STATE-OF-THE-ART IN FARSI SCRIPT RECOGNITION

Javad Sadri, Sara Izadi, Farshid Solimanpour, Ching Y. Suen, Tien D. Bui

CENPARMI, Computer Science and Software Engineering Department
Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, Quebec
Canada, H3G 1M8, Tel: (514)-848-2424-Ext:7950, Fax: (514)-848-2830
Emails: {j_sadri, s_izadin, f_solima, suen, bui}@cse.concordia.ca

ABSTRACT

In this paper, a brief history of the evolution of Farsi (Persian) script is presented, including how the Farsi alphabet was derived from the Arabic alphabet. Important features, similarities and dissimilarities between Farsi and Arabic scripts, from the Optical Character Recognition (OCR) point of view are discussed. Also, a brief review of some of the state-of-the-art techniques in off-line and on-line Farsi script recognition and segmentation, as well as challenges ahead are briefly presented.

1. INTRODUCTION

Farsi (Persian) and Arabic are two important cursive scripts used mainly in the Middle East and some other neighboring countries. Farsi is the main language used in Iran and Afghanistan, and it is spoken by more than 110 million people, including some people in Tajikistan, and Pakistan. Arabic is spoken in all Arab countries, both in the Middle East and in Africa, and it is used by 234 million people worldwide. In western countries, it is commonly thought that Farsi and Arabic scripts are the same. Although, the Farsi alphabet was derived from the Arabic alphabet, minor yet important differences exist in their alphabets and their styles of writing. Due to these differences, a system adjusted for automatic recognition of one script might not perform well for the other one. While much research on Arabic recognition has been published and introduced internationally, most of the research in Farsi recognition has been presented only in Farsi Journals and Iranian conferences. To the best of our knowledge, research in Farsi script recognition has not yet been widely introduced to the research community. This paper is organized as follows: Section 2 presents a brief history of the evolution of Farsi script, and its historic connection to Arabic script. In Section 3, similarities and dissimilarities among Farsi and Arabic scripts from the OCR point of view are highlighted. Section 4 presents the general structure of a Farsi Script recognition system. Sections 5-7 review some of the research efforts towards off-line on-line Farsi script recognition/ segmentation. Section 8 introduces some of the existing databases for research on Farsi script recognition. Section 9 lists some of the remaining challenges in Farsi script recognition, and draws the conclusion.

CENPARMI (Center for Pattern Recognition and Machine Intelligence)

2. HISTORY AND EVOLUTION OF THE FARSI (PERSIAN) SCRIPT

In this section, we briefly review the history of Persian script and its connection to Arabic script. Historically, evolution of the Persian script falls into three periods: Old, Middle, and Modern.

2.1. Old Persian Script (550 to 330 B.C.)

Old Persian script was a cuneiform type script dating from the time of the Achaemenid dynasties in Persia (6th-4th century B.C.) [1],[2] . In that script, characters were made of strokes, which could be impressed upon soft materials by a stylus with an angled end. Figure 1 shows the alphabet and numbers used in Old Persian script.

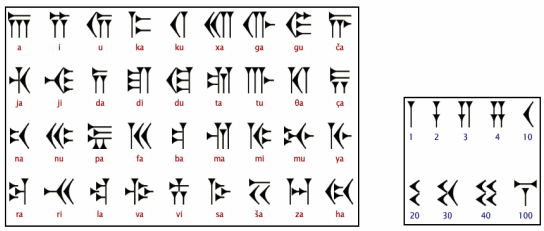

Fig. 1. Old Persian (cuneiform type): alphabet and numbers.

2.2. Middle Persian (300 B.C. to 900 A.D.)

Middle Persian includes the Iranian dialects as they appeared from about 300 B.C. to about 900 A.D. Middle Persian is generally called Pahlavi (a derivative of the old Persian word 'Parthian'). It was the language of quite a large body of Zoroastrian literature, the state religion of Sassanid in Iran. An example of Pahlavi Script is shown in Figure 2.

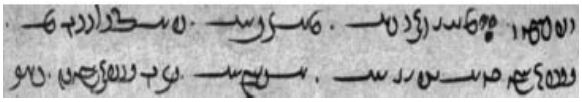

Fig. 2. A sample of Pahlavi (Middle Persian) script.

