

# Introducing a very large dataset of handwritten Farsi digits and a study on their varieties

Hossein Khosravi<sup>a,b,\*</sup>, Ehsanollah Kabir<sup>a</sup>

<sup>a</sup> Department of Electrical Engineering, Tarbiat Modarres University, Tehran, Iran

<sup>b</sup> Research and Development Unit, HODA System Co., Tehran, Iran

Received 5 September 2005; received in revised form 24 September 2006

Available online 21 February 2007

Communicated by A. M. Alimi

## Abstract

A very large dataset of handwritten Farsi digits is introduced. 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. A method for finding variety of handwritten digits in a typical dataset is proposed. Based on this method, training and test subsets are provided to facilitate sharing of results among researchers as well as performance comparison.

© 2007 Elsevier B.V. All rights reserved.

**Keywords:** Farsi digits; Persian; Arabic; Handwriting; Dataset; Digit variety

## 1. Introduction

Recognition of handwritten characters is one of the most interesting topics in pattern recognition domain. In OCR applications, handwritten character recognition, especially digit recognition, is dealt with in postal mail sorting, bank check processing, form data entry, etc. In recent decades many researchers worked on this topic (for example see (Liu, 2003; Mayraz and Hinton, 2002; Oliveira, 2002; Trier et al., 1996)). During these years standard datasets were developed to help researchers sharing their results on the same dataset and comparing performance of their classifiers.

For some scripts such as English, there are standard datasets available that we will review some of them in the next section, but in the case of Farsi script there exists no freely or commercially available large dataset.

The rest of this paper is organized as follows. In Section 2, we review some existing English datasets. Section 3 describes the background of this work and the procedure of data collection. In Section 4, a method for finding variety of samples in the dataset is described. In Section 5, we use the result of this method to divide the dataset into training and test sets. Conclusion follows in Section 6 and finally in Appendix A, we describe the dataset format.

## 2. Handwritten digit datasets

There are several datasets of digits and letters in English. We review those commonly used by researchers.

The CENPARMI<sup>1</sup> digit dataset (Suen, 1992) is available from CENPARMI, Concordia University. It contains 6000 digits collected from the envelop images of USPS<sup>2</sup>, scanned at 166 dpi. In this dataset 4000 images, 400 samples per class, are specified for training and the remaining 2000 images for test.

\* Corresponding author. Address: Department of Electrical Engineering, Tarbiat Modarres University, Tehran, Iran. Fax: +98 21 8800 5040.

E-mail addresses: [hosseinkhosravi@modares.ac.ir](mailto:hosseinkhosravi@modares.ac.ir), [hosseinkhosravi@gmail.com](mailto:hosseinkhosravi@gmail.com) (H. Khosravi), [kabir@modares.ac.ir](mailto:kabir@modares.ac.ir) (E. Kabir).

<sup>1</sup> Center for Pattern Recognition and Machine Intelligence.

<sup>2</sup> United States Postal Service.

Table 1  
Some popular digit datasets

Dataset	dpi	Training samples	Test samples	Total samples
CENPARMI	166	4000	2000	6000
CEDAR	300	18,468	2711	21,179
MNIST	Normalized into 20 * 20	60,000	10,000	70,000
USPS	300	7291	2007	9298

The CEDAR<sup>3</sup> digit dataset is available from CEDAR, SUNY<sup>4</sup> at Buffalo. The images were scanned at 300 dpi. The training and test sets contain 18468 and 2711 digits, respectively. The number of samples in both training and test sets differ for each class. Since some images in the test set are poorly segmented, a subset of 2213 well-segmented images are also provided for testing (Liu, 2003).

The MNIST, modified NIST,<sup>5</sup> dataset (LeCun et al., 1995) was extracted from the NIST datasets SD3 and SD7. The training and test sets are composed from both SD3 and SD7. Samples are normalized into 20 \* 20 gray-scale images with aspect ratio reserved, and the normalized images are located in a 28 \* 28 frame. The dataset is available from LeCun. Number of training and test samples are 60,000 and 10,000 respectively.

At last the USPS digit dataset has 7291 training and 2007 test samples (Hull, 1994). Table 1 lists these datasets briefly.

In the case of Farsi language, although much work has been carried out in character recognition (for example (Soltanzadeh and Rahmati, 2004; Azmi and Kabir, 2001; Dehghan, 2001; Nabavi et al., 2005)), almost all of these works relied on privately collected datasets and there is no standard dataset of handwritten digits or letters.

### 3. Farsi digit dataset

#### 3.1. Data collection

Although Farsi is a right to left script, its digits are written from left to right. Sample handwritten digits are shown in Fig. 1.

