مجلة صبراتة للعلوم البحرية والشاملة
Sabratha Journal
for Marine and Total Sciences
عدد خاص بالمؤتمر
العلمي الدولي الخامس
لتكنولوجيا علوم البحار
February 2025 فبراير

Comparative Analysis of Feature Extraction Techniques for Handwritten Arabic Character Recognition: using CCOB and Eigenvector-based Approaches

1Musa Omar, 2Omar Balola Ali, 3Salahaddin Sahboun
1 Computer Science, Faculty of Information System , Ajdabia University – Ajdabia, Libya
2 Computer Engineering, Faculty of Science Engineering, Bright Star University – El-Brega, Libya
3Mechanical and Industrial Engineering, Faculty of Technical Engineering, Bright Star University – El-Brega, Libya
1musa.omar@uoa.edu.ly, 2omar.balula@bsu.edu.ly, 3s.sahboun@bsu.edu.ly

Abstract:

This paper compares two feature extraction techniques for the automatic recognition of cursive handwritten Arabic characters. The first technique, known as CCOB, incorporates statistical features such as the center of mass, crosshair count, outliers (Right, Left, Top, and Down), and black ink histograms. The second technique involves extracting eigenvectors and eigenvalues from the shifted mean of resized cropped images. The comparison was conducted using the SUST-ARG dataset, which consists of 6,800 images of Arabic handwritten characters. All experiments utilized the Adaptive Neural Network Fuzzy Inference System (ANFIS) classifier. The experimental results demonstrated the performance of both techniques. The recognition rate achieved with the CCOB features was 96.1%, while the second technique yielded a slightly lower recognition rate of 95.94%. Overall, the comparison indicates that both techniques are effective in recognizing Arabic handwritten characters, with only a marginal difference in recognition rates. Notably, the first technique slightly outperformed the second in terms of recognition rate.

Keywords: Handwritten Character Recognition (AHC), Neural Network Fuzzy Inference System (ANFIS), Neuro-fuzzy

1. Introduction and related work

Recognizing Arabic letters poses a challenge due to the need for descriptors that account for the presence of similar-looking characters. For instance, certain characters may have identical body shapes but differ only in diacritical points. In the subsequent sections, we will provide a thorough explanation of the feature extraction methods employed and present the corresponding results..

Handwritten recognition in general involves two types of feature extraction: statistical and structural. Statistical features include histograms of transition and projection profile, histograms of gray level distribution, Fourier descriptors, and chain code. On the other hand, structural features consider factors like the presence of loops, the number and position of diacritical points, and the orientation of curves. [1].

The literature review includes several studies on Arabic character recognition. In 2006, [2] proposed a system that utilized wavelet transform for feature extraction and a neuro-fuzzy approach for character recognition. The system achieved a recognition rate of 95.64%. In 1997, [3] developed an on-line Arabic handwriting character system using a fuzzy neural network as a classifier. The training dataset consisted of 2000 characters written by a single writer.

[4] introduced a new method for off-line recognition of handwritten Arabic characters based on structural characteristics and a fuzzy classifier. They employed Fuzzy ARTMAP neural network and Five Fuzzy ARTMAP neural networks, achieving a recognition rate of 93.8%.

[5] presented an off-line Multiple Classifier System (MCS) for Arabic handwriting recognition. The MCS combined two individual recognition systems based on the Fuzzy ART network. The best combination ensemble achieved a recognition rate of 90.1%.

These studies demonstrate different approaches to Arabic character recognition, including neuro-fuzzy approaches, fuzzy classifiers, and multiple classifier systems. The recognition rates achieved in these studies highlight the effectiveness of these methods in recognizing Arabic characters.

The literature review encompasses various studies on offline Arabic character recognition, focusing on different statistical feature extraction methods. Some of the key findings from these studies are summarized as follows:

Statistical features played a significant role in pattern recognition. [6, 7, and 8] used pixel densities, sum of black and white pixels, and detection of black and white points as statistical features, respectively. Vertical and horizontal projections were utilized in [9] and [10], where [9] specifically used the longest spike to represent the baseline.[11] used branch, start-point, and end-point features of a character. [12] and [13] utilized features related to character body and secondary parts, position relative to other parts, loops, and

Radon transforms. Pseudo-Zernike moments, size, rotation, and translation invariant features were used in [14].[15] and [16] employed features such as center of mass, crosshair, outliers, and black ink histogram.

These studies demonstrate the diverse range of techniques and features used for offline Arabic character recognition, highlighting the importance of statistical features in achieving accurate recognition results.

Section 2 provides an introduction to the recognition system, a detailed description of the preprocessing, normalization, and feature extraction methods, and a short introduction to the HMM based recognition process. Section 3 follows with results using the different feature sets.

2. Recognition system

Figure 1 shows a block diagram of the recognition system. In the following subsections we describe normalization, feature extraction, and ANFIS recognizer in more detail.

