A Hybrid Recognition System for Islamic Annotation and Historical Arabic Handwritten Manuscripts

Omar Balola Ali[1], Adnan Shaout [2]

[1] Sudan University for Sciences and Technology, Sudan

[2] The University of Michigan, Dearborn, USA

[1]Abuamer1975@gmail.com, [2]shaout@umich.edu

ABSTRACT

In this paper, a multi neural-fuzzy recognition system with two combined statistical features, to solve the recognition problem of Historical Arabic Handwritten (HAH) manuscripts will be presented. The first set of statistical features are center of mass, crosshair, outlier and blank ink histogram (CCOB). The second feature is the principal component analysis (PCA). The new method will use two stages (levels) which are based on two classifiers, one public and one private, according to the similar features among characters. In the first level, we built a public classifier to deal with all character groups. Each group contains characters with overlapped features. The public classifier classifies the characters in the segmented character data set, which is captured from HAH manuscripts to specified groups. In the first stage, the system was applied to 34 Arabic characters and achieved 97.15% recognition rate for the tested dataset. In the second level, we created a private classifier for each group to recognize and classify the characters within a group which achieved 99.34% recognition rate for the tested dataset using the two level model.

***Keywords*: Historical Arabic Handwritten, Islamic Annotation, Principal Components Analysis (PCA), Multi Neural-Fuzzy, Adaptive Neural Network Fuzzy Inference System (ANFIS).**

1. INTRODUCTION

Large collections of historical handwritten manuscripts are now available as electronic images. For these electronic manuscripts to be efficiently and effectively accessible to scholars, new techniques for automatic indexing, classifying and retrieving electronic manuscripts must be developed. Therefore, recently, automatic reading of handwritten text has become an important issue. The Arabic and Islamic civilization has for many centuries influenced many regions of the world with different cultures. This great civilization has left a huge collection of valuable historical HAH manuscripts that are of significant importance to scholars around the world (Al Aghbari and Brook, 2009).

Transliteration and understanding historical manuscripts are two challenging problems. This is especially the case with Arabic manuscripts, which exhibit a wide variety of handwriting styles (Gacek, 2009).

The handwriting recognition problem arouses great interest in researchers, since there is a high level of ambiguity and complexity in such kind of images, and because of the importance of Optical Character Recognition (OCR) in office automation and many other applications.

Recognition of cursive handwritten text is one of the most difficult cases in the domain of OCR. However, the large number of potential applications for cursive handwritten text makes it a very popular research subject (Somaya Al-Ma'adeed, 2004).

2. RELATED WORKS

Many classifier techniques have contributed to solving more complex problems in many areas including prediction, approximation and pattern recognition. In the HAH manuscripts, the classifiers that were used by researchers includes Neural Network (NN), Hidden Markov model (HMM) and others.

In (Farrahi Moghaddam et al, 2010), the authors extracted a dataset from an Arabic handwritten manuscript in two phases, one in an automatic way (using word spotting clustering), and the other by human experts. They used the stroke map (SM), the edge profile (EP), the stroke gray level (SGL), and the estimated background. They use a set of topological features adapted to document images. The first transformation (T1) assigns a set of features to each branch point (BP) of connected components (CC). A BP is connected to a loop which is connected to an end point (EP) and is connected to another. The second transformation (T2) generates topological features associated with EPs. It is set based on whether or not the EP is connected to a BP which is connected to another EP, and based on its vertical state with respect to its vertical component. They used a radial basis function (RBF) kernel as a classifier.

The study in (Al Aghbari and Brook, 2009) presented a technique to classify Historical Arabic handwritten (HAH) manuscripts. They used a hard copy HAH manuscript with 300 by 300 resolution scanner and transformed it into a digitized image. Then, applied four pre-processing steps on the digital image as follows: binarization, noise removal, smoothing and thinning. In segmentation, the study used the Naskh style technique. The feature extraction techniques used were of two categories; structural (such as connected part upper/lower profile and projection profile) and statistical (such as punctuation count and ration between punctuation and the main connect part). Their experiments were performed on a manuscript entitled" كشف اللثام عن وجه السلام ". The overall number of words in the dataset is about 2000. The average accuracy of the whole word technique was 79.5%.