We used 11,942 registration forms of two types for this dataset. Two sample forms and their fields of interest are shown in Fig. 2. The first type was the registration form for Iran university entrance examination for the M.Sc. degree, filled by senior B.Sc. students, and the second type was the registration form for the university entrance examination for B.Sc. degree filled by senior high school students. There were 5393 forms of type 1 and 6549 forms of type 2. All these forms were scanned at 200 dpi in 24 bit color format with a high speed scanner, Axiome 4300.

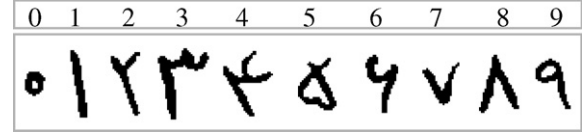


Fig. 1. Sample handwritten Farsi digits.

There were several fields in both types of forms. We used two digit fields from type 1, including *Postal Code* and *National Code*, each of 10 digits length and three digit fields from type 2 including *Record Number*, *Identity Certificate Number* and *Phone Number* that at most have 26 digits, while in average about 20 digits. Both forms are in color. In both types, handwritten texts are in blue or occasionally in black.

#### 3.2. Digit extraction and recognition

To extract the digits, we must find the regions of interest. There were at least two reference marks (squares) in each form (circled in Fig. 2). We first search for these marks using a simple and fast algorithm shown in Fig. 3. If they are not found, the form is rejected. This situation occurs rarely, e.g. when the paper is scanned upside down or the reference square is too noisy. Then, if the reference squares are not in their expected positions, the form is rotated and shifted so that these squares are placed in the right positions.

Now we find the regions of interest from their coordinates. Handwritten digits are in blue or black and the background is almost light red. So we apply appropriate thresholds to color components and binarize the image. Finally a 3 \* 3 median filter is applied to remove grain noise.

In the next stage, these regions, forming a single black and white image (Figs. 2c and d), are sent to the connected component labeling module. Then the components which are close to each other within a specified threshold are merged to prevent broken digits. Closeness was defined by five distances: a distance between the centers of two adjacent components and four distances between the four sidewalls of their surrounding frames.

$$\begin{aligned}
 d_1 &= |C_i - C_j| \\
 d_2 &= \min\{|L_i - R_j|, |L_j - R_i|\} \\
 d_3 &= \min\{|L_i - L_j|, |R_i - R_j|\} \\
 d_4 &= \min\{|T_i - B_j|, |T_j - B_i|\} \\
 d_5 &= \min\{|T_i - T_j|, |B_i - B_j|\}
 \end{aligned} \tag{1}$$

where subscripts  $i$  and  $j$  stands for two different regions and  $C$ ,  $R$ ,  $L$ ,  $T$  and  $B$  stand for Center, Right, Left, Top and Bottom side of the region, respectively. Fig. 4 shows these distances for two regions. If  $(d_1 < T)$  or  $((\min(d_2, d_3) < T) \text{ and } (\min(d_4, d_5) < T))$ , the two regions will be merged.  $T$  is an empirical threshold which is set to 10 pixels.

<sup>3</sup> Center of Excellence for Document Analysis and Recognition.

<sup>4</sup> The State University of New York.

<sup>5</sup> National Institute of Standards and Technology.