2.3. Modern Persian Script (after 7th century)

After the fall of the Sassanids and conquest of Persia (Iran) by the Arab Muslims (in the 7th century - 650 A.D), Pahlavi script gradually gave way to the Arabic script. The introduction of Islam brought a massive infusion of loaned words from Arabic to the Persian language. The Arabic alphabet was gradually adopted as the Persians' new alphabet, as the script of politics, religion, and education. An example of modern Persian script is shown in Figure 3.

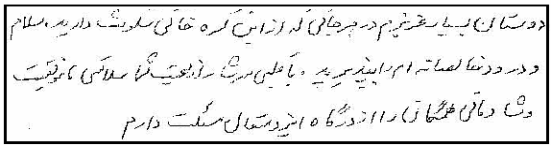

Fig. 3. A sample of Modern Persian script.

3. SIMILARITIES AND DISSIMILARITIES OF FARSI AND ARABIC SCRIPTS

3.1. Farsi and Arabic Alphabets

The Farsi alphabet has four more letters than the Arabic alphabet. Figure 4 shows the Farsi alphabet, which has 32 letters; whereas, the Arabic alphabet has only 28. In the specified figure, the letters that are identified by a closed circle are not included in the Arabic alphabet. These letters represent phonemes that do not appear in Arabic phonology. In both Farsi and Arabic Alphabets there are several letters that share the same basic form and differ only by a small complementary part. The complementary part could be a dot, a group of dots or a slanted bar. It can lie above, below or inside the letter. In fact, all the letters in the Farsi alphabet are derived from 18 basic shapes. These letters are distinguished by placing one dot (10 cases), two dots (3 cases) or three dots (5 cases) above or below of them. The vowels for A, E and O have no letters and may be shown as a small diagonal underbar stroke (for E), an over-bar stroke (for A), or a small comma (for O). Therefore, another characteristic of both Farsi and Arabic scripts is the existence of some diacritics marks written above or below the letters. These diacritics indicate vowels, where consonant letters are connected together to make their pronunciation easier. This is also another important feature which must be considered when designing OCR systems for these scripts

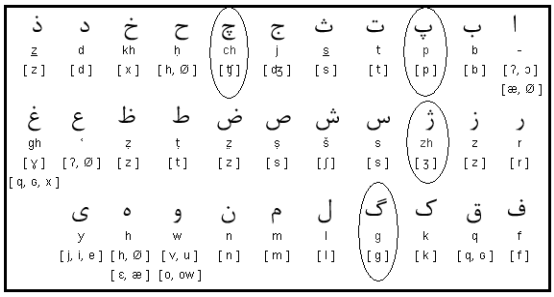

Fig. 4. Farsi alphabet. Letters identified by a closed circle are not included in the Arabic Alphabet.

3.2. Cursiveness of the Words and Connection of the Letters

The Arabic scripts and all of its derived forms (including Farsi Script) are inherently cursive. In both of these scripts, the position of each letter in the word and its preceding or following letter in the same word (if there is any), are the factors that determine the shapes of the letter. In the Farsi alphabet, similar to the Arabic alphabet, a letter can appear in four different forms: detached, initial, medial, and final. All Farsi letters (with seven exceptions / six in Arabic) can be connected to other letters from both the right and the left sides. The seven exceptional characters can only be connected to other letters from the right side. Therefore, if any of those seven letters appear in the middle of a word, there will be a gap in connectivity.

3.3. Farsi and Indian (Hindi) Digits

Farsi and Indian (Hindi) digits look very similar; however, there are minor but important differences between Farsi and Indian (Hindi) handwriting of digits [3]. Indian (Hindi) digits are used in most Arabic countries. Farsi digits are used mainly in Iran, and normally form 13 classes of shapes because of the two different ways of writing the numbers 0, 4 and 6. At the same time, Indian (Hindi) digits normally form 11 classes because of the two different ways of writing the number 3. Also the number 5 is written differently in these two scripts. Figure 5 compares Arabic, Farsi and Indian (Hindi) handwritings of digits. Note that Arabic digits (used mainly in Latin and English countries), are written differently from Farsi and (Indian) Hindi digits.

Fig. 5. (a) Arabic digits (used mainly in Latin and English countries), (b) Farsi digits (used mainly in Iran), (c) Indian (Hindi) digits (used in most Arabic countries).

In the next sections, we briefly explain some of the state-of-the-art techniques and developments in off-line,

and on-line Farsi script recognition systems.

