

Writer Identity Recognition and Confirmation Using Persian Handwritten Texts

Aida Sheikh ^{1*}, Hassan Khotanlou ^{2*}

1. Department of Computer, Qazvin Branch , Islamic Azad university, Qazvin, Iran.

Aida.sheikh@gmail.com

2. Department of Computer, Bu-Ali Sina University, Hamedan, Iran.

Hassan.khotanlou@gmail.com

Abstract

There are many ways to recognize the identity of individuals and authenticate them. Recognition and authentication of individuals with the help of their handwriting is regarded as a research topic in recent years. It is widely used in the field of security, legal, access control and financial activities. This article tries to examine the identification and authentication of individuals in Persian (Farsi) handwritten texts so that the identity of the author can be determined with a handwritten text. The proposed system for recognizing the identity of the author in this study can be divided into two main parts: one part is intended for training and the other for testing. To assess the performance of introduced characteristics, the Hidden Markov Model is used as the classifier; thus, a model is defined for each angular characteristic. The defined angular models are connected by a specific chain network to form a comprehensive database for classification. This database is then used to determine and authenticate the author.

Keywords: *Persian handwriting recognition, authentication, Off-line, Hidden Markov Model*

1. Introduction

In this era, security and information protection is considered a big challenge of humanity in the modern world. Identifying writers from their handwriting has recently become a considerable and interesting subject in identity recognition. Among behavioral features, handwriting is easily achieved and studies show that different individuals have different handwritings [1,2]. Therefore, identifying and confirming the identity of individuals using their handwriting has been a research focus in recent years and its major application is in security and legal issues, controlling systems access, and financial activities. The identity recognition problem aims to specify the identity of the writer, given a handwritten text and the identity confirmation problem aims to specify whether two given handwritten texts have been written by one individual. Generally, graphology experts analyze and investigate handwritten texts. Although human intervention in solving this problem is an effective approach, it is costly and tiring due to the nature of human beings.

Most conducted studies in writer identity recognition and confirmation have been offline and signature-based online methods are more common in identity recognition. Most studies in identity recognition focus on the English language and there has been few studies regarding Arabic and Persian handwritten texts in comparison to the English language. Accordingly, this paper briefly studies the widely used methods in Latin and Persian texts and compares these methods and application for the Persian language. Finally, an online method is introduced and presented for identity recognition and confirmation for Persian texts.

2. Previous Works

Reviewing conducted studies and papers shows that they mostly investigate writer identity recognition based on handwritten texts and a limited number of methods are introduced for determining and confirming identity. This may be because introducing an efficient solution for identity recognition can also include a reliable approach for identity confirmation. In what follows, the most important proposed methods are introduced in the domain of handwriting recognition.

Horizontal and vertical views have been widely used in identifying English and Persian texts. Accordingly, Zois et al. [3] propose a text dependent method for identity recognition. In this method, the vertical view histogram of a word's image is processed and the considered features are extracted using morphology operators.

Bensifa et al. [4] investigate identity recognition in form of a text-based information retrieval problem. In this method, features are retrieved based on the type of the used data and thus, this method is easily generalized to other languages, including Arabic, Persian, etc.

Reference [5] proposes a method for identity recognition and confirmation, which is based on Markov's hidden model. In this model, for each writer, an HMM model is considered and each corresponding model is trained using samples of the writer's handwriting. Therefore, each HMM model works as an

expert, who is able to recognize the handwriting of a certain individual.

Reference [6] introduces a method for writer identity recognition that extracts features from the image of the handwritten text using edges information. The introduced features indicate direction changes during the writing process. In fact, the main notion of this method is that the writing process constitutes of a set of pen hits and the specifications of this process can be achieved by extracting the edges information.

In a paper by author [7], a new method is proposed for feature extraction from handwritings. One of the proposed methods is the n feature extraction method that extracts features indicating straight lines and curves used in handwritings.

The aforementioned methods are the basic propose algorithms in handwriting recognition. Table 1 briefly presents some of the writer identity recognition papers using different algorithms.