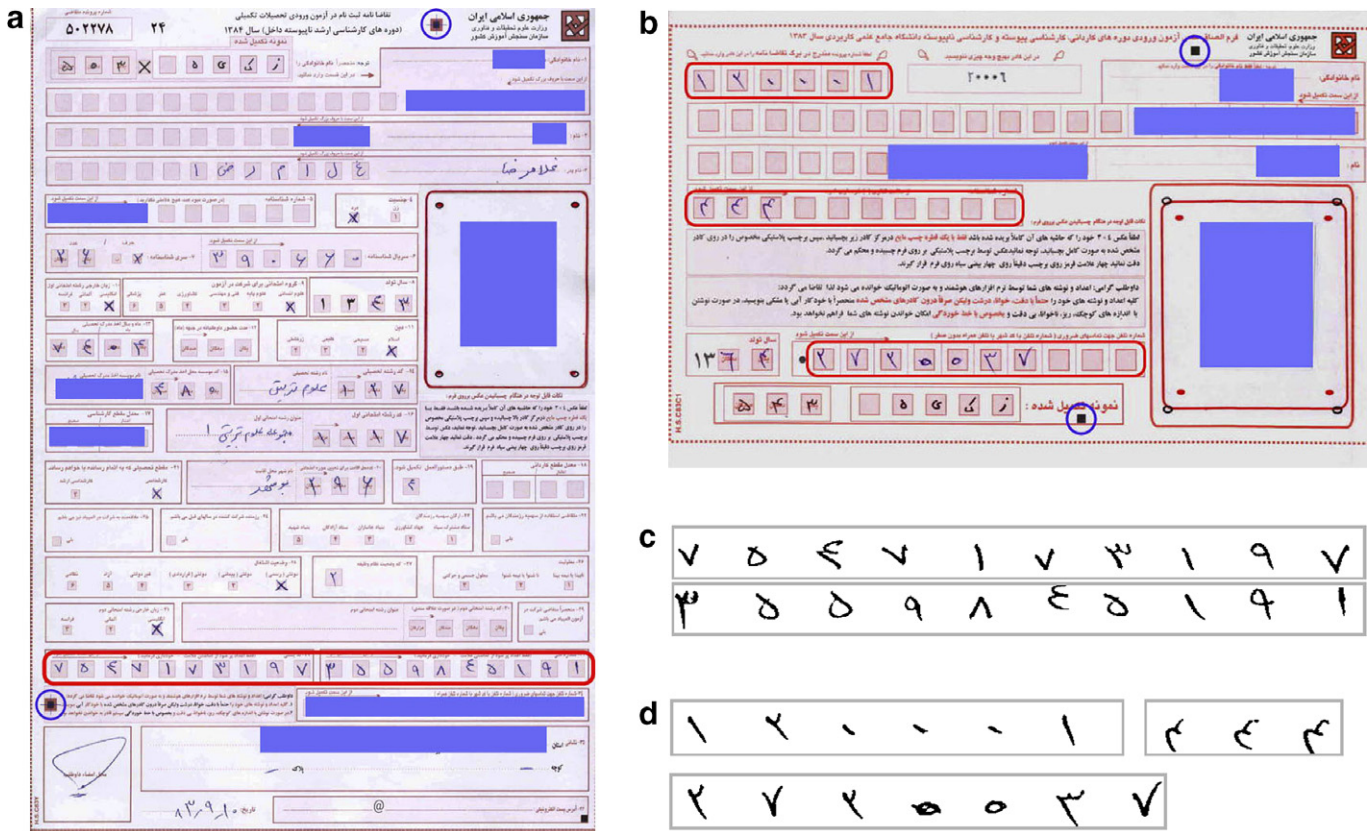


Fig. 2. Sample form images. (a) Type 1. (b) Type 2. (c) and (d) Binary images extracted from fields of interest.

An example of a digit, broken to three parts due to binarization, that is merged by these distance rules is shown in Fig. 5. In this figure, regions 1 and 2 satisfy both conditions, so they are merged. Also regions 2 and 3 satisfy the second condition and are merged.

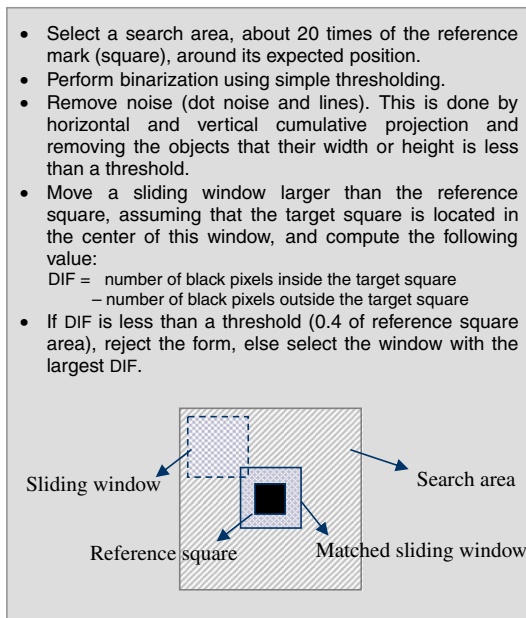


Fig. 3. Procedure of searching for a reference square.

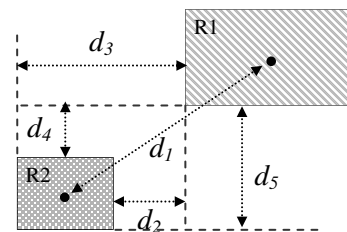


Fig. 4. Distances between two regions R1 and R2.

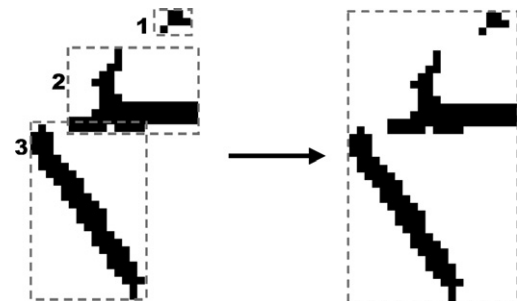


Fig. 5. A broken digit of 3 parts that is merged by the distance rules.