4. OFF-LINE FARSI SCRIPT RECOGNITION

The earliest work in off-line Farsi script recognition was conducted by two Iranian researchers: Parhami and Taraghi [6]. Their system was demonstrated by analyzing newspaper headlines; sub-words were segmented and recognized according to features such as concavities, loops, and connectivity. This section briefly reviews some of the research efforts towards off-line Farsi script recognition based on the types of features used for recognizing isolated Farsi digits or letters: structural features, and statistical features.

4.1. Structural Features

Structural features are those incllude loops, branch-points, endpoints, and dots, which are based on the instinctive aspects of writing. One of the reasons that structural features are more common for the recognition of Farsi scripts is that the primary shape of many Farsi letters are alike and only their number of dots and the locations of their dots are different. Therefore, to differentiate such letters, structural features are used for capturing dot information explicitly.

In a sample paper found in the literature ([13]) structural features are used. After thinning, each digit is segmented into lines. Then for each half of the line, average and variance of the coordination differences of points on the line are calculated separately for directions of x, and y. Therefore, for each line, 8 features are extracted which are then normalized using their maximum values. These features are invariant to scaling and location, but are sensitive to rotation. This way the feature vectors extracted for the digits could have a length of 16 to 72. To equalize the length of feature vectors, PCA method was used to make the length of feature vector equal to 75. The result of 94.44% is attained using 2240 samples for training, 1600 samples for testing.

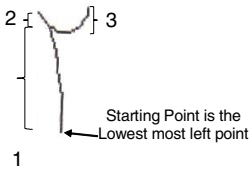

Fig. 6. Structural features are extracted from the skeleton of digit '2'.

4.2. Statistical features

Statistical features are numerical measures, computed over the images or regions of the images. Some examples of statistical features used in Farsi script recognition include: pixel densities, geometric moments, Zernike moments, Pseudo Zernike moments, Wavelets coefficients, histograms of chain code directions, Fourier descriptors, and features extracted from profiles. In Figure 7, samples of statistical features extracted from outer profiles, projection histograms, and crossing counts are shown, which were used in [3], [8], and [15].

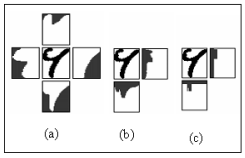

Fig. 7. (a) Outer profiles from four directions, (b) projection histograms from two directions, (c) Crossing counts from two directions.

In a study on Farsi/Arabic handwritten isolated digit recognition [8], authors generalized the method in [3], they used profiles features, also they combined with three other types of features including: Crossing counts, Projection histograms, and Size feature (for distinguishing zeros from other digits). These features are shown in Figure 8. They used different orientations for the profile features, and they tested three different sizes using Support Vector Machines (SVM) as classifiers. They collected their own database of samples which has 5000 training samples and 4000 testing samples of handwritten digits written by 90 writers. The best results that they reported was 99.57% recognition rate with 257 features, on their collected database, using SVM classifiers with RBF kernels.

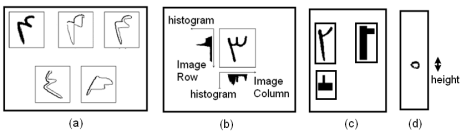

Fig. 8. A combination of four different features used in [8]: (a) profiles, (b) histograms, (c) crossing counts, (d) size (for distinguishing zeros from other digits).

In another paper [12] the idea of showing Farsi digits using a 12-Segment display pattern was employed in order to extract features. Figure 9 depicts the pattern. Although this pattern does not look like to be an appropriate segmentation pattern, it is utilizable. The proposed pattern has one degree of freedom (control point). Figure 10 shows how the control point can move. Using moment of inertia and center of gravity of each image in the training set, and weight functions for each digit, an optimized control point model was found. For extracting features, after a 12-Segment pattern is mapped on the image matrix in the bounding box, the control point's location is extracted based on the described model. Now to form the feature vector, for each segment the number of black pixels is extracted, and divided by the whole number of pixels in that segment (which makes it invariant to scaling). For training the neural network, 230 samples from 23 writers were used, and for testing the recognition system, 500 samples

from 50 writers were fed into the system. The achieved recognition rate was 97.6%.

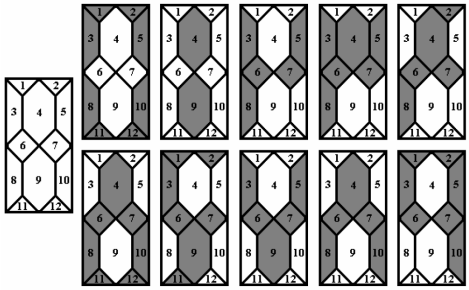

Fig. 9. Segmentation pattern and mapping of all Farsi digits from 0 to 9 on to the pattern.