Tabel1: writer identity recognition papers using different algorithms

<i>Author's Name</i>	<i>Database</i>	<i>Feature Extraction</i>	<i>Classification Methods</i>	<i>Language-Accuracy</i>
M.Bulacu [8]	650 Writers	Edge based directional PDFs as features (Textural and allograph prototype approach)	K-nearest neighbor and a feed forward neural network classifier	English- 92%
Neils et al. [9]	43 Writers	Allograph prototype matching approach using the dynamic time warping (DTW) distance function	af-iwf (allograph frequency inverse writer frequency) measure	English- 60%
B.Helli, al. [10]	100 Writers	Point-based (speed, acceleration, vicinity linearity, vicinity slope), stroke-based	Tey proposed an LCS (longest common subsequence) based classifier	Persian- 95%
Soleymani Baghshah [11]	128 Writers	Fuzzy approach	Fuzzy learning vector quantization	Persian- 95%
Helli, B., Moghaddam [12]	100 Writers	XGabor filter	Weighted Euclidian distance	Persian-77%
Sadeghi [13]	50 Writers	Fuzzy clustering & gradient features	MLP (multi layer perceptron)	Persian-77%
Golnaz Ghiasi [14]	180 Writers	Codebook	2-fold cross validation, Leave-oneout cross validation	Persian-93%
Schlapbach et al. [15]	100 Writers	X-Y coordinates	Hidden Markov Models	English- 96%
Bensefia et al. [16]	88 French Writers 150 English Writers	A textual based Information Retrieval model, local features such as graphemes extracted from the segmentation of cursive handwriting	Cosine similarity	French- 95% English- 86%
Rafiee and Motavalli [17]	100 Writers	Baseline and width structural features	Feed forward neural network	Persian-95%

3. The Structure of the Recognition System

Handwriting as a behavioral feature is very dependent on brain performance and it is affected by several

complex factors, like training, physical status, environmental and mental conditions of the individual. Therefore, in writer identity recognition, features should be discussed that indicate the differences between

different handwritings. In addition, Arabic and Persian handwritings have several differences from Latin handwriting. Persian and Arabic handwritings are inherently continuous and overheads and underlines prevent the application of methods that require full segmentation. The shapes of Persian and Arabic alphabets are a function of their location in the word and there may be different forms of each alphabet in different locations, including the beginning, middle, and ending of the word or as a separate alphabet. Some Persian alphabets have one, two, or three dots above or below them. Persian and Arabic handwritings have many print fonts and handwriting styles.

3.1. The Proposed Algorithm

Since our goal is to propose an automatic writer identity recognition method and there are no constraints in the considered handwritings, we cannot use methods that require automatic and full segmentation of text into words and alphabets. The general structure of this method includes four stages: collecting the handwriting samples, preprocessing and create the binary image, feature extraction, writer identity recognition or confirmation using pattern recognition methods. The proposed method considered the handwritten text as an image. This method is evaluated using the handwritings of 70 individuals and results indicate that the proposed method is highly efficient for Persian handwritten texts.

3.1.1 Sample Collection

In order to evaluate the proposed method and the other investigated approaches, 70 individuals with different education, age, and gender were asked to fill the designed forms with their normal handwriting. These individuals were selecting from ordinary people. For evaluation, the individuals were asked to write two determined texts, which are in one paragraph and include various words. The determined texts are presented in figure 1 and 2.

align a straight line. After collecting the forms, all were scanned as grayscale images with resolution 300dpi.

3.1.2 Preprocessing and create binary images

One of the stages that is performed in most image processing systems is binarization. Transforming a grayscale image into a binary one has several advantages. The size of a binary image is far smaller than a grayscale one and since, one of the important characteristics of personality recognition systems for handwritten texts is their speed, and the smaller size of the binary image facilitates and accelerates all stages. Therefore, create binary image reduces image sizes, as well as removing redundant information.

3.1.3 Feature Extraction

In order to propose an efficient method for writer identity recognition, features should be considered that indicate the differences of different handwritings. In this method, angular features (in 8 angles, 0, 20, 40, 60, 80, 100, 120, and 140 degrees) are first extracted and compared for each image page.