The extracted digits were passed to a recognition engine to be labeled and stored. The recognition engine was a multiple classifier system consisting of four MLP classifiers created based on AdaBoost.M2 algorithm (Freund and



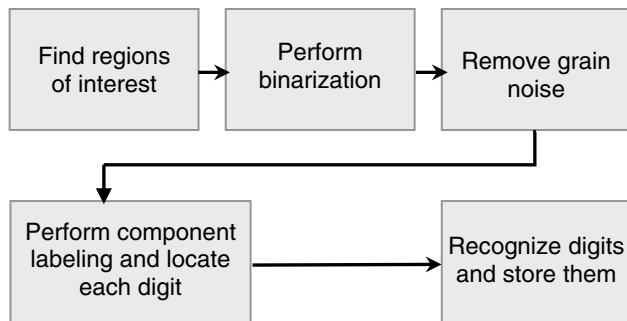


Fig. 6. Process diagram of extracting the digits from registration forms and recognizing them.

Schapiro, 1996). These classifiers were trained with a private dataset of digits containing 15,000 training and 5000 test samples. We used a modified gradient technique for feature extraction (Trier et al., 1996). The recognition rate was 100% on the training set and 98.8% on the test set. The process diagram for dataset collection is shown in Fig. 6.

### 3.3. Manual refinement

After processing all forms, we had 94,530 digits from type 1 and 128,203 digits from type 2. We randomly took 60,000 digits from each form. Then we corrected the errors manually. This process was done using a verification program that shows the digits of same class on the screen so that the user can simply spot the errors and correct them using keyboard (Fig. 7). The process took about 16 man-hours. The digits with very poor quality, not recognizable by human, and noise objects were removed (see Fig. 8).

After manual refinement stage, we had 102,352 digits of reasonable quality with correct labels. Some samples of dif-

ferent qualities are shown in Fig. 9. Number of digits in each class is listed in Table 2.

In Section 4 we will describe a method for finding different variety of samples in the dataset.

## 4. Finding variety of digits in each class

There are some previous works in the field of style recognition (e.g. Franke and Oberlander, 1993; Crettez, 1995; Ma and Doermann, 2004; Tanprasert and Tang, 1999), mainly in word or paragraph level. Franke and Oberlander (1993) proposes a method to find whether a script is handwritten or machine printed. Crettez (1995) and Ma and Doermann (2004) propose two methods that distinguish between different handwriting styles in word level. In (Tanprasert and Tang, 1999) a method for finding different font styles, regular, bold and italic, in a Thai printed script is proposed.

In (Suen et al., 1977) a quantitative measure, dispersion factor, was proposed for the quality of handprinted characters. This measure was based on the frequency diagram. In (Downton et al., 1988), the idea of frequency diagram was used to define a similarity factor for character recognition. In this paper, we use a modified version of frequency diagram and propose a similarity factor to find the variety of the digits in our dataset (Fig. 10). This helps us in dividing the dataset into training and test sets.

### 4.1. The modified frequency diagram

The frequency diagram depicts the density of the black pixels of the samples. The lower the frequency is, the greater the samples differ from one another (Suen et al., 1977). To compute frequency diagram, we normalized the

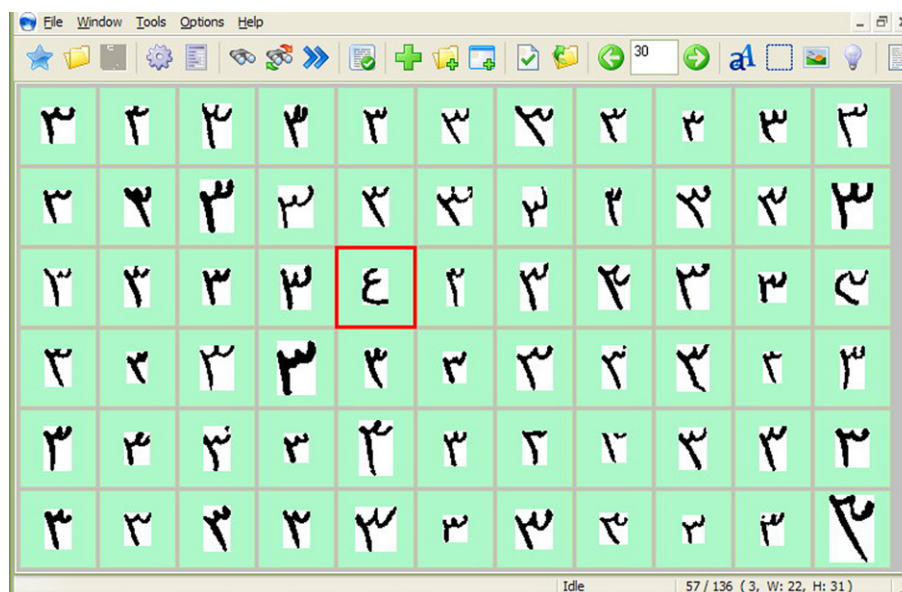


Fig. 7. Screen designed for manual refinement.