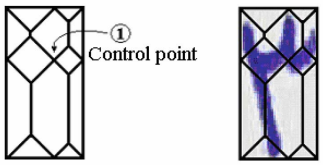

Fig. 10. Adjusting the control point for a digit.

4.3. Combining classifiers for off-line Farsi Script Recognition

In a system described in [9], authors used a combination of three Neural Network classifiers for recognition of isolated handwritten Farsi digits. The architecture of their systems has been shown in Figure 11. First, they extracted 81 features for each digit based on concavity features, then they reduced the dimensionality of the feature vectors to 15 features, using a PCA (Principal Component Analysis) transformation. They trained three Multi Layer Perceptron (MLP) Neural Network with a different number of neurons in their hidden layers, as three different experts (classifiers). Then they trained another neural network as the combiner of the decisions of these three classifiers, and producing the final decision of the system. Their experiments on 2430 Farsi digits (1000 samples for training experts, 900 samples for adjusting the combiner, and 530 samples for testing the complete system) show 87%, 85%, 83% recognition rates of for the experts, and 91% recognition rate for the complete system (including combiner of the experts) on the testing set.

5. ONLINE FARSI SCRIPT RECOGNITION

In a recent attempt in the area of online Farsi handwriting recognition, Razavi et al. [10] designed a system for isolated character recognition. Isolated Farsi letters were divided into 11 classes based on the number of dots, shapes and locations of their diacritic marks, as shown in Figure 12.

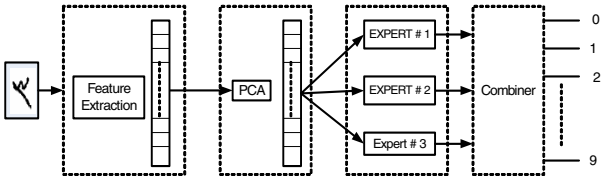

Fig. 11. A system which uses a combination of three experts (classifiers) for the recognition of isolated Farsi digits.

Group Number	Letters in the group	complementary part
1	آ	Tilda-shape upper stroke
2	ع ل م و ه ی ا ح د ر س ص	none
3	ب ج	One point below
4	خ ذ ز ض غ ف ن	One point above
5	ت ق	Two points above
6	ث ژ ش	Three points above
7	پ چ	Three points below
8	ک	One slanted bar above
9	ی	Two slanted bars above
10	ط	One vertical bar
11	ظ	One vertical bar and one point above

Fig. 12. Farsi isolated characters Divided into 11 groups

In order to recognize an input pattern in the suggested system, first, the complementary parts and their locations are recognized by two multi-layer perceptron neural networks. If the character belongs to a group which has more than one member, the each character's main shape will be analyzed by another neural network for further recognition. For this purpose, for any of groups 2 to 8 (in Figure 12), a three-layer neural network is used with 20 and 10 neurons in the input layer and hidden layer, respectively. The characters main shape, after size normalization is represented in ten points. These coordinates are considered as features in the second stage of recognition. In the methodology used, the first group is discriminated from the rest of the groups based on the aspect ratio and the number of strokes. The reported recognition rate was 93.9% (for 4,144 letters) using the database reported in [11]. Both complementary and main shape recognition had misclassifications; 198 and 53 cases respectively. Figure 13 shows some misclassified cases due to the mistakes made by the main shape recognizer. This method assumes that the character's main body is written in the first stroke and also fixed numbers of strokes are assumed for each character. These assumptions, although valid in most of the cases, are not always guaranteed. Using the number of strokes as a discrimination criterion decreases the robustness to the writing variations even in the scope of isolated characters.

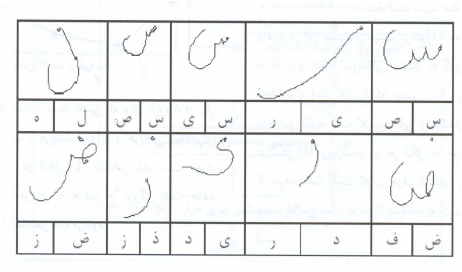

Fig. 13. Erroneous cases due to mistakes in main shape recognition.