In (Khorsheed, 2003), the authors presented a new method for off-line recognition of handwritten Arabic scripts. The method trains a single hidden Markov model (HMM) with the structural features extracted from the manuscript words. VQ was used to form a codebook of the symbols, dots, endpoints, branches and loops. This process implements the k-means clustering algorithm to partition 4615 feature vectors. Word samples used here belong to the manuscript entitled '' جمهرة النسب لابن الكلبي' ' . The average image size of a single character from this manuscript is 35 pixels in height and 24 pixels in width. The recognition rates of the proposed method are (72 to 78) and (81 to 97).

The study in (A. Alaei et al., 2010) described a technique for the recognition of Persian handwritten isolated characters with a two-stage SVM based scheme. In the first stage, similar shaped characters were characterized into eight groups which were used to obtain recognition results. In the second stage, selected groups containing more than one similar shape of characters were considered further for the final recognition. They used feature techniques based on the under- sampled bitmaps technique and modified chain-code direction frequencies. They computed 49 dimension features based on under-sampled bitmaps and 196 dimension chain-code direction frequencies. The system achieved an accuracy rate of 96.68%.

In (Abandah et al., 2008), the authors used principal component analysis to select the best subset of features out of a large number of extracted features. They utilized both parametric and non-parametric classifiers to determine the best set of features.

In this paper, a multi neural-fuzzy system with two combined statistical features to solve the recognition problem of HAH manuscripts will be presented. Fuzzy and neural networks are used. The system works by interpreting the fuzzy rules in terms of the neural network. Fuzzy sets are taken as weights, while fuzzy rules, and input and output variables are taken as neurons.

The organization of the paper is as follows: In section 3, the recognition process of the document images is presented, together with the pre-processing and feature extraction based on the CCOB and PCA feature techniques. In section 4, the classification stage is presented, while section 5 describes the approach architecture, the experiments conducted and the results obtained. Finally, the conclusions will be given in section 6.

3. THE RECOGNITION PROCESS

Three main processes are included in the proposed system; pre-processing, feature extraction, and classification, as shown in Figure 1. The system uses a multi-classifier system classification for handwritten isolated Arabic characters as shown in Figure 2. In the proposed system the characters that were segmented from a handwritten manuscript document will be used as shown in Figure 2.

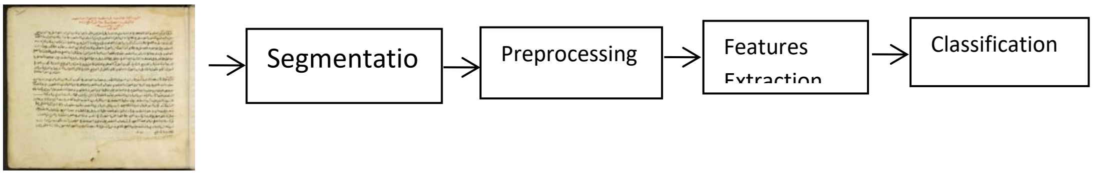


Figure 1. System phases for classification of manuscripts

The following will explain the processes involved in the classification for handwritten isolated Arabic characters:

