2012 International Conference on Advanced Computer Science Applications and Technologies

A New Approach to Feature Selection in Handwritten Farsi/Arabic Character Recognition

Mohammad Amin Shayegan and Chee Seng Chan
Department of Artificial Intelligence
Faculty of Computer Science and Information Technology
University of Malaya, 50603 Kuala Lumpur, Malaysia
mashaygan@siswa.um.edu.my

***Abstract* -- Feature extraction and feature selection are very important steps in pattern recognition systems. However, finding an optimal, effective, and robust feature set is usually a difficult task. In this paper, with the use of a comprehensive study on offline handwritten Farsi/Arabic digit recognition systems, a set of well-known features were extracted. Then, by employing one- and two-dimensional spectrum diagrams for standard deviation and minimum to maximum distributions, an optimal subset of initial features set was selected automatically. Experimental results, according to one of the biggest standard handwritten Farsi digit datasets, the HODA, had shown 95.70% accuracy with the proposed method. The achieved results showed a salient improvement in system precision in comparison to using other state-of-the-art approaches.**

Keywords: Farsi/Arabic Handwritten OCR, Feature Extraction and Selection, Principal Component Analysis, Spectrum Diagram.

I. INTRODUCTION

Pattern Recognition (PR) is one of the most challenging branches in the field of artificial intelligence, and Optical Character Recognition (OCR) has been one of the more attractive branches that researchers are faced in the recent years. Among the various stages in pattern recognition systems, Feature Selection (FS) and Feature Extraction (FS) play a vital role in recognition and also system accuracy [1]. An efficient FE method is usually invariant to different affine transformations such as scaling, translation, rotation, and skewing [2].

Generally, Principle Component Analysis (PCA) is one of the most common approaches [3] in feature selection operation. PCA is a classical statistical linear transform which has been widely used for different pattern recognition applications such as data compression [4], face recognition [5], as well as character recognition [6]. However PCA suffers from high computational cost. In addition, PCA generates a non-orthogonal feature space, and finding the order of most effective to least effective features in this space is not possible.

In Farsi/Arabic character recognition issue, state-of-the-art works usually find a set of features heuristically and employ them in pattern recognition systems. However, there is no direct way to identify that how much the proposed features are correlated to each other.

In this paper, we proposed a new approach to perform feature selection for handwritten Farsi/Arabic character recognition. Empirical results with the largest handwritten Farsi dataset – the HODA – have shown the effectiveness of the proposed method.

The rest of the paper is structured as follows. Section II reviews the related works in this area. Section III discusses our proposed method. Section IV shows the experimental results and finally section V draws the conclusions of this work.

II. RELATED WORKS

A. Feature Extraction

The Feature Extraction (FE) process is one of the most important and critical stages in any PR system such as OCR. The output of this stage directly influences the performance of the next stage, i.e. recognition [1]. In a PR system, the objects of interest are often represented by a set of numerical features with a goal to remove the redundancy from them. Extracting appropriate and robust features is the most difficult and basic point in an OCR systems like other PR applications,

A feature vector for a class should be insensitive to noise, scale changes, rotation, and other probably changes related to patterns as much as possible. Feature vectors should not be similar, redundant and repetitive.

Until now, a large number of various features have been introduced by researchers for printed/handwritten offline/online OCR systems. Usually, features are categorized into global transformations [7], structural [8], statistical [9], and template matching and correlation [10].

B. Feature Selection (FS)

Many different features can be found or calculated for an object in a PR system. However, it is possible some of the features will be correspond to very small details of the patterns, or maybe some will be the combination of other features (non-orthogonal features). Irrelevant or redundant features may degrade the recognition results and reduce the speed of learning algorithms significantly [11]. Hence, following the feature extraction process, Feature Selection (FS) issue arises which reduces problem dimensionality. FS is typically a search problem for finding an optimal subset with *m* features out of the original *M* features [12].

978-0-7695-4959-0/13 $25.00 © 2013 IEEE
DOI 10.1109/ACSAT.2012.77

Based on removing strategies, feature selection methods are categorized into 3 groups. The first category is the Sequential Backward Selection (SBS). In this approach, features are deleted one by one and the systems' performance is measured to find feature performance. However, finding the order of feature deletion one by one is very important. It means that a system's derived efficiencies after deleting features A, B, and C are not equal to the same system's derived efficiencies after deleting the features in order A, C, and B, or B, C, and A, and so on [13].