6. SEGMENTATION OF TOUCHING DIGITS

A method for segmentation of handwritten numeral strings in Farsi presented in [5], is briefly described here. This method is based on combining features from the foreground and background of the image. The method uses an new algorithm for extracting features in the foreground and background which is called skeleton tracing. Also for finding background features it combines vertical top and bottom projection profiles and their skeletons. Based on a combination of this information, and some global information from the string image, this algorithm tries to construct the best segmentation paths for separating the touched digits. Figures 14, and 15 show an example of this algorithm for feature extraction in the foreground and background.

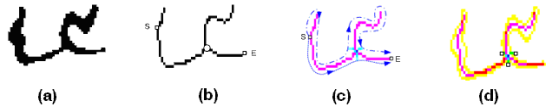

Fig. 14. (a) Original image (4 touching 3), (b) Original skeleton, starting point, and ending point are depicted by S, and E, respectively, (c) From starting point (S), skeleton is traversed in two different directions (clockwise: dashed arrows, and counter-clockwise: dotted arrows) to the end point (E), (d) Mapping of intersection points on the outer contour by bisectors to form foreground-features.

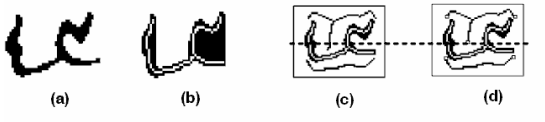

Fig. 15. (a) Pre-processed and slant corrected image, (b) Background region, (c) Top/Bottom background skeletons, (d) Top/Bottom-background-skeletons after removing skeleton parts which are lower/higher than the middle line.

After finding all the feature points in the foreground (denoted by in Figure 14-d), and background features (denoted by in Figure 15-d), this algorithm assigns feature points from top to bottom, or from bottom to top, alternatively together to construct all possible segmentation paths (denoted by dashed lines in Figure 16.

Fig. 16. Feature points on the background, and foreground (from the top and bottom) are matched, and assigned together to construct segmentation paths.

This algorithm was able to correctly segment 75% of the testing samples in a testing set of 150 touching Farsi digits collected by the authors.

7. DATABASES FOR FARSI SCRIPT RECOGNITION

Standard databases are essential requirements of research and development in the field of off-line and on-line handwriting recognition. Recently, some databases were released for Arabic handwriting recognition such as CENPARMI Arabic Cheques in 2003 [14]. Among the latter two, CENPARMI's database consists of legal and courtesy amounts on bank cheques, and isolated handwritten digits. Since the digits in Arabic and Farsi are alike; CENPARMI's Indian digits and Arabic courtesy amounts databases can also be used for Farsi handwritten numeral recognition. Until recently, there was no standard Farsi database available for researchers; however, very recently two standard databases have been developed for research on Farsi off-line handwritten recognition in [15] and [16]. In [15] authors collected and generated a comprehensive standard databases in Farsi language which consists of 6 sets including isolated Farsi digits (1888 samples), isolated Farsi letters (11900 samples), Farsi numerals (7350 samples), Farsi legal amounts (8575 samples), Farsi dates (175 samples) and a small set of English digits (3500 samples) written by 175 different writers in Iran. This database also includes both gray scale and binary versions of these sets. The only online Farsi database was developed in 2004 by contribution of 124 writers [10]. One thousand of the most frequent sub-words were extracted from the online archives of some newspapers in Iran, containing more than 300,000 words as candidates for data collection. We hope these databases will become popular and will be used by researchers in the area of Farsi script recognition. Also, we hope more comprehensive databases which cover all aspects of Farsi script will be soon available for research on both off-line and on-line Farsi recognition.

8. CONCLUSION AND CHALLENGES AHEAD

This paper reviews the evolution of Farsi (Persian) script and its relationship to Arabic script. It discusses similarities and dissimilarities between these two scripts from the

OCR (Optical Character Recognition) point of view. A brief review of some of the state-of-the-art techniques in off-line and online Farsi script recognition/segmentation, as well as available databases are also presented. There are still many unsolved problems in Farsi and Arabic script recognition systems. The main problem is the lack of standard database with huge number of samples of words, letters, digits for off-line and on-line Farsi script recognition. So far, developed methods have been mainly tested on the small and private databases collected by the authors and have not always been available to others. We hope introducing new developed databases can fill this important gap in this research area. Cursiveness of Farsi/Arabic handwriting makes segmentation of words and numerical strings a very difficult challenge. There has been little research in the segmentation of handwritten words and numerals in Farsi. A large part of the work in both off-line and online recognition consists of isolated character recognition, which is hard to generalize for cursive word recognition and related real world applications. Finally, most of the research in Farsi script recognition during the past 10 years has been presented in Iran in Farsi local journals/conferences, and the international research community does not know very much about those research works and their achievements. The lack of comprehensive surveys on Farsi recognition (recognition methodologies, segmentation, databases, etc.) sometimes yields to repetition of the same research by different groups in different labs. We hope the results of the efforts in Farsi and Arabic script recognition will be unified and will be made available internationally in order to yield high performance recognition systems for these two important scripts.