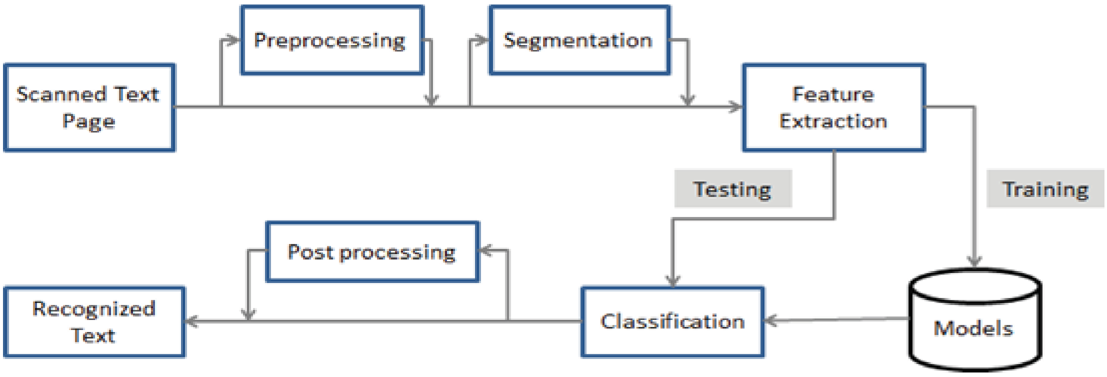


Figure 2. Multi-classifier system classification for handwritten isolated Arabic characters

3.1 Pre-processing

In this section binarization, morphology and cropping techniques for images of ancient Arabic documents will be described. A handwritten character will be sampled on A4 size paper. The characters will be scanned using a scanner with a resolution of 300dpi. These characters, are then segregated according to their own character group and stored as gray scale images. The following pre-processing techniques were used:

A. Binarization

Binarization is the process of converting a grayscale image (0 to 255 pixel values) into a binary image (0 and 1 pixel values) by selecting a global threshold that separates the foreground from the background. Each pixel is compared with the threshold value, and if it is greater than the threshold it is set to 1, otherwise it is set 0. Binarization is commonly reported to be performed either globally or locally (Anuradha and Koteswarrao, 2006).

The author in (R. Wisnovsky, 2004) utilized an adaptive method to compute a threshold value based on the values of image pixels by adding the minimum pixel value, min(p), and the maximum pixel value, max(p), in an image and divide the sum by 2.

B. Normalization

The process of changing the intensity value of the pixel to within the range of [0, 1] is called normalization in image processing. The conversion of various dimension images into fixed dimensions is also called normalization. Normalization is used in digital signal processing.

C. Morphological

The morphological process is used in image processing and involves the task of extracting image components that are useful in the representation and description of regional shapes. Morphology traces the outline of an object in a binary image. Nonzero pixels belong to an object and 0 pixels constitute the background. Figure 3 shows an example of the morphological process for the Arabic letter "j".

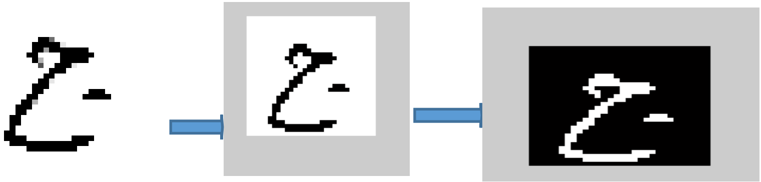
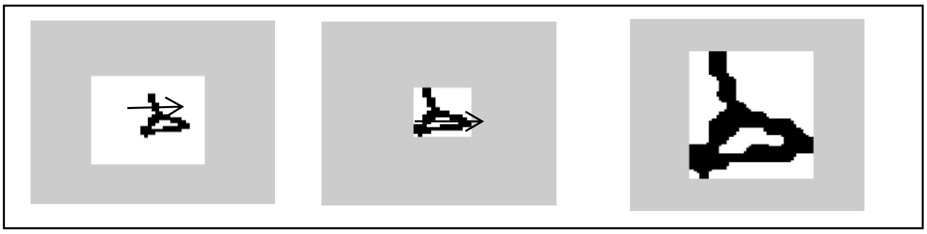


Figure 3. Illustrates a Morphological image after Binarization of grayscale

D. Cropping the Image

Cropping refers to the removal of the outer parts of an image to improve framing, to accentuate subject matter or to change the aspect ratio. The unwanted background parts of the image are omitted, leaving the focus only on the number. This makes the feature more accurate and accordant. The image is scanned row by row. The last set of rows from the top with all pixels corresponding to a value of "255" is obtained. Next, the last set of rows from the bottom with all pixels corresponding to a value of "255" is obtained. Following this process, the top and bottom parts of the image would have been cropped. This is followed by a step to crop the left and right sides of the image. Figure 4 shows a cropping example for the Arabic character 'Ta'.