The second group comprises the random search methods such as Genetic Algorithms (GA). GA is a random method which uses the global evolution theory for problem solving. Azmi et al. used GA in a handwritten Farsi OCR system [11]. The initial number of features in their system was 81 and the recognition rate was 77%. After applying GA, the number of features was reduced to 55 and the recognition rate ascended from 77% to 80%.

The main problem concerning GA methods is that they always select chromosomes one by one with the best recognition percentage, and move this chromosome (feature) to the next stage. However, it is possible that when a good characteristic feature gets combined to another feature, the overall performance will not be as good as the individual performances.

The third method for feature selection is Principle Component Analysis (PCA), a statistical method that has been applied to find important patterns in high-dimension input data [6]. Gesualdi and Seixas employed PCA along with neural networks for license plate recognition [4]. They used PCA for data compression in the feature extraction part. They reduced the number of features from 30 to 4. Reported results showed that the achieved accuracy for digits recognition was acceptable, but the accuracy for characters recognition was degraded significantly.

Ziaratban et al. proposed a novel statistical description for the structure of isolated Farsi handwritten characters [14]. They used a PCA algorithm to reduce and equalize the feature vectors' lengths. A 93.15% recognition rate was achieved on a dataset with 7647 test samples.

To recognize handwritten Arabic isolated letters, Abandah et al. extracted 95 features from all feature categories [6]. After that, only the first 40 features were selected from the PCA process result. Finally, 5 different classifiers were employed and 87% accuracy was achieved on average in the best case.

III. THE PROPOSED METHOD

A. Introducing One-Dimensional Standard Deviation (1D_SD) Spectrum

In the 1D_SD method, a spectrum line corresponding to a specific feature is drawn from mean-SD to mean+SD for each class. For example, in Farsi/Arabic digit recognition, there are 10 classes corresponding to 10 digits 0 to 9. Hence for each feature from initial feature set, a 1D_SD diagram was plotted with 10 spectrum lines corresponding to digits 0 to 9. **Fig. 1_a** and **1_b** show the 1D_SD distribution diagrams corresponding to the '*Maximum Vertical Crossing Count'* and '*Aspect Ratio'* feature for Farsi/Arabic digits, respectively. In **Fig. 1_a**, all spectrum lines are in an overlapping range (2.5, 8), meaning that the '*Maximum Vertical Crossing Count'* feature cannot discriminate existing classes from each other in the feature space. In **Fig. 1_b**, the spectrum line corresponding to class (digit) 1 is completely separate from other spectrum lines, indicating that the '*Aspect Ratio'* feature can completely discriminate class 1 from other classes. Therefore it can be considered as a candidate feature in the final features vector.

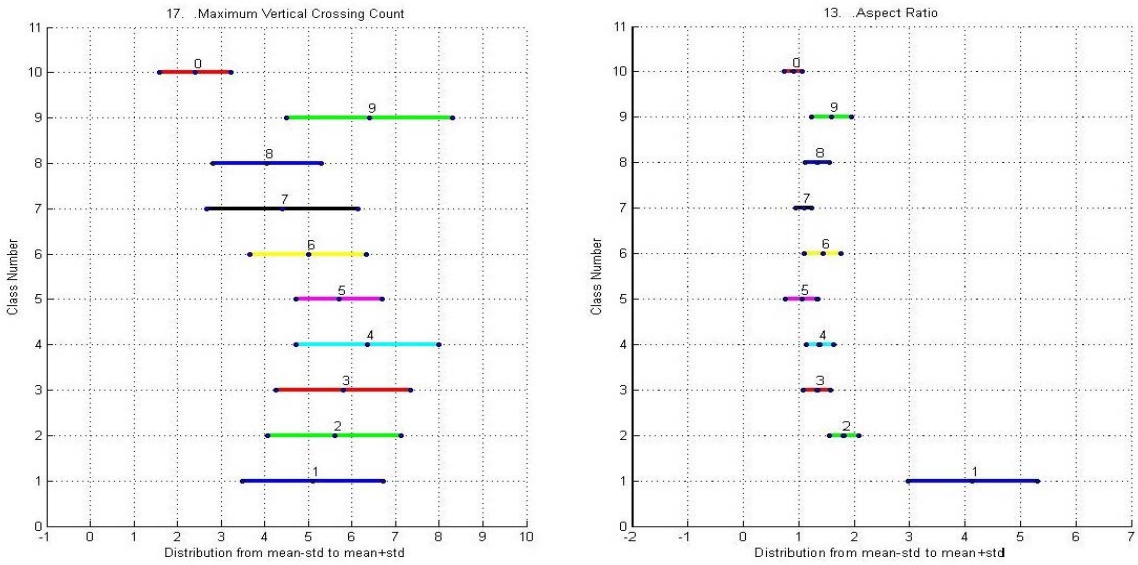
B. Introducing One-Dimensional Minimum to Maximum (1D_MM) Spectrum

