writers	Training Set	Verifying Set	Testing Set
175	175	105	20	50

2.6. English digits

English digits have already been collected and included in different databases; however, a small set was included in the first form (each digit from 0 to 9 was repeated twice in each form) in order to capture the style of writing English digits by non-native English speakers (Iranians). Table 6 shows statistics of this database.

Table 6. Statistics of the isolated digits database.

Classes	Writers	Training Set	Verifying Set	Testing Set
10	175	2100	400	1000

3. Data Extraction

3.1. Preprocessing

Each form was completely scanned using a Lexmark-P3180 scanner whose resolution was set to 300 dpi at a grey level of 8 bits. The images were saved in PNG (Portable Network Graphics) indexed-color format files. PNG provides a patent-free replacement for GIF and also replaces many common uses of TIFF. [9]

All the databases consist of grayscale and binary versions of images and each set is included in a separate folder. First, grayscale images were extracted, and then all were converted to binary in a separate folder keeping the same filenames and the same folder structure. To convert each file to binary, the threshold of a grayscale image is calculated using the gray-level histogram [10], and then all the pixels with brightness less than that value are set to black, and the rest to white.

Before starting the process of extracting images from scanned forms, their salt and pepper noise was remove using the algorithm presented in [11].

3.2. Data Preparation

A computer program was developed to automatically extract images of the fields from the pre-processed scanned forms using a template that was manually designed for identifying the data entry fields relative to the anchoring marks at the corners of the forms. The program first recognized edge identifier anchor marks on the scanned image by a simple template matching technique. It then tried to match the template coordinates to anchor marks of the image by scaling and/or rotating the template if necessary. After that, all the fields were cut from the image, based on the boundaries in the matched template. The fields were saved as individual image files using the set they belonged to and the naming convention of the database.

To determine the set to which an image belongs, the writers were selected from different ages, genders, and jobs to serve in the training, testing, or verifying set. All the images extracted from each particular writer's form, were saved to the same set for making sure that the data sets are totally unrelated. For each image, a record was inserted into a Microsoft® Access™ database that includes the path to the image file relative to the base folder, the label of the image, the number of characters in the image, the number of words in the image, the type of the contents (numerical, date, cursive worded number or letter), and some other information. By querying this type of detailed information, future researchers will be able to find the proper set of images more easily.

4. Experimental Results

In order to show the application of our databases, we conducted some experiments on the recognition of handwritten isolated Farsi digits. We used our isolated digits database which contains 1100 training, 200 verifying, and 500 testing samples per digit.

4.1. Feature Extraction

In order to compare our results with some previous works, we used the features presented in [12]. Eight sets of features were used to represent images of digits: the outer profile from four directions; crossing counts; and a projected histogram from each of two directions.

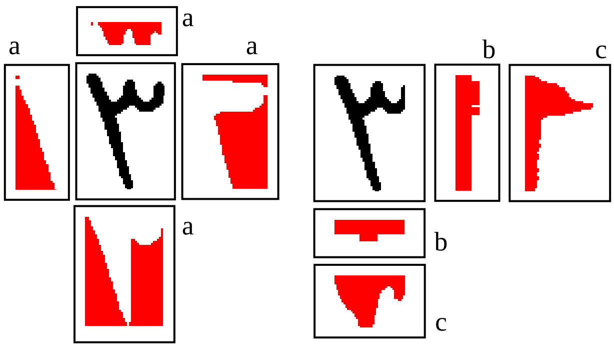

Figure 10. A sample of the features. a: outer profiles, b: crossing counts, c: projection histogram.

Each set produced an array that was later normalized to an array of size eight. The normalization was done using linear interpolation for up-sampling and averaging for down-sampling the array. The combination of all the feature sets produced a 64-member array that was used as our feature vector. A sample of features used in our experiment can be viewed in Figure 10.

4.2. Classification

For classification we used Support Vector Machines (SVM) [13] and Radial Basis Function (RBF) kernel. The parameter C was set to 11 and the parameter σ was set to 0.05. To find the best parameter values, we adjusted the parameters on the training set, and tested them on the verifying set. Parameters that gave the best results on the verifying set were used for classifying on the testing set. We used LIBSVM [14] for the implementation of our SVM classifier. Table 7 shows the overall results of our classifier compared to the results of [12] and [15]. The confusion matrix of the testing set is also shown in Table 8. Each row of this table shows how isolated digits in the testing set were classified or misclassified.

Table 7. Our results compared with [12] and [15].

	Our Results	Results of [12]	Results of [15]
Training Set	11000	4500	7390
Verifying Set	2000	-	-
Testing Set	5000	3600	3035
nSV*	1577	629	-
Training Error	**0.85%**	0.00%	0.24%
RR**	**97.32%**	99.44%	94.14%

* Number of Support Vectors, ** Recognition Rate

Table 8. The confusion matrix of the testing set using SVM with polynomial kernel.