A. Ta letter before cropped B. Ta letter after cropped C. Cropped Ta letter after resized

Figure 4. Grayscale character image cropping and resizing.

E. Resize the Image

This process is used to reduce an image size to a size smaller than the original, and to find the medial axis which defines as a set of pixels, S. Those pixels have an equal distance from the boundary pixels around it. The output of this process is a skeleton of the handwritten word. This process saves the geometry and the connections between the characters and the location of the original character.

3.2 Features Extraction

The purpose of feature extraction is to reduce the original data set by measuring certain properties, or features, that distinguish one input pattern from another pattern. Many character features based on cropped gray level and digitized image have been used in this work.

In our proposed system, we combine the feature extraction of two methods of two types.

3.2.1 First Set of Features (CCOB)

The first type of features are as follows: The area that is formed from the projections of the upper and lower parts as well as of the left and right character profiles is calculated. The center mass of the character image is represented as (xt, yt). The number of transitions and outliers (right, left, top and down) and black-ink histograms are determined. The first set of features are based on the extract mean of the resized cropped images, the calculated covariance's, computed eigenvectors and eigenvalues. Statistical features for extracting features of Arabic handwritten characters were used as listed in the following:

1. **Center of mass**, the center of mass feature, f_m , is the relative location (relative to the height and width of the image) of the center of mass of the black ink. The center of mass of the Arabic

letter ‘jeem’ is shown in Figure 5a. Given an image and a letter, where c and c’ are their centers of mass, respectively, the center of mass penalty used is pm = $1/(1 + d^{1/2}$ (c, c,)), where $d^{1/2}$ gives the square-root of the Euclidean distance (a commonly used measure).

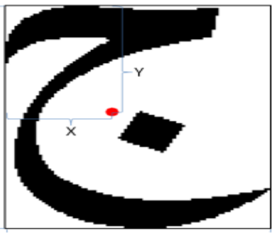


Figure 5a. The center of mass of the letter ‘jeem’ is marked by the (red) dot.

The center of mass feature is (X/W, Y/H).

2. **Cross feature**, counts the number of transitions from background to foreground pixels along the vertical and horizontal lines through the character image. Figure 5b shows the cross hair feature for the Arabic letter ‘jeem’.

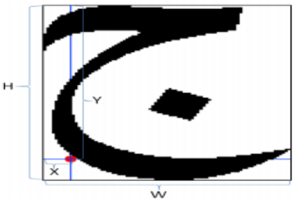


Figure 5b. The cross hair of the letter ‘jeem’ is marked by the (red) dot.

3. Outliers (Right, Left, Top and Down) calculates the distances of the first image pixel detected from the upper and lower boundaries of the image along the vertical lines and from the left and right boundaries along the horizontal lines. Figure 6 shows the top outline of the letter ‘jeem’.

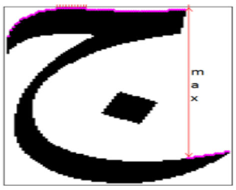


Figure 6. The top outline of the letter ‘jeem’ marked by the semi-led (purple) dots.

4. Black ink histograms (Horizontal and Vertical). Each image has a horizontal and vertical black ink histogram. The horizontal black ink histogram feature is represented by $f_h = (h_1, . . . , h_H)$, where H is the height of the bounding box of the black ink and is calculated as follows:

a. For i = 1, ..., H, let b_i be the number of black ink pixels in row i.

b. For i = 1, ..., H, let h_i be $b_i/ \max\{b_i\}$.

The vertical black ink histogram feature (f_v) is calculated in a similar manner.

3.2.2 Principal Components Analysis (PCA)

PCA is a very popular technique for dimension reduction. In (R. Sharma and Patterh, 2015), de-noised images are given to the next process in order to calculate the score values using the PCA technique. The score values obtained from the PCA techniques are then used by the Adaptive Neural Network Fuzzy Inference System (ANFIS) classifier for accomplishing the training process. Based on the predefined threshold value, the image under test is indicated as recognized or not recognized. In (G. A. Abandah, 2008), researchers proposed five Arabic handwritten character recognition classifier systems. They used 95 feature vectors, secondary components features, main body features, skeleton features, and boundary features; before applying the Principal Component Analysis (PCA).