In the 1D_MM plot, a spectrum line corresponding to a specific feature is drawn from the minimum to the maximum value of that specific feature for each class. A shorter spectrum line concerning a specific feature indicates that the existing samples in a particular class have more similarity to each other with respect to that feature. Hence a shorter spectrum line is better than a longer one. In addition, a distribution diagram with farther class centers (locations of classes' means) is better than one with closer class centers. In this case, a classifier separates existing clusters better.

Nevertheless, finding a set of separated spectrum lines using only 1D_SD distribution diagrams is not enough to create an optimum feature vector, because the outlier samples in each class are not placed in the range of 1D_SD spectrum lines but they are put in the 1D_MM range. **Fig. 2** illustrates 1D_MM spectrum lines for the same feature, *'Aspect Ratio'* from min to max. It is obvious that some samples of class 1 overlap with some samples in classes 2 and 9. In other words, in the recognition phase, it is possible for these samples in class 1 to get misclassified into class 2 or 9 and vice versa, if only the *'Aspect Ratio'* feature is employed.

C. Introducing Two-Dimensional Standard Deviation (2D_SD) and Two-Dimensional Minimum to Maximum (2D_MM) Spectrums

Similar to the 1D_SD distribution diagram and 1D_MM spectrum, Two-Dimensional Standard Deviation (2D_SD) distribution diagram and Two-Dimensional minimum to maximum spectrum for two features are made by mapping one feature on the X axis and another feature on the Y axis. In these cases, an ellipse (or rectangular) is plotted for each couple of features. For example, in 2D_SD, the values of the main ellipse diagonals (or the length and width values in the rectangular case) are mean-SD to mean+SD for two features. As such, the [n*(n+1)/2] 2D_SD distribution diagram can be generated for *n* independent features.