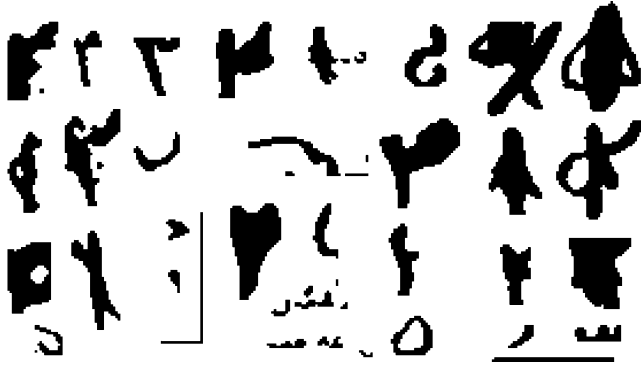


Fig. 8. Some deleted samples.

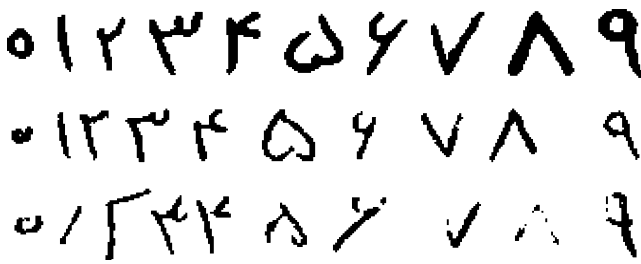


Fig. 9. Sample digits of different qualities from final dataset.

Table 2  
Number of digits in each class

0	1	2	3	4
10070	10331	9927	10336	103333
5	6	7	8	9
10110	10256	10364	10264	10373

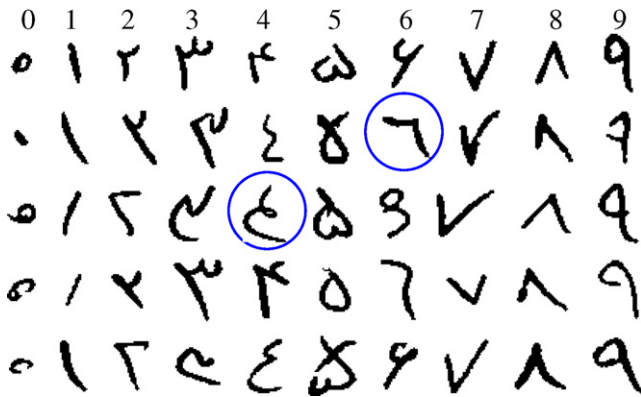


Fig. 10. Some different styles of handwritten Farsi digits.

samples to  $40 \times 40$  with aspect ratio reserved. An example of frequency diagram is shown in Fig. 11.

We modified this frequency diagram in a way that any black pixel causes an increase in frequency, and any white pixel causes a decrease in frequency. Now, the lower the

absolute value of frequency, the greater the dissimilarity within samples. The modified frequency diagram of Fig. 11 is shown in Fig. 12. The modified frequency diagram for class  $i$  can be formulated as follows:

$$D_i(x, y) = \frac{1}{N_i} \sum_{n=0}^{N_i} (F_n(x, y) * 2 - 1) * 100 \quad (2)$$

$$x = 0, 1, \dots, 39, y = 0, 1, \dots, 39$$

where:

$F_n(x, y)$  is the pixel value of  $n$ 'th sample at coordinates  $(x, y)$  which is 0 or 1.

$N_i$  is the number of samples from class  $i$ .

In this way, we construct a frequency diagram for each class using all samples of that class.

#### 4.2. Similarity factor

Now we define a similarity factor. We use  $D_i$  as the template for class  $i$ . This template is used later in Section 4.3 to analyze the variety of samples in each class. The similarity factor between input sample,  $F(x, y)$ , and class template  $D_i$  is as follows:

$$S_i = \sum_{y=0}^{39} \sum_{x=0}^{39} \text{sgn}[(F(x, y) \times 2 - 1) \times D_i(x, y)] \times |D_i(x, y)| \quad (3)$$

$S_i$  increases either when a black pixel of  $F(x, y)$  occurs at positive values in  $D_i$  or when a white pixel occurs at negative values in  $D_i$ .