In this paper, the combined statistical features, CCOB and PCA were used as a feature extraction technique. The process of PCA proceeds by taking the mean of the data matrix and subtracting the mean from the data matrix, followed by determining the Eigenvalues and Eigenvectors. In the last step, the desired number of Eigenvectors are selected as shown in Figure 7.

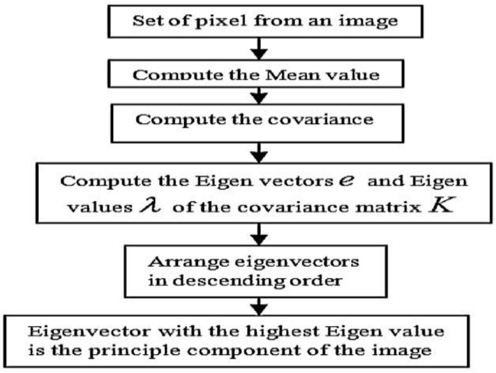

Set of pixel from an image
Compute the Mean value
Compute the covariance
Compute the Eigen vectors e and Eigen values λ of the covariance matrix K
Arrange eigenvectors in descending order
Eigenvector with the highest Eigen value is the principle component of the image

Figure 7. Flowchart of the principle component analysis technique.

4. CLASSIFICATION STAGE

The classification process is carried out at the final stage to recognize characters. It assigns an input character to one of many pre-specified classes which are based on the extracted features. For the classification process, ANFIS will be use in this paper with different algorithms. We will combine all feature vectors extracted from each image to a set of features (PCA and CCOB) for training and testing the data set. Figure 8 shows the proposed classification system with the combined feature steps.

4.1 ANFIS Architecture

The structure of the ANFIS consists of four inputs and a single output. The four inputs represent the different character features calculated from each image by using many combined statistical features and applied PCA features reduced to four features. Each of the training sets forms a fuzzy inference system with eighty one fuzzy rules. Each input was given three generalized triangular membership functions and the output was represented by three nonlinear membership functions. The outputs of the eighty one rules are condensed into one single output that represents the system output for that input image. The proposed NEURO-FUZZY topology is shown in Figure 9. Many experiments have been conducted to find the best values of the output (z).

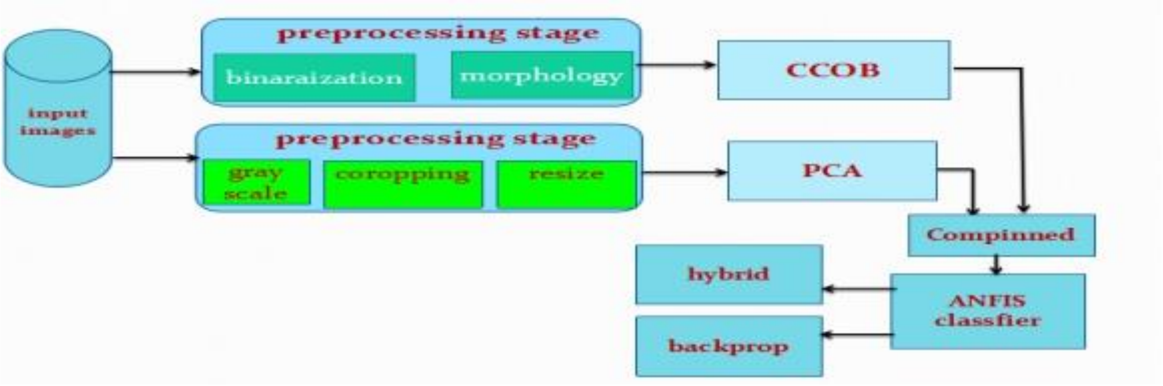


Fig 8. The proposed classification system steps with combined features.