a) 'Maximum Vertical Crossing Count' feature **b) 'Aspect Ratio' feature**

Figure 1. 1D_SD distribution diagrams for Farsi/Arabic digits

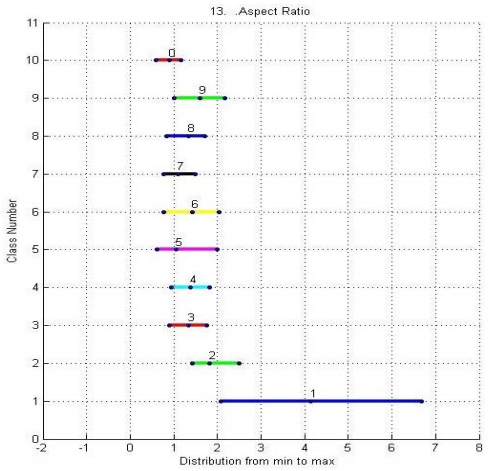


Figure 2. 1D_MM distribution diagrams for 'Aspect Ratio' feature

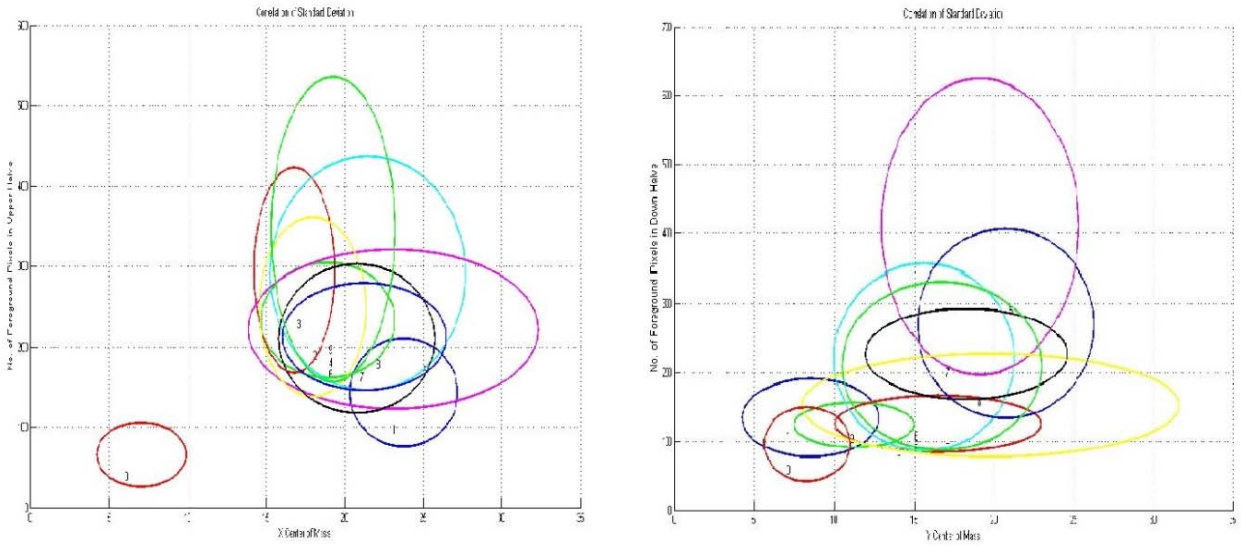
Fig. 3_a shows a 2D_SD distribution diagram for two features, namely *'X coordinate center of mass*' and '*No. of foreground pixels in upper half of image.* As it can be seen, the ellipse for class (digit) 0 is completely distinct from other ellipses. Hence the feature pair (X coordinate center of mass, No. of foreground pixels in upper half of image) is a good choice for membership in features vector (to distinguish class (digit) 0 from other classes (digits)).

Fig. 3_b shows another 2D_SD distribution for two features: '*Y Center of Mass*' and '*No. of foreground pixels in Upper half of image'*. It is completely clear in this case that the mentioned features are highly correlated, and therefore are not suitable couple features for membership in the features vector.

D. Feature Extraction

Due to the vast diversity in writing style, handwritten characters are placed in a high-dimension data category. Hence, finding an optimal, effective, and robust feature set for usage in the recognition phase of an OCR system is usually a complex task. In the first step and based on the literature, an initial feature vector S1 including 133 common features for Farsi/Arabic digit recognition process were extracted from pattern images. Some of these features are:

Aspect Ratio, Total Image Area, Image Perimeter, Diameter, Solidity, Thinness Ratio, Euler Number, Image Extent, Eccentricity, Center of Mass (COM), Centroid distance, Pixel distribution density in up, down, left, and right halves, Pixel distribution density in upper and lower main diagonal, Density of different parts of the normalized image, The ratio of pixel distribution in different quarters of the normalized image to each other, Ratio of horizontal variance histogram to vertical variance histogram, Ratio of upper half variance to lower half variance of an image, Normalized horizontal and vertical transition, Maximum horizontal and vertical crossing counts, Average of multiplication distance X and Y from COM, Average of distance X and Y from boundary, Variance of horizontal and vertical histogram, Ratio of major to minor axis length, Convex area, Number and location of start, end, branch, corner, and crossing points in image skeleton, Holes, Normalized invariant moments to order 7, Discrete cosine transform coefficients related to the main image to order 9, Image's outer boundary, and all image profiles, Top, Down, Left and Right Profile Histograms, Top, Down, Left, and Right concavities in the image's skeleton, Horizontal and Vertical Projection Histograms, Number of modified horizontal and vertical transitions, Average distance and



a) **b)**

3_a) X Center of Mass' feature vs. 'No. of Foreground Pixels in Upper Half of image' feature
3_b) 'Y Center of Mass' and 'No. of Foreground Pixels in Lower Half of image'
Figure 3. 2D_SD distribution diagrams for Farsi/Arabic digits

Average angular distance of each foreground pixel in a sub-image from a virtual origin.

E. Feature Selection

Using all samples in the training part of each class, the value of an specific feature was defined as $f_k(S_{i,j})$ where f_k is the value of k'th feature from initial feature vector and $S_{i,j}$ represents the j'th sample in class i . After that the min, max, mean, and standard deviation for all features in initial feature vector were computed.