#### 4.3. Variety analysis

To find different varieties of digits in this dataset, the following procedure was carried out. At first we constructed a template,  $D_{i1}$ , for each class from the whole dataset using Eq. (2) (Table 3, first row). Then we compared each sample with its class template,  $D_{i1}$ , using similarity factor (Eq. (3)). If similarity was greater than 0.75, an empirical threshold, the sample was marked as TRUE else it was marked as FALSE. All samples of the dataset were tested in this manner. About 73% of all samples marked as TRUE and others as FALSE. In this stage we named the TRUE samples as S1 and constructed a new template from them (Table 3, second row). The representative images show that 73% of peoples write the digits almost in this way.

In the next step we sat aside S1 from the dataset and performed this method on the remaining 27% of samples, S'1, i.e. we created new templates from S'1,  $D_{i2}$ , and again compared all samples of S'1 with these new templates. In this step 37% of the samples S'1 that is 9.83% of the total samples, were marked as TRUE. We named these samples as S2. This procedure can be applied sequentially, each time on the samples of the previous step which marked as FALSE. We performed this procedure 3 times. The results are shown in Table 3. It can be seen that the second row of

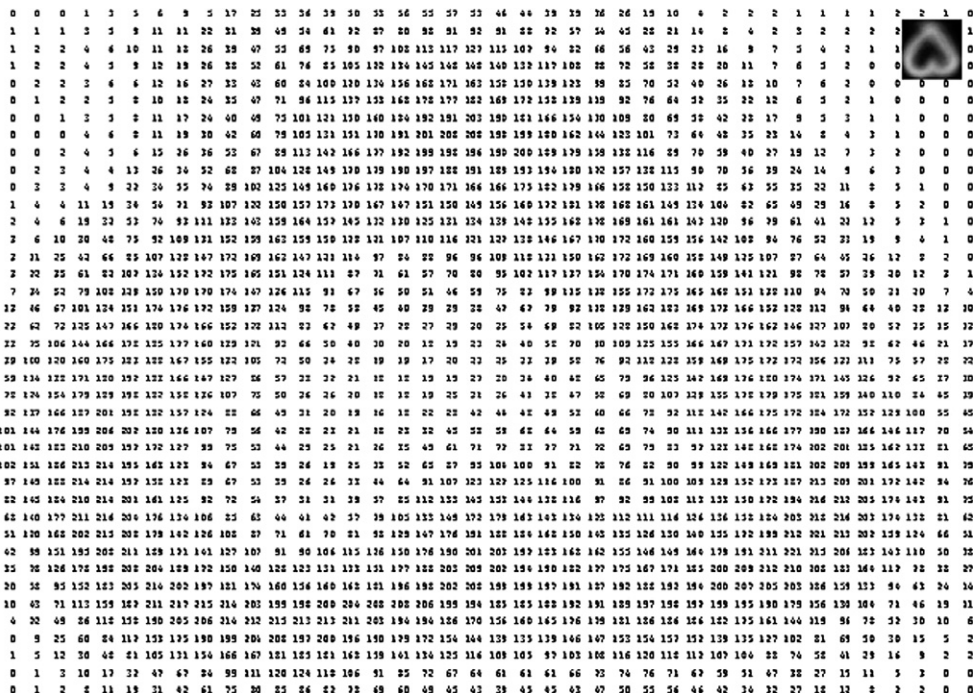


Fig. 11. The Frequency diagram of 300 samples of digit '5'. The corresponding grayscale image is shown on top-right corner.