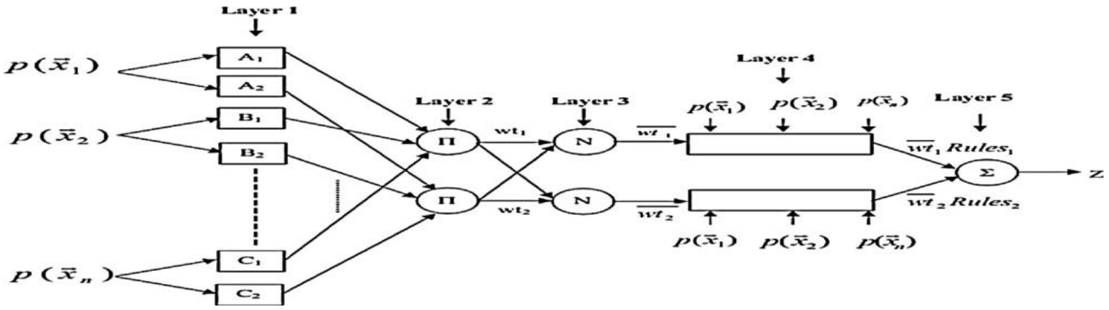

Layer 1
$p(\bar{x}_1)$
$p(\bar{x}_2)$
$p(\bar{x}_n)$
A₁
A₂
B₁
B₂
C₁
C₂
Layer 2
Π
Π
wt₁
wt₂
Layer 3
N
N
$\overline{wt}_1$
$\overline{wt}_2$
Layer 4
$p(\bar{x}_1)$ $p(\bar{x}_2)$ $p(\bar{x}_n)$
$p(\bar{x}_1)$ $p(\bar{x}_2)$ $p(\bar{x}_n)$
Layer 5
$\overline{wt}_1\,Rules_1$
$\overline{wt}_2\,Rules_2$
Σ
z

Fig 9. ANFIS architecture for Sugeno type reasoning.

4.2 Dataset

The dataset of handwritten Arabic characters prepared is as shown below in Figure 10. Those characters are scanned using a scanner and will be segregated according to their own character groups. A set of Arabic handwritten characters was selected concentrating on the basic Arabic letters, which consists of 28 characters and the related characters as (أ، ئ إ ء ،لا، ؤ).

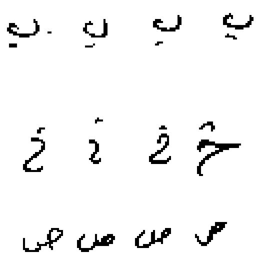


Figure 10. Characters samples for (ba), (kha) and (sad).

5. EXPERIMENTS AND RESULTS

The dataset (the set of Arabic handwritten characters) was divided into three levels as shown in Figure 11. The first level contained all characters in one class. The second level includes a group of the same characters in one class, which have generated fifteen different classes. The last level placed each character in a separate class called the letter class. For the experiments, we have selected 75% training and 25% testing datasets for all levels as shown in Table 1.

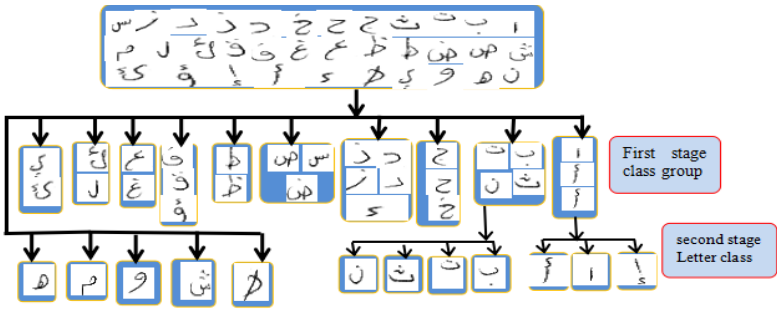


Figure 11. The proposed two-stage Arabic Handwritten character recognition system.

Table 1. Dataset types and sizes for training and testing.