To find the initial reduced subset of the initial features set S1, the 1D_SD distribution diagram along with the 1D_MM spectrums were employed. The output of this stage was a reduced version of feature vector S2 with 94 features, which satisfied the criteria necessary for membership in the final features vector. Finally, by using 2D_SD distribution diagrams and also 2D_MM spectrums on the reduced version of feature set S2, a feature vector S3 with 58 members was selected as the final features vector.

The following are among the selected features in set S3: 'X Coordinate of Center of Mass', 'No. of Foreground Pixels in Upper half', 'No. of Foreground Pixels in lower half', 'Ratio of Foreground Pixels to Area of Bounding Box', 'Ratio of number of foreground pixels upper main diagonal to number of foreground pixels under main diagonal', 'Aspect Ratio', 'Normalized Horizontal Transition', 'Maximum Horizontal Crossing Count', 'Normalized Vertical Transition', 'Variance for vertical histogram', 'Equiv_Diameter', 'Solidity', 'Perimeter', 'Ratio of Major to Minor Axis Length', 'Convex Area', 'Number of end points', 'Number of end points in lower half', 'Number of end points in different zones of image bounding box', 'Number of branch points', 'Some Discrete Cosine Transform Coefficients such as (1,1), (1,2), (1,4), (1,5), (2,1)', 'Some Discrete Cosine Transform Coefficients of Image Profile such as (1,4), (2,3), (2,5), (3,3)', 'Some Discrete Cosine Transform Coefficients of outer boundary such as (2,1), (2,7), (3,3), (3,4)'.

IV. EXPERIMENTAL RESULTS

A. Choosing a Dataset

This research was conducted on Farsi/Arabic digit recognition specifically. Hence, in order to test the effectiveness of the proposed method, the digit part of one of the biggest Farsi handwritten standard datasets was chosen, namely HODA [15]. The HODA dataset has two parts: digits and characters. The digit section of the HODA dataset was prepared in 2007 by extracting the digits' images from 11,942 registration forms in Iran. Those forms were scanned at 200dpi in 24bit color format. The digits were extracted from the *postal code, national code, record number, identity certificate number,* and *phone number* fields on each form. **Fig. 4** shows sample digits from this dataset.

Figure 4. Some sample digits in the HODA dataset

The digit section of the HODA dataset is divided into two parts, namely training and testing. There are 6000 and 2000 samples for each digit in the training and testing parts of this dataset, respectively.

B. Pre-processing

The performance of an OCR system heavily depends upon the quality of original data. Hence, before applying any operations, some important pre-processing operations such as noise removal, dimension normalization, skew and slant correction using common powerful techniques are first performed on the prototypes. We applied a median filter with 3x3 window [16] and also morphological opening and closing operators ((1) and (2) in order) using dilation and erosion operators ((3) and (4)) for noise removal [17].

$$A \square B = (A \ominus B) \oplus B \quad (1)$$

$$A \bullet B = (A \oplus B) \ominus B \quad (2)$$

$$A \oplus B = \{c / c = a + b , a \in A , b \in B \} \quad (3)$$

$$A \ominus B = \{ c / c = (B)_c \subset A \} \quad (4)$$

A : initial image,
B : unit structural element 2x2

Fig. 5 demonstrates an example of filtering, opening and closing processes on four of Farsi digits (2, 5, 6, and 7). Also, we used our own method and that of [18] to connect the broken image segments.

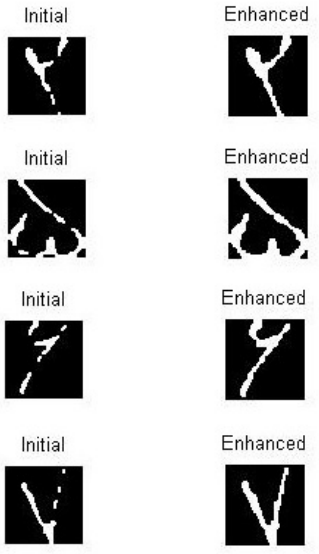


Figure 5. Applying morphological filtering and connectivity operation on Farsi/Arabic digits

Without any changes to the aspect of image ratio, the size of each image was normalized, to 50 pixels and the image was located in the center of a 50x50 bounding box. In order to correct the slant angle of each image, Hanmandlu et al.'s method was used [19].