$I(x, y)$ is the resulting image from preprocessing, which is a binary image. First, closing morphological operations are applied to binary image I . g_i is the i -th image resulting from closing operations with constructing element S_i , which has the same dimensions as the input image I . I : the binarized input image. S_i : the i -th component. Components are direct lines with different angles. Figures 3 to 5 presents an example of angular specifications extracted from each image.

As we can see, for most angles, these specifications have different patterns for different images and by inferring these differences, features can be introduced that can recognize the handwriting. Some of the usable features include 1- the number of remaining black objects in the image resulting from angle extraction operation. For instance for a 40 degrees angle, it is clear that the image has more black spots. 2- The length of the remaining objects from the feature extraction operation. 3- The total area of black locations of the image resulting from the feature extraction operation.

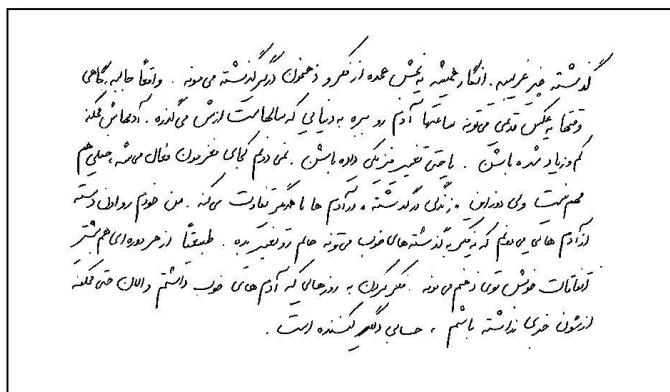


Figure 1. The image of the first handwritten text

The forms do not have lines; however, the individuals were asked to write along straight horizontal lines. Nonetheless, it was observed was some forms had tilted lines; in most forms, however, the lines were almost