Dataset Level	Dataset size	Training dataset	Testing dataset
All dataset with 34 characters in 34 classes	6800	5100	1700
Grouping dataset into 15 different classes	3000	2250	750
Similar characters (1 to 5) class for each character	Each character with 200 images	150*class number	50*class number

A two stage classifier was used in the proposed system. The first stage was based on features extracted from groups of similar characters. The second stage has sub-classifiers. A sub-classifier is a multiple adaptive neuro-fuzzy classifier to recognize the characters of only one group. In the experiments, we used an ANFIS classifier with hybrid learning algorithms. The handwritten character images were taken from the dataset. These images are firstly binarized before morphology is used. The second passed to the feature extraction techniques and had extracted four features as CCOB features. With the second set of features, the image characters were cleaned by removing its outer parts using the cropping technique. Thereafter, the unwanted background parts of the image were omitted, the focused was then on the number and resized the cropped images to 7 by 5. The two sets of features (4+35) were combined and reduced to four features using the PCA technique. Then, the feature values which were obtained from the PCA techniques were used by the ANFIS classifier for performing the training process.

Four features were used as inputs to the system, three numbers of triangular membership functions were also used, and a nonlinear output with 1000 epochs. A set of different experiments were performed on the training data set and the testing dataset for all types of dataset, as shown in Table 2. Table 3 shows the recognition accuracy results for any class that was obtained with ANFIS classifiers for grouped classes in the first stage.

The proposed system has improved the first stage, with a recognition accuracy of 97.15% for isolated Arabic handwritten characters in grouped classifiers, while a recognition rate of 99.34% was obtained for the second stage; an increase of about 3.5% from the recognition accuracy achieved using a single classifier system as shown in Table 2.

Table 2. The accuracy rate for the all datasets with a second set of features.

Dataset type	Training accuracy %	Testing accuracy%
All dataset	95.8167	95.8173
Grouping dataset	97.18	97.15
Similar dataset	99.6	99.34

Table 3. The accuracy for the grouped classes of characters.

Group number	Similar dataset	Training accuracy %	Testing accuracy%
Groups 1	أ إ ا	99.66	99.42
Group2	ن ث ت ب	99.34	99.21
Group3	خ ح ج	99.52	99.23
Group4	ء ز ر ذ د	99.21	99.40
Group5	ض ص س	99.59	99.32
Group7	ظ ط	99.63	99.13
Group8	غ ع	99.78	99.05
Group9	ؤ ق ف	99.51	99.35
Group10	ك ل	99.98	99.70
Group14	ي ئ	99.77	99.60

4. CONCLUSION

In this paper, a two-stage classifier system with combined statistical features based solution for HAH manuscripts recognition was proposed. The multi neuro-fuzzy recognition system was proposed to process Islamic annotation and HAH manuscripts. The dataset that was used in the experiments were isolated Arabic character sets applied after segmenting the manuscript documents. We aim to design other data sets that contains groups of similar letters to improve the recognition rate. The method that was described in this paper for isolated Arabic handwritten character recognition can be extended for other Arabic character positions by including a few other pre-processing activities.

The experiments used 34, 15 and (1 to 5) character classes. The highest accuracy result for testing a dataset with 15 classes was 97.15% and yielded 99.34% as the average recognition rate for similar characters as shown in table 2. The results of recognition rates shows that there was an improvement in performance using the proposed system. The proposed system will be used in the future on a dataset of handwritten Arabic letters for ancient manuscripts, with the expectation of similar recognition rates.

5. REFERENCES

Al Aghbari, Zaher, and Salama Brook. (2009) "HAH manuscripts: A holistic paradigm for classifying and retrieving historical Arabic handwritten documents." Expert Systems with Applications 36.8 (2009): 10942-10951.

A. Gacek. (2009) “Arabic Manuscripts: A Vademecum for Readers”, Handbook of Oriental Studies. Section 1 The Near and Middle East, 98. Leiden; Boston: Brill, 2009. ISBN-10: 90 04 17036 7.

Somaya A. Al-Ma’adeed, (2004) “Recognition of Off-line Handwritten Arabic Words”, thesis submitted to The University of Nottingham for the degree of Doctor of Philosophy, June 2004.