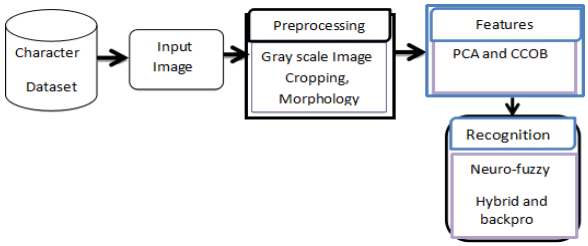

Figure 1. Illustrates the recognition system steps.

2.1 Dataset

This paper used a dataset , namely the SUST-ARG dataset, which stands for Sudan University for Sciences and Technology Arabic Recognition Group (SUST-ARG). The dataset comprises a total of one hundred and forty-one forms that were completed by various individuals.

2.2 Preprocessing

After scanning a document some basic preprocessing tasks like image binarization, characters segmentation, and noise reduction have to be performed. Due to the fact that we use the cropped grayscale characters images coming from the (SUST-ARG) - dataset, morphology and cropping techniques for images of ancient Arabic documents used. A handwriting character has been sampled on A4 size paper. The characters w scanned using a scanner with a resolution of 300dpi. These characters, then segregated according to their own character group and stored as gray scale images.

2.3 Statistical Features

Statistical pattern recognition draws from constructed concepts in statistical decision theory to distinguish among data from different classes based upon quantitative features of the data. There are wide types of statistical techniques that can be used within the description task for feature extraction, ranging from simple descriptive statistics to complex transformations [17]. The following shows the major statistical features used for character representation**:**

CCOB Features: Vertical and horizontal crossings are found by counting the number of white-Black -white transfers when scanning the image's pixels on a vertical line and a horizontal line, respectively [18]. Outliers and blank ink histogram [19]. **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. **Cross feature** Count the number of transitions from background to foreground pixels along vertical and horizontal lines through the character image as shown in Figure 2. **Outliers** (Right, Left, Top and Down) - calculate 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. **Black ink histogram** is calculate the black ink histogram features as shown in figure 3.

Figure 2 Vertical and Horizontal character, crosshair and center of mass.

Figure 3: Right, Left, Top and Down character, Outliers and Blank ink.

2) **Principal Component Analysis (PCA):** The principal component analysis is one of the oldest and most famous of the multivariate analysis techniques. Put the basis first by

Pearson (1901), and developed by Hoteling (1933). However, it has not been used widely until the advent of electronic computer, but now use in the statistical software package [20].

Principal Component Analysis (PCA) is a widely used method for feature extraction in various applications, including image processing. The primary purpose of PCA is to reduce the dimensionality of data matrices to more manageable sizes. The PCA process typically starts by calculating the mean of the data matrix, followed by computing the covariance of the data. Eigenvalues and Eigenvectors are then estimated [21].

The main objective of PCA is to identify the directions in which the data exhibits the highest variance. These directions, represented by the Eigenvectors, form a space that captures the most significant features of the data. By reducing the dimensionality of the data using PCA, it becomes easier to analyze and work with the data while still preserving important information related to the overall variance in the dataset [22].

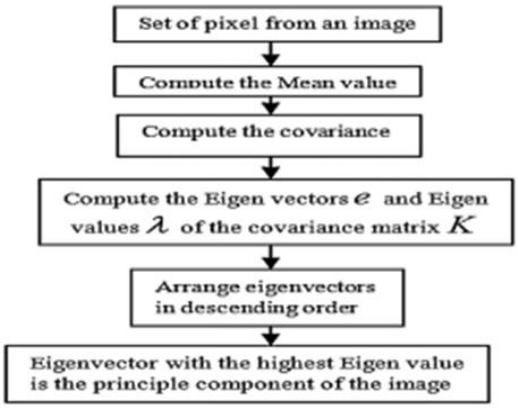

Figure 4. The flow chart that illustrates the PCA.

2.4 Adaptive Neuro Fuzzy Inference System (ANFIS)

ANFIS [23] implements an FIS model and has a five layered architecture as shown in Figure 5. The first layer received inputs from external environment and the input variables are fuzzified. The second layer computes the rule antecedent part. The third hidden layer normalizes the rule strengths followed by the fourth hidden layer where the consequent parameters of the rules are determined. The last layer is the output layer which computes the summation of all input signals. ANFIS uses back-propagation learning to adjust the premise parameters and Least Mean Square estimation to control the consequent parameters.

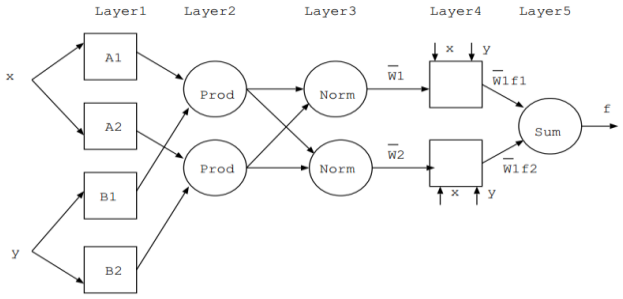

Figure 5. ANFIS architecture for a 2-inputs Sugeno fuzzy model.