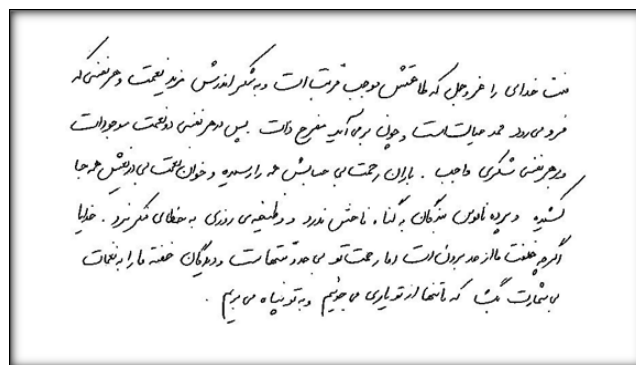


Figure 2. The image of the Second handwritten text

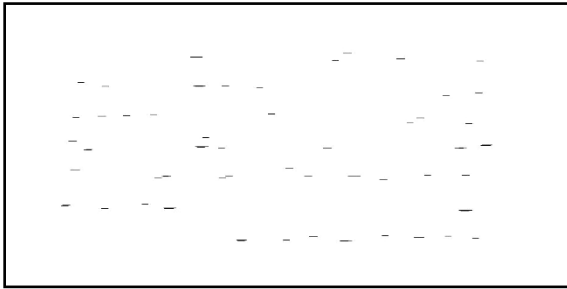


Figure 3. Angular specifications of 0 degrees for the image of a handwriting

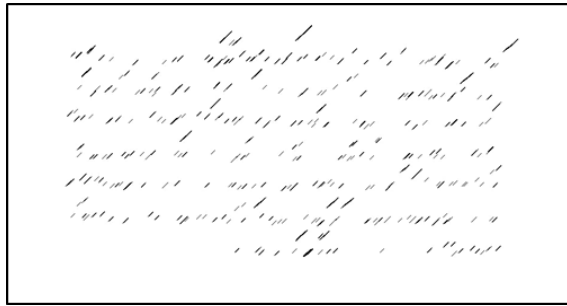


Figure 4. Angular specifications of 40 degrees for the image of a handwriting

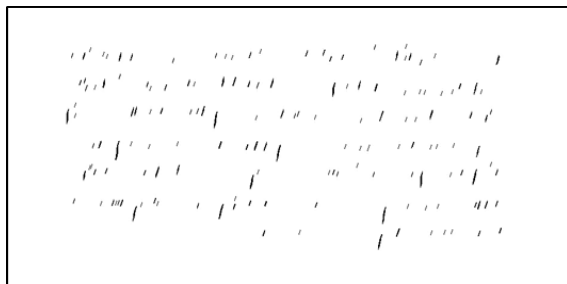


Figure 5. Angular specifications of 80 degrees for the image of a handwriting

Subsequently, the angular specification of each image is extracted for different angles of the number 2 handwriting of the collected samples as figure 2. Moreover, we can also use a combination of the aforementioned features. For the feature extraction stage, the number of length of the remaining black objects is used as the component. This feature is computed and compared for the handwritings of two individuals that there are two different handwritings from each individual. As we can see in table 2, the difference of features between the handwritings of an individual is smaller than that of two different individuals.

After the pre-processing stage, feature extraction is performed for each image extracted from angular specifications, separately with components in form of lines with different lengths, which are considered 3 to 40 in this research, so that the observation sequence (feature vector) of each handwritten text is achieved. Finally, this observation sequence is sent to a chain network to recognize the identity of the writer; however, it should be normalized before sending it to the classification stage. Moreover, when the discrete hidden Markov model is

used for classification, the input values should be discrete. Therefore, a quantization operation is used. A feature vector with integer values in range [0, 100] is considered in the implemented identity recognition system for each angular specification.

Table 2. Calculating features for the handwritings of two individuals

Angles	Number of components remained from First Handwriting of first person	Number of components remained from Second Handwriting of first person	Number of components remained from First Handwriting of Second person	Number of components remained from First Handwriting of Second person
0	34	50	64	74
20	166	163	147	159
40	199	214	146	166
60	104	101	35	40
80	83	79	29	33
100	25	23	13	11
120	5	7	4	7
140	15	15	10	8

4. Proposing a Model for Angular Specification

The proposed writer identity recognition system is divided into two main sections used for training and testing. In order to increase accuracy and security, the constructed models are embedded in a chain network, which presents a credible text image for angular specifications. In this research, each angular specification is considered as a separate model. The discrete left to right hidden Markov model [18] with a forwarding state is used to implement each angular specification model. In order to train the model, it is necessary to compute the related parameters, $\lambda = \{A, B, \pi\}$, using equations 1, 2, and 3. Therefore, it is only necessary for the considered angular specification after the feature extraction stage and providing the feature vector (observation sequence o_i) to use the three equations above to extract the parameters of the model.

$$\bar{\pi}_i = \gamma_1(i), 1 \leq i \leq N \quad (1)$$

$$\bar{a}_{i,j} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i,j)}, 1 \leq i \leq N \quad 1 \leq j \leq N \quad (2)$$

$$\bar{b}_j(k) = \frac{\sum_{t=1, o_t=v_k}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)}, 1 \leq j, \quad 1 \leq k \leq M \quad (3)$$

4.1 Feature Extraction

The Baum Welch algorithm is used to train HMM parameters [18]. It means that $\lambda = \{A, B, \pi\}$ is used to present each angular specification model. The initial state distribution of the models are $\pi = \{\pi_i\}$, where $\pi_1 = 1$ and $\pi_i = 0$ for $1 < i \leq N$. Moreover, N is the number of states in the corresponding model. For the transition

probability matrix $A = \{a_{i,j}\}$, initially $a_{i,j} = a_{i,i+1} = a_{i,i+2} = 0.5$ for $1 \leq i < N$ and $a_{i,j} = 0$ where $j \neq i + 1$, $j \neq i$, and $j \neq i + 2$. Moreover, $a_{N,N}$ is equal to one.