Farrahi Moghaddam, Reza, et al. (2010) "IBN SINA: a database for research on processing and understanding of Arabic manuscripts images." Proceedings of the 9th IAPR International Workshop on Document Analysis Systems. ACM, 2010.

Khorsheed, Mohammad S. (2003) "Recognizing handwritten Arabic manuscripts using a single hidden Markov model." Pattern Recognition Letters 24.14 (2003): 2235-2242.

A. Alaei, P. Nagabhushan and Umapada Pal, (2010) “A New Two-stage Scheme for the Recognition of Persian Handwritten Characters”, 12th International Conference on Frontiers in Handwriting Recognition 2010, 978-0-7695-4221-8/10. DOI 10.1109/ICFHR.2010.27.

Abandah, G.A., Younis, K. S. and Khedher, M. Z. (2008) “Handwritten Arabic character recognition using multiple classifiers based on letter form”, In Proc. 5th IASTED Int'l Conf. on Signal Processing, Pattern Recognition, & Applications (SPPRA 2008), 2008.

R. Wisnovsky. (2004) “Philosophy, Science and Exegesis in Greek, Arabic and Latin Commentaries”, volume 2, chapter: The nature and scope of Arabic philosophical commentary in post-classical (ca. 1100-1900 AD) Islamic intellectual history: Some preliminary observations, pages 149-191. Institute of Classical Studies, London, 2004.

Anuradha B, Koteswarrao B, (2006) “An efficient Binarization technique for old documents”, 2006 Proc. Of International conference on Systemic, Cybernetics, and Informatics (ICSCI2006), Hyderabad, pp771-775.

M. Rashad and N. Semary, (2014) "Isolated Printed Arabic Character Recognition Using KNN and Random Forest Tree Classifiers," in Advanced Machine Learning Technologies and Applications, Springer, 2014, pp. 11-17.

A. Rosenberg and N. Dershowitz, (2012) "Using SIFT Descriptors for OCR of Printed Arabic," Tel Aviv University, 2012.

M. Dahi, N. A. Semary and M. M. Hadhoud, (2015) “Primitive Printed Arabic Optical Character Recognition using Statistical Features”, IEEE Seventh International Conference on Intelligent Computing and Information Systems, 2015.

R. Sharma, M.S. Patterh (2015) “A new pose invariant face recognition system using PCA and ANFIS”, Elsevier GmbH 2015.

G. A. Abandah, K. S. Younis and M. Z. Khedher, (2008)"Handwritten Arabic Character Recognition Using Multiple Classifiers Based On Letter Form", In Proc. 5th IASTED Int'l Conf. on Signal Processing, Pattern Recognition, & Applications (SPPRA 2008), Feb 13-15, Innsbruck, Austria.

BIODATA

	Mr. Omar Balola was born in Sudan. He discussed completion in the PhD program in Computer Science at the University of Sudan for Science and Technology. He held a M.Sc. in Computer Science from Sudan University of Science and Technology, Sudan in 2011. He got B.Sc. (Computer Science) from Omdurman Islamic University, Sudan in 2001. In 2004, he joined the Department of Computer Science, Omdurman Islamic University as a Teaching Assistant, and in 2011 became a Lecturer. At present he teaches courses in Advanced Database, Visual Programming Languages, AI and Intelligent Systems. His current research interests include pattern recognition, image processing, neural network and fuzzy logic.
	Dr. AdnanShaout is a full professor in the Electrical and Computer Engineering Department at the University of Michigan - Dearborn. At present, he teaches courses in embedded systems, cloud computing, software engineering methods, fuzzy logic and engineering applications and computer engineering (hardware and software). His current research is in embedded systems, applications of fuzzy set theory, software engineering, artificial intelligence and cloud computing. Dr. Shaout has more than 34 years of experience in teaching and conducting research in the electrical and computer engineering fields at Syracuse University and the University of Michigan - Dearborn. Dr. Shaout has published over 210 papers in topics related to electrical and computer engineering fields. Dr. Shaout has obtained his B.Sc., M.S. and Ph.D. in Computer Engineering from Syracuse University, Syracuse, NY, USA, in 1982, 1983, 1987, respectively.