	0	1	2	3	4	5	6	7	8	9
0	459 91.8%	33 6.6%	1 .2%	2 .4%		3 .6%			2 .4%	
1	12 2.4%	483 96.6%	2 .4%				2 .4%			1 .2%
2		1 .2%	493 98.6%	2 .4%	3 .6%		1 .2%			
3	1 .2%		17 3.4%	472 94.4%	8 1.6%	2 .4%				
4			3 .6%	4 .8%	492 98.4%		1 .2%			
5	8 1.6%					492 98.4%				
6	1 .2%	2 .4%	3 .6%	1 .2%	3 .6%	1 .2%	484 96.8%			5 1%
7								500 100%		
8							1 .2%		499 99.8%	
9							5 1%		3 .6%	492 98.4%

5. Discussion

This research effort has produced six databases. Each database is divided into training, verifying, and testing sets, which includes approximately 60%, 12%, and 28% of the available data respectively. All the databases are available in grayscale and binary versions. Table 9 and Table 10 show a comparison between our two important databases (Farsi isolated digits and Farsi isolated letters) and other similar available databases. Although the result of our recognition rate in Section 4 is a little bit lower than [12], our databases were not the same, and our isolated digits database has more samples compared to them. Also we used unseen data to test our classifier and in [12] testing set was used for adjusting parameters of the classifier. As our database is available for the research community, we hope that it can function as a standard comparison basis for Farsi handwritten recognition research.

Table 9. Comparison of number of samples in our Farsi isolated digit database with other databases.

Isolated Digits				
Database	Set	Training	Verifying	Testing
MNIST	English Digits	60,000	0	10,000
CEDAR	English Digits	5,802	0	707
CENPARMI	English Digits	4,000	0	2,000
CENPARMI	Arabic Digits	10,536	0	4,324
Our Database	**Farsi Digits**	**11,000**	**2,000**	**5,000**

Table 10. Comparison of number of samples in our Farsi isolated letters database with other databases.

Isolated Letters				
Database	Set	Training	Verifying	Testing
CEDAR	English Letters	19,145	0	2,183
Our Database	**Farsi Letters**	**7,140**	**1,360**	**3,400**

6. Conclusion and Future Works

We have presented six new standard databases consisting of handwritten Farsi numerical strings, digits, letters, legal amounts and dates which can serve as a basis for future research in offline Farsi handwritten recognition. These databases are available to the research community upon request to the Center of Pattern Recognition and Machine Intelligence (CENPARMI) of Concordia University. Our database contains binary and grayscale versions of the images allowing for experimentation and comparison with both grayscale and binary preprocessing and recognition techniques. In the future, the databases may be expanded by collecting more data entry forms, and adding more sets such as Farsi words, sub-words and sentences. Furthermore, the sets may be easily adopted for Farsi-based cheque-processing systems. Later, we would like to develop sophisticated segmentation and recognition algorithms for processing samples of these databases.

7. References

[1] I. Guyon, R. Haralick, J. Hull, and I. Phillips, "Database and benchmarking," In H. Bunke and P. Wand, editors, *Handbook of Character Recognition and Document Image Analysis*. World Scientific, 1997, Chapter 30, pp. 779–799.

[2] R. Wilkinson, J. Geist, S. Janet, P. Grother, C. Burges, R. Creecy, B. Hammond, J. Hull, N. Larsen, T. Vogl, and C. Wilson. The first census optical character recognition systems conf. #NISTIR 4912, The U.S. Bureau of Census and the National Institute of Standards and Technology, Gaithersburg, MD, 1992.

[3] J. Hull, "A database for handwritten text recognition research," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, May 1994, Volume 16, Issue 5, pp. 550–554.

[4] C. Y. Suen, C. Nadal, R. Legault, T. Mai, and L. Lam, "Computer recognition of unconstrained handwritten numerals," *Proc. of the IEEE*, 1992, Volume 7, Issue 80, Pages 1162–1180.

[5] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet, "Unipen project of on-line data exchange and benchmarks," *Proc. of the 12th IAPR Int. Conf on Pattern Recognition*, Jerusalem, Israel, Oct. 1994, pp. 29–33.

[6] Yousef Al-Ohali, Mohamed Cheriet, and C.Y. Suen, "Databases for recognition of handwritten Arabic cheques," *Proceedings of the Seventh Int. Workshop on Frontiers in Handwritten Recognition*, Sep 2000, pp. 601-606.

[7] F. Jelinek, "Self-organized language modeling for speech recognition," In A. Waibel and K.-F. Lee, editors, *Readings in Speech Recognition*, Morgan Kaufmann Publishers, Inc., 1990, pp. 450–506.

[8] D. Kim, Y. Hwang, S. Park, E. Kim, S. Paek, and S. Bang, "Handwritten Korean character image database PE92," *In Proceedings of the Second Int. Conference on Document Analysis and Recognition*, 1993, pp. 470–473.

[9] Chris Lilley, *PNG (Portable Network Graphics)*. The World Wide Web Consortium (W3C), Details available at http://www.w3.org/Graphics/PNG/

[10] N. Otsu, "A thresholding selection method from gray-level histogram," *IEEE Transactions on Systems, Man, and Cybernetic*, 1979, Volume 9, pp. 62-66.

[11] Jae S. Lim, *Two-Dimensional Signal and Image Processing*, Englewood Cliffs, editor, Prentice Hall, USA, 1990, pp. 469-476.

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

[13] C.J.C Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, 1998, Volume 2, pp. 121–167.

[14] Chih-Chung Chang, Chih-Jen Lin, *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[15] J. Sadri, C. Y. Suen, T. D. Bui, "Application of Support Vector Machines for Recognition of Handwritten Arabic/Persian Digits," Proceedings of the Second Conference on Machine Vision and Image Processing & Applications (MVIP2003), Vol. 1, pp. 300-307, Feb. 2003, Tehran, Iran.