4.2 Proposing a Chain Network for Training Models

A chain network of angular specifications is used to increase the accuracy and security of the writer identity recognition system's output and constrain the search space. The chain network is divided into a set of sub-models $\{D_1, D_2, \dots, D_n\}$ based on the angles. Subsequently, each model is indexed as $SWD_{k,i}$ based on its parts. Sub-models are embedded in a chain network called $WN_{k,i}$, which presents the text images, and they are connected through specific paths. The performance of the identity recognition system for searching a writer on the chain network is as follows. First, using the method, which was explained in the feature extraction method, the observation sequence $O_s = [o_1, o_2, \dots, o_k]$ is obtained for the image of a handwritten text.

In this equation, $p(SW_i|O_i) = p(O_i|SW_i)p(SW_i)/p(O_i)$ [18]. Moreover, for simplicity, we assume that all parts of the text image occur with the same probability along the chains' path. Therefore, $p(O_i)$ is the same for all parts of the image text and the problem is reduced to maximizing $p(O_i|SW_i)$, which can be effectively computed by the Viterbi Algorithm for network $WN_{k,i}$. Finally, equation 4 is used to search writer w in D_k .

$$W = \arg\max \prod_{i=1}^k P(O_i|SW_i, WN_{k,i}) \quad (4)$$

5. Implementation of the Proposed Algorithm

In this research, all simulation is performed using Matlab R2013a on windows 7 operating system and 2.30GHz Intel Core i5 processor. Results of extracting the components based on angular specifications show that for Persian handwritten texts, features are about the same throughout the text. For instance, for a 40 degrees angle, images with maximum component have different pixel lengths. After the pre-processing stage, for each image extracted separately from the angular specification with line components and different pixel lengths, which is considered 3 to 40 in this research, feature extraction is performed to provide the observation sequence (feature vector) corresponding to each handwritten text.

Subsequently, for each writer, a left to right HMM is considered in which the initial node is zero degrees and the end node is 140 degrees with a 20 degrees step. After the feature extraction stage, for the first handwriting of each writer, the state change probability matrix and symbol observation probability matrix is created, which is considered the input of the HMM. Parameters are trained using the Baum-Welch algorithm on the extracted features. In these computations, the maximum learning iteration is considered 100. In order to classify a sequence to one of the K classes, K HMM models are

trained for each class. Subsequently, LL values are computed using the HMM Toolbox [18] of MIT. If the i -th model is the most similar, the sequence belongs to class i . In statistics, the log likelihood function is a function of statistical model's parameters that plays a key role in inferential statistics.

In order to run the propose algorithm to recognize and confirm the identity of the writer, a specific handwriting of a writer is selected; the second handwriting of the same writer is then compared using HMM Toolbox. Subsequently, the second handwriting of another writer is selected and its LL measure is computed. In this section, the output of the proposed identity recognition system is compared under different circumstances. In these experiments, from the total 70 samples, 50 are selected for training and the remaining 20 are selected for testing. This comparison is performed for different training conditions, including the parameters of components' angles, training duration, etc. the precision is computed based on the percentage of correctly recognized samples divided by the total number of recognitions. Table 3 presents these results. As we can see, the highest precision rate of identity recognition for the training data is resulted when the number of component's angle changes is equal to 7 or 12 and the component length is equal to 9 or 10. Whereas, for the test data, the number of component's angle changes is equal to 12 and the component length is equal to 9.

Table 3. Comparison of results for different HMM training conditions

number of component's angle changes ($\Delta\theta$)	component length	Learning Time (Second)	precision rate of identity recognition	
			Learning Data	Test Data
4	17	804	99%	63%
7	13	444	98%	75%
10	10	320	98%	78%
12	9	266	94%	81%
15	8	213	89%	77%
17	7	177	84%	74%
20	6	160	78%	73%
25	5	124	71%	65%
30	4	106	61%	57%

6. Evaluation

Since, there has been no standard database for Persian handwriting texts and the relevant published papers each have a different database, the Arabic handwriting database in reference [20] is used to evaluate the proposed algorithm.

This database consists of 4 pages in Arabic of 1000 forms of handwritings, which are collected from 1000 female and male writers in the country with different ages, 15 to 70. About 65% of these forms are filled by individuals with ages in range 16 to 25. For each individual, there are four images of handwritings with

different contexts in resolutions 300 and 600dpi. Reading signs and symbols are rarely observed in these texts.