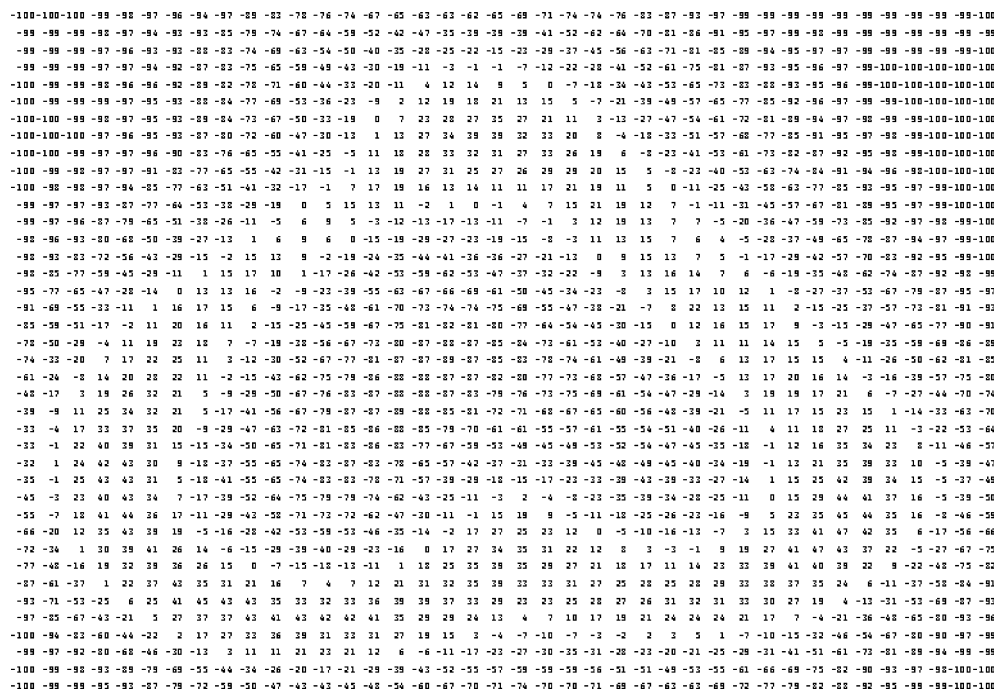




















































Fig. 12. The modified frequency diagram of 300 samples of digit '5'.

Table 3, S1, are almost like the first row of Fig. 9. Also other samples in Fig. 9 can be seen in this table. For example, digits 4 and 6 in the third row, S2, are similar to those of Fig. 9, shown in circle. This table also shows that some digits like 6 and 4 have more variety of writing

styles than others (see columns for digits 4 and 6). More exactly we can see that digit 6 has the most diversity and digits 4, 3, 2 and 5 rank afterwards. The highlighted fields in Table 3 are those styles that have more than 10% of population.

Table 3  
Different styles of writing Farsi digits

Digit		0	1	2	3	4	5	6	7	8	9	Total
All samples	Quantity	10070	10330	9923	10334	10333	10110	10254	10363	10264	10371	102352
	Percentage	100	100	100	100	100	100	100	100	100	100	100%
	Frequency Diagram											
Samples S1	Quantity	4548	10147	9133	8552	8112	3487	7543	7373	7186	9116	75197
	Percentage	<b>45.16</b>	<b>98.23</b>	<b>92.04</b>	<b>82.76</b>	<b>78.51</b>	<b>34.49</b>	<b>73.56</b>	<b>71.15</b>	<b>70.01</b>	<b>87.90</b>	73.47%
	Frequency Diagram											
Samples S2	Quantity	626	150	503	878	1061	797	1345	1796	1937	965	10058
	Percentage	6.22	1.45	5.07	8.50	<b>10.27</b>	7.88	<b>13.12</b>	<b>17.33</b>	<b>18.87</b>	9.30	9.83%
	Frequency Diagram											
Samples S3	Quantity	72	24	182	421	238	225	426	395	388	148	2519
	Percentage	0.71	0.23	1.83	4.07	2.30	2.23	4.15	3.81	3.78	1.43	2.46%
	Frequency Diagram											
Samples S4 (remaining samples)	Quantity	4824	9	105	483	922	5601	940	799	753	142	14578
	Percentage	<b>47.90</b>	0.09	1.06	4.67	8.92	<b>55.40</b>	9.17	7.71	7.34	1.37	14.24%
	Frequency Diagram											

## 5. Choosing the training and test sets

To facilitate sharing of results on this dataset between researchers, we provide two distinct datasets for training and test.

From Table 3 it can be seen that the most usual styles are fallen into samples S1, and other varieties are fallen into S2, S3 and S4. So we tried to select most of training samples from S1. To be more accurate we selected from each category a number of samples equal to their proportion in total samples, i.e. 73.47% of training samples were selected from S1, 9.83% from S2 and so on. Then we set aside training samples and select test samples from the remaining samples, randomly. In this way the training set is a true representation of the whole population, while the test set is selected without any predefined information.

We selected 60,000 samples for training set and 20,000 for test. The remaining samples are also available in another subset (see Appendix A).

## 6. Conclusion

In this paper, we introduced a very large dataset of handwritten Farsi digits. The dataset samples were extracted from about 12,000 registration forms of two types. The procedure of preprocessing, finding areas of interest and digit extraction was described. We also proposed a method for finding different digit varieties in a typical dataset, based on modified frequency diagram and an appropriate similarity factor. We applied this method on Farsi digit dataset to divide it into two distinct subsets of training and test. The final subsets are as follows.