C. Feature Selection

Several experiments were carried out to test the effectiveness of the proposed method for the feature selection operation. In all experiments, a multi-layer perceptron network with back propagation was trained with 94 (or 58) neurons in the input layer (corresponding to the number of features in sets S2 and S3), 30 neurons (found experimentally) in the hidden layer, and 10 neurons (corresponding to 10 different classes of digits 0 to 9) in the output layer, respectively.

In the first experiment, neural network was employed with 94 (number of features in set S2) neurons in the input layer. The network was trained with only 3000 samples of each class (total of 30,000 samples from the training part of the HODA dataset) and then tested with 1000 samples for each class from the testing part of the HODA dataset (total of 10,000 samples). This operation was repeated 10 times and finally, 92.60% accuracy was achieved on average in this stage.

To compare the performance of the proposed method against other well-known techniques such as PCA – as state of the art for feature selection operation - a general PCA technique was applied on the initial reduced feature set S2 with 94 features. PCA changed the order of features in the new orthogonal feature space and generated a new re-ordered feature set S4 with the same 94 features. The complete reordered feature set S4 was fed into the same MLP_NN and a final accuracy of 94.04% was achieved. This result was 1.44% higher than the previous result, portraying the superiority of the PCA technique for feature selection.

In the second experiment, the system was trained with the proposed final version of feature set S3 with only 58 features. The correct recognition rate increased from 92.60% to 95.12% on average, clearly indicating the superiority of reduced feature set S3 with the proposed method against the initial reduced feature set S2 with 94 features. To find the superiority of feature set S3 compared with other subsets of S2 which have 58 members, we made set S5 with the first 58 members of set S4 generated by PCA. The recognition rate declined dramatically from 94.04% to 89.00%. This result obviously shows the effectiveness and superiority of our proposed technique in comparison with other available feature selection techniques. In **Table 1** the experimental result outcomes are given.

TABLE 1. Recognition rate corresponding to different feature vectors

Feature Set	Number of Features in Feature Vector	Feature Selection Method		Accuracy
		Proposed Method	PCA Method	
S2	94	*		92.60%
S4	94		*	94.04%
S3	58	*		95.12%
S5	58		*	89.00%

D. Result Comparison

Some researchers from literature used PCA for feature reduction [14]. However, based on OCR application comparison, the results of the proposed approach were compared only with [20], because the same dataset and recognition engine were used in both research works. Enayatifar et al. used 48 features and achieved 94.30% accuracy only for the 3000-sample test. Our method used 58 features and reached 95.12% accuracy for a 10,000-sample test (more than 3 times that of [20]). In another experiment using only 3000 (equal to the number of samples in [20]) samples of testing set, the proposed approach achieved to 95.70% accuracy.

V. CONCLUSION

In this paper, a new method for feature selection was introduced and applied in a handwritten OCR application. At first, a big set of well-known common features – based on the literature - were extracted from the training patterns. Then, by using one and two-dimensional standard deviation distribution diagrams as well as one and two dimensional minimum to maximum distribution diagrams, spectrum lines were plotted for each feature separately. By choosing only the features whose spectrum lines are not overlapping at all (or less than a small threshold) for at least one of the available classes, a final reduced feature vector with 58 features was selected.

The mentioned algorithm was implemented in an OCR system for recognizing one of the biggest standard handwritten Farsi/Arabic digit datasets, HODA. Using an MLP-NN with back propagation technique as classifier, we achieved 95.12%, accuracy when only 1/2 of the training datasets (30,000 samples) and 1/2 of the testing datasets (10,000 samples) were used in operations. Also when only 3000 test samples were used, the recognition accuracy increased to 95.70%.

The accuracy was more than 6% compared to a similar experiment which used PCA as state of the art in the feature selection operation. The result clearly indicates the superiority of the proposed method for feature selection over other available techniques such as PCA.

According to the results, the proposed technique is completely effective for OCR application as a subcategory of PR systems. Although the results have been derived to use the proposed method in OCR application, this new method can also be used for other PR systems with different database types.

REFRENCES

[1] A.M. Al-Tameemi, L. Zheng and M. Khalifa, “offline Arabic Words Classification using Multi Set Features,” Information Technology Journal, 2011, pp. 1 - 7.