After scanning the forms and performing pre-processing operations, the researchers, who have collected this database [19], have transformed them into binary images using the Otsu thresholding algorithm and they have also removed the noise. The tilt modification operations are performed before extracting paragraphs from the images. After identifying the possible lines, the X-Y coordination of each extracted line and their regression is obtained to compute their slopes. At the next stage, paragraphs are extracted. Since, there are many dots and overheads between the lines in Arabic, the subsidiary components are first removed and only the primary components remain (in the reference paper, this part is called Part of Arabic Word (PAW)). The image is then divided into several vertical strips. The lines of each strip are identified and finally, dots and overheads are added to the strips. The words of each line are identified by finding the white spaces between words. Moreover, the division positions are identified by findings a hypothetical space more than 10 pixels. At the next stage, the text inserted in the forms is evaluated in three phases. This is performed by human personnel. In the performed research, results of 4808, 938, and 966 lines are respectively dedicated to training, validation, and test sets. The handwritten text is recognized using discrete left to right HMM with Bakis topology and using HTK. In fact, the shape of each character is considered as a class during training. The identified characters with similar shapes are placed in one class. In sum, 149 classes are considered. Several statistics features are extracted from the lines images using the sliding window technique, where windows are considered with different lengths. However, the best result is achieved by the sliding window with length 4 and shared space 2. After the extraction, features are quantized in the linear codebook using top-down clustering. Table 4 presents the recognition rate results of the algorithm in reference [19], the multi-layer perceptron neural network, and the proposed algorithm.

Table 4. Comparison of the precision results of the proposed algorithm on the handwriting database

<i>Methods</i>	<i>Training</i>	<i>Test</i>
	<i>Acc. (%)</i>	<i>Acc. (%)</i>
Reference [19]	51.4	51.73
MLP	28	34
Proposed algorithm	61.5	58.7

Since this research aimed to propose a dynamic (smooth and automatic) method and no limitations are considered for the type of the investigated handwritings, we do not consider methods that require the automatic segmentation of the text into words and letters. The characteristic of the proposed method is compatible with the structure of Persian handwritten texts.

This method is introduced as an approach to text-dependent identity recognition. Of course, considering the images resulting from angular specifications, it is clear that these features are about the same throughout the text and thus, this method can also be used as an independent method.

7. Conclusions

Feature extraction is one of the determining stages of increasing the identity recognition rate in identity recognition systems. Extracted features should indicate the different of an individual's handwriting from that of others and there should also be minimal difference between handwritten texts of an individual. This research employs a combination of edge direction distribution and edge axis distribution features to extract features, as well as the hidden Markov Model. Angular specification models are connected through a specific chain network to increase the accuracy and security of the identity recognition system. Moreover, a database provided by 70 different writers was used to train and evaluate the system.

Considering the results, the advantage of the proposed algorithm is as follows:

- 1- Using a combination of edge direction distribution and edge axis distribution features at the feature extraction stage.
- 2- The autonomy of the proposed identity recognition system, particularly at the feature extraction stage (since, manual feature extraction is overwhelming and time-consuming).
- 3- Increasing the accuracy and credibility of the identity recognition system due to using the chain network at the training stage.
- 4- Eliminating the need for the segmentation stage at the feature extraction stage.

Since the writer identity recognition problem is performed for the first time using Persian handwritten texts with the proposed feature extraction method, the results of this paper are promising. Moreover, the proposed features are global features and the results can be improved by extracting and combining different features.

References

- [1] S.N. Srihari, H.Arora, S.H. Cha and S.lee, "Individuality of handwriting," Journal of Forensic Sciences, vol.47, no.4, pp.1-17, 2002.
- [2] S.N. Srihari, H.Arora, S.H. Cha and S.lee, "Individuality of handwriting: a validation study," IEEE Proc. Of 6th Int. conf. on Document Analysis and Recognition, pp.106-109, 2001.