Total samples:	102,352 digits
Training set:	60,000 digits
Test set:	20,000 digits
Remaining samples:	22,352 digits

## Appendix A. Dataset specification and availability

The dataset is available in four separate files, **Total.cdb**, **Training.cdb**, **Test.cdb**, **Remaining.cdb**. The file format is described here with a pseudo code:

**Skip Header** (1024 bytes)

while not End of File

{

    read **Start Byte**: (1 byte) 0xFF that specifies the start of new image

    read **Label**: (1 byte) character label

    read **Width**: (1 byte) character width

    read **Height**: (1 byte) character height

    read **Byte Count**: (2 bytes) number of bytes for this character.

    //Runlength coding on each row

    for  $y = 0$  to Height

        while( $x < \text{Width}$ )

        {

            read **NumOfWhitePixels**,

            read **NumOfBlackPixels**;

        }

    }

Source codes for reading the dataset files are available in Matlab, C++ and Pascal. To get the dataset please contact [kabir@modares.ac.ir](mailto:kabir@modares.ac.ir). or see the homepage <http://www.modares.ac.ir/eng/kabir>.

## References

- Azmi, R., Kabir, E., 2001. A new segmentation technique for omnifont Farsi text. *Pattern Recognition Lett.* 22, 97–104.
- Crettez, J. 1995. A set of handwriting families, style recognition. In: Third International Conference on Document Analysis and Recognition, pp. 489–494.
- Dehghan, M. et al., 2001. Unconstrained Farsi handwritten word recognition using fuzzy vector quantization and hidden Markov models. *Pattern Recognition Lett.* 22, 209–214.
- Downton, A.C., Kabir, E., Guillevis, D. 1988. Syntactic and contextual post-processing of handwritten addresses for optical character recognition. In: 9th International Conference on Pattern Recognition, pp. 1072–1076.
- Franke, J., Oberlander, M. 1993. Writing style detection by statistical combination of classifiers in form reader applications. In: Second International Conference on Document Analysis and Recognition, pp. 581–584.
- Freund, Y., Schapire, R.E. 1996. Experiments with a new boosting algorithm. In: International Conference on Machine Learning, Italy, pp. 148–156.
- Hull, J.J., 1994. A database for handwritten text recognition. *IEEE Trans. Pattern Anal. Machine Intell.* 16 (5), 550–554.
- LeCun, Y., MNIST OCR data, <http://www.research.att.com/yann/exdb/mnist/index.html>.
- LeCun, Y. et al., 1995. Comparison of learning algorithms for handwritten digit recognition. In: International conference on Artificial Neural networks, France, pp. 53–60.
- Liu, C.L. et al., 2003. Handwritten digit recognition: Benchmarking of state-of-the-art techniques. *Pattern Recognition* 36, 2271–2285.
- Ma, H., Doermann, D. 2004. Adaptive word style classification using a Gaussian mixture model. In: 17th International Conference on Pattern Recognition, pp. 606–609.
- Mayraz, G., Hinton, G.E., 2002. Recognizing handwritten digits using hierarchical products of experts. *IEEE Trans. Pattern Anal. Machine Intell.* 24 (2), 189–197.
- Nabavi, S.H., Ebrahimpour, R., Kabir, E. 2005. Recognition of handwritten Farsi digits using classifier combination. In: 3rd Conference on Machine Vision, Image Processing and Applications, Tehran, pp. 116–119 (in Farsi).
- Oliveira, L.S. et al., 2002. Automatic recognition of handwritten numerical strings, a recognition and verification strategy. *IEEE Trans. Pattern Anal. Machine Intell.* 24 (11), 1438–1454.
- Soltanzadeh, H., Rahmati, M., 2004. Recognition of Persian handwritten digits using image profiles of multiple orientations. *Pattern Recognition Lett.* 25 (14), 1569–1576.



- Suen, C.Y. et al., 1992. Computer recognition of unconstrained handwritten numerals. *Proc. IEEE* 80 (7), 1162–1180.
- Suen, C.Y., Shinghal, R., Kwan, C.C. 1977. Dispersion factor, a quantitative measurement of the quality of handprinted characters. In: *International Conference of Cybernetics and Society*, pp. 681–685.
- Tanprasert, C., Sae Tang, S. 1999. Thai type style recognition. In: *IEEE International Symposium on Circuits and Systems*, pp. 336–339.
- Trier, O.D., Jain, A.K., Taxt, T., 1996. Feature extraction methods for character recognition – a survey. *Pattern Recognit.* 29 (4), 641–662.