[2] G.A. Abandah and N. Anssari, “Novel Moment Features Extraction for Recognizing Handwritten Arabic Letters,” Journal of Computer Science, Vol. 5, No 3, 2009, pp. 226-232.

[3] F. Bouchareb, R. Hamdi and M. Bedda, “Handwritten Arabic character recognition based on SVM classifier,” 3rd International Conference on Information and communication Technologies : From Theory to Application, 2008, pp. 1-4.

[4] D.R. Gesualdi and J.M. Seixas, “Character recognition in car license plates based on principal components and neural processing,” Proceeding of VII Brazilian Symposium on Neural Networks, 2002, pp. 206-211.

[5] A. Bansal, K. Mehta and S. Arora, “Face Recognition using PCA & LDA Algorithms,” Second International Conference on ACCT, 2012, pp. 251-254.

[6] G.A. Abandah, Kh.S. Younis and M.Z. Khedher, “Handwritten Arabic Character Recognition Using Multiple Classifiers Based on Letter Form,” 5th International Conference on Signal Processing, Pattern recognition & Application, 2008, pp. 128-133.

[7] J.H. Al-Khateeb, J. Jiang, J. Ren, F. Khelifi and S.S. Ipson, “Multiclass Classification of Unconstrained Handwritten Arabic Words Using Machine Learning Approaches,” The Open Signal Processing Journal, Vol. 2, 2009, pp. 21-28.

[8] J. Shanbehzadeh, H. Pezashki and A. Sarrafzadeh, “ Feature Extraction from Farsi Handwritten Letters,” Proceeding of Image and Vision Computing, 2007, pp. 35-40.

[9] L.M. Lorigo and V. Govindaraju, “Offline Arabic handwriting recognition: a survey,” IEEE Transaction on PAMI, Vol. 28, No. 5, 2006, pp. 712-724.

[10] A.D. Rafaa and J. Nordin, “Offline OCR System for Machine Printed Turkish using Template Matching,” Journal of Advanced Materials Research, Vol. (341-342), 2011, pp. 565-569.

[11] R. Azmi, B. Pishgoo, N. Norozi, M. Koohzadi and F. Baesi, “A hybrid GA and SA algorithms for feature selection in recognition of handprinted Farsi characters,” IEEE International Conference on ICIS, Vol. 3, 2010, pp. 384-387.

[12] I. Guyon and A. Elisseeff, “An Introduction to variable and feature selection," JOURNALof Machine Learning Research, Vol. 3, No. 1, 2003, pp. 1157-1182.

[13] Y. Elglaly and F. Quek, “Isolated Handwritten Arabic Character Recognition using Multilayer Perceptrons and K Nearest Neighbor Classifiers,” 2011.

[14] M. Ziaratban, K. Faez and M. Ezoji, “Use of Legal Amount to Confirm or Correct the Courtesy Amount on Farsi Bank Checks,” ICDAR’07, 2007, pp. 1123–1127.

[15] H. Khosravi and E. Kabir, "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties," Pattern Recognition Letters, Vol. 28, No. 10, 2007, pp. 1133-1141.

[16] M. Dehghan, K. Faez and M. Shidhar, "Handwritten Farsi (Arabic) word recognition : A holistic approach using discrete HMM," Pattern Recognition, Vol. 34, No. 5, 2001, pp. 1057-10651.

[17] Sh. Alirezaee, M. Ahmadi, H. Aghaeinia and K. Faez, “An Efficient Restoration Algorithm for the Historic Middle-age Persian (Pahlavi) Manuscripts,” IEEE International Conference on Systems, Man and Cybernetics, vol. 3, 2005, pp. 2114-2120.

[18] M. Ziaratban and K. Faez, "Detection and compensation of undesirable discontinuities within the Farsi/Arabic subwords," International Arab Journal of Information Technology, Vol. 8, No. 3, 2011, pp. 293-301.

[19] M. Hanmandlu, K.R. Murali Mohan, S. Chakraborty, S., Goyal and D. Roy Choudhury “ Unconstrained handwritten character recognition based on fuzzy logic,” Pattern Recognition, Vol. 36, 2003, pp. 603-623.

[20] R. Enayatifar and M. Alirezanejad, “Offline Handwriting Digit Recognition by using Direction and Accumulation of Pixels,” International Conferenceon Computer and Software Modeling, l4, 2011, pp. 214-220.