- [3] E.N.Zois,V.Anastassopoulos, "Morphological waveform coding for writer identification," Pattern Recognition, vol.33, pp.385-398, 2000.
- [4] A.Bensifa, T.Paquet and L. Heutte, "Handwritten document analysis for automatic writer recognition," Electronic Letters on Computer Vision and Image Analysis, vol.5, no.2, pp.72-86, 2005.
- [5] A.Schlapbach and H.Bunke, "Using HMM-based recognition for writer identification and verification," IEEE Proc of 9th Int. Workshop on frontiersin handwriting Recognition, pp.167-172, 2004.
- [6] M. Bulacu, L.Schomaker, "Writer style from oriented edge fragments," proc.10th Int. Conf. Computer Analysis of images and Patterns, pp. 460-469, 2003.
- [7] Golnaz Ghiasi, Reza Safabakhsh, "An Efficient Method for Offline Text Independent Writer Identification" International Conference on Pattern Recognition, 2010.
- [8] Bulacu, M., Schomaker, L., Vuurpijl, L. Writer identification using edge-based directional features, in: Seventh International Conference on Document Analysis and Recognition (ICDAR).2005.
- [9] Niels, R., Gootjen, F. Vuurpijl, L. "Writer Identification through Information Retrieval: TheAllograph Weight Vector," in International Conference on Frontiers inHandwriting Recognition, pp. 481-486, 2008.
- [10] B. Helli, M.E.Moghaddam, "A text-independentPersian writer identification system using LCS based classifier", in: IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2008.
- [11] Soleymani Baghshah, M., Bagheri Shouraki, S. Kasaei, S. a novel fuzzy classifier using fuzzy LVQ to recognize online Persian handwriting, in: Second IEEE Conference on Information and Communication Technology (ICTTA), 2006.
- [12] Helli, B.Moghaddam, M.E. Persian writer identification using extended Gabor filter, in: International Conference on Image Analysis and Recognition (ICIAR), 2008.
- [13] R.Sadeghi .S.Moghaddam, M.E. Text-independent Persian writer identification using fuzzy clustering Approach, in: International Conference on Information Management and Engineering (ICIME), Malaysia.2009.
- [14] Golnaz Ghiasi, Reza Safabakhsh, "An Efficient Method for Offline Text Independent Writer Identification" International Conference on Pattern Recognition, 2010.
- [15] Schlapbach, A., Bunke, H. Using HMM based recognizers for writer identification and verification, in: Proceedings–International Workshop on Frontiers in Handwriting Recognition, IWFHR, Tokyo, pp. 167–172, 2004.
- [16] A. Bensifa, T. Paquet and L. Heutte, "A Writer identification and verification system," Pattern Recognition Letters, vol.26, no.13, pp.2080-2092, 2005.
- [17] Rafiee, A., Motavalli, H. Off-Line Writer Recognition for Farsi text. In: 6th Mexican International Conference on Artificial Intelligence, Special Session,pp. 193-197, 2007.
- [18]http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm_usage.html
- [19] KHATT: S. A. Mahmoud, I. Ahmad, W.Al-Khatib, M. Alshayeb,M. Tanvir Parvez, V. Märgner, G.A. Fink, "An open Arabic offline handwritten text database", Pattern Recognition, vol. 47, no. 3, pp. 1096–1112, March 2014.
- [20] <http://khatt.ideas2serve.net/index.php>

Aida Sheikh received her B.Sc. in computer engineering from department of computer Engineering, Hamedan Branch, Islamic Azad University, in 2007, and her M.Sc. in computer software engineering from department of computer, Qazvin Branch, Islamic Azad University, in 2015.Her research interests includes Image processing, knowledge management and intelligent systems.

Hassan Khotanlou is an associate professor in department of computer at Bu-Ali Sina University, Hamedan, Iran. He received his B.Sc. degree in computer engineering from IUST University in 1995, and his MSc. degrees in artificial intelligence engineering from Shiraz University in 1997 and his Ph.D. degrees in artificial intelligence engineering from Pierre & Marie Curie University (Paris VI – TELECO) in 2007.His main research interests are Image processing, Fuzzy Systems, Acoustic, Data Mining, Computer Networks, Medical Image processing and Artificial Intelligence.