9. REFERENCES

[1] "Wikipedia, The Free Encyclopedia," Wikimedia Foundation Inc. Aug. 2006 http://en.wikipedia.org/wiki/Main_Page

[2] C. Y. Suen, S. Izadi, J. Sadri, F. Solimanpour, "Farsi Script Recognition-A Survey", Proceedings of International Summit on Arabic and Chinese Handwriting, Sept. 2006, University of Maryland, USA, pp. 101-110.

[3] J. Sadri, C. Y. Suen, T. D. Bui , "Application of Support Vector Machines for recognition of handwritten Arabic/Persian digits," Proceeding of the Second Conference on Machine Vision and Image Processing & Applications (MVIP), Vol. 1, Feb. 2003, Iran, pp. 300-307.

[4] J. Sadri, C. Y. Suen, and T. D. Bui, "A new approach for segmentation and recognition of handwritten numeral strings," Proceeding of IS&T/SPIE International Symposium on Electronic Imaging (EI), Vol. 5657, USA, Jan. 2005, pp. 92-100.

[5] J. Sadri, C. Y. Suen, and T. D. Bui, "Segmentation of handwritten numeral strings in Farsi and English languages," Proceedings of the Third Iranian Conference on Machine Vision and Image Processing & Applications (MVIP) 2005, Vol. 1, Feb. 2005, Tehran, Iran, pp.305-311.

[6] B. Parhami and M. Taraghi, "Automatic recognition of printed Farsi texts," Pattern Recognition, Vol. 14, No. 1-6, 1981, pp. 395-403.

[7] Y. S. Chen and W. H. Hsu, "A new parallel thinning algorithm for binary image," Proceedings of National Computer Symposium, 1985, pp. 295-299.

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

[9] S. H. Nabavi Kahrizi, R. Ebrahim Pour, E. Kabir, "Application of Combining Classifiers in the Recognition of Isolated Farsi Digits", Article in Farsi, Proceedings of the Third Iranian Conference on Machine Vision and Image Processing & Applications (MVIP) 2005, Vol. 1, Feb. 2005, Tehran, Iran, pp.115-119.

[10] S. M. Razavi and E. Kabir, "Online Persian Isolated character recognition," The Third Conference on Machine Vision, Image Processing & Applications (MVIP) 2005, In Farsi, Vol. 1, Tehran, Iran, Feb, 2005, pp.83-89.

[11] S. M. Razavi and E. Kabir, "A Data base for Online Persian Handwritten recognition," 6th Conference on Intelligent Systems, In Farsi, Kerman, Iran, 2004.

[12] A. Harifi, and A. Aghagolzade, "A New Pattern for Handwritten Persian/Arabic Digit Recognition," International Journal Of Information Technology, Volume 1, Number 4, 2004, pp.293-296

[13] M. Zeyaratban, K. Faez, S. Mozzafari, M. Azvaji, "Presenting a new structural method based on partitioning thinned image for recognition of Handwritten Farsi/Arabic numerals," Proceedings of the Third Conference on Machine Vision, Image Processing and Applications, In Farsi, Vol. 1, Iran, Feb 2005, pp.76-82.

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

[15] F. Solimanpour, J. Sadri, and C. Y. Suen, "Standard databases for recognition of handwritten digits, numerical strings, legal amounts, letters and dates in Farsi language, Proceedings of the 10th Intl Workshop on Frontiers in Handwriting Recognition (IWFHR), Oct 2006, France, pp. 3-7.

[16] S. Mozzaffari, K. Faez, F Faradji, M. Ziaratban, and M. Golzan, "A Comprehnsive Isolated Farsi/Aarabic Charachter Database for Handwritten OCR Research", In the Proceedings of the 10th Int'l Workshop on Frontiers in Handwriting Recognition, Oct. 2006, France, pp. 385-389.