3. Experimental result

The (SUST-ARG) data set for Arabic handwritten character together with the mentioned ANFIS recognizer was used to test the two different feature sets. These dataset was partitioned into two distinct levels. These levels differ in terms of their size and the number of classes they encompass.The first level comprises the entire dataset, with each individual character belonging to its own unique class. In the second level, the dataset is organized into groups of characters that share similarities, and each group is assigned to a single class . The dataset was divided into two levels as shown in table 1. This partitioning approach was employed in a study conducted by Balola and Shaout [24].

Table 1: Data set types and sizes for training and testing.

Dataset Level	Dataset size	Training dataset	Testing dataset
First level 34 class	6800	5100	1700
Second level 15 classes	3000	2250	750

In the experiments, have been employed two types of statistical features. The first type of features (CCOB) are the Center mass (xt, yt) of the character image, Cross hair count the number of transitions, Outliers (Right, Left, Top and Down) and Black ink histograms. The second sets of features used PCA to select the desired number of eigenvectors and eigenvalues from resize cropped images.

3.1 Experiments with First Statistical Set of Features using CCOB Technique

The AHC images are taken from the SUST-ARGG dataset. These images are both binarized and morphologicalized, second are passed to the feature extracted techniques and extracted four (CCOB) features.

features were used as inputs to the system, three numbers of triangular membership functions were used and single output with 1000 epochs.

Analyzing results show the improvement at the second level were the recognition accuracy is 96.1% for isolated Arabic handwritten character with testing data set in grouped classifiers with an increase of about 5.6% from the recognition accuracy achieved by a single classifier system as shown in table 2.

Table 2. Results using the CCOB features.

Dataset levels	Training Accuracy %	Testing Accuracy %
All dataset(single classifier)	90.7	90.5
Grouping dataset(grouped classifiers)	96.12	96.1

3.2 Experiments with Second Statistical Set of Features using PCA Technique

The handwritten character images are taken from the SUST-ARGG data set. The outer parts of these images are removed using the cropping technique. After the unwanted background parts of the image are omitted, the cropped images are resized to 7 by 5. The images are then passed to the feature extraction process in order to calculate the score values using the principal component analysis (PCA) technique. The score values which were obtained from the PCA techniques are then used by ANFIS classifier for doing the training process. Features were used as inputs to the system.

A experiments were applied on the training data set and the testing data set for data set and the results are reported in table 3.

Table 3. Results using the PCA features.

Dataset levels	Training accuracy %	**Testing accuracy%**
All data set	92.58	**92.11**
Grouping data set	96.75	**95.94**

The recognition rates of character recognition calculated by using classifier methods are shown in table 4. The capability of the ANFIS model in recognition rate is better when using the CCOB feature technique. We noted that there is an improvement in the recognition result whenever we used CCOB Technique features.

Table 4. Highest accuracy results using CCOB and PCA.

Classifiers levels	CCOB	PCA
Single classifier	93.5	92.11
Group classifier	96.1	**95.94**

4. Conclusion

In this research paper, we presented an approach for recognizing Arabic isolated handwritten characters using the Adaptive Neuro-Fuzzy Inference System (ANFIS). We compared two different feature extraction processes to determine their effectiveness in handwritten character recognition.

Our approach utilized two types of statistical features. The first type involved calculating the area formed by the projections of the upper and lower character profiles, as well as the left and right profiles. We also determined the center of mass (xt, yt) of the character image, counted the number of transitions, identified outliers in different directions (Right, Left, Top, and Down), and analyzed the black ink histograms. The second set of features was based on extracting eigenvectors and eigenvalues from resized and cropped images. We used cropped grayscale images of Arabic handwritten characters and applied PCA (Principal Component Analysis) for feature extraction. Morphology was used to binarize the images, and the second set of features was applied. The classification method employed a neuro-fuzzy classifier learning algorithms. The dataset was divided into two levels, with characters grouped based on their similarity.

Through our experiments, we achieved a high accuracy rate for the testing dataset, with the CCOB (Center of Character's Outer Boundary) feature yielding the highest result accuracy of 96.1%.

The findings indicate that the CCOB-based feature extraction technique outperforms the second set of statistical features based on PCA.

In future research, we aim to improve the recognition accuracy of neuro-fuzzy for handwritten Arabic character recognition by exploring more advanced feature extraction techniques.

References

[1] Y. Boulid, A. Souhar, and M. Y. Elkettani, “Handwritten character recognition based on the specificity and the singularity of the arabic language,” International Journal of Interactive Multimedia and Artificial Inteligence, vol. 4, no.4, pp. 45-53 Regular Issue, 2017

[2] N. Ben Amor, M. Zarai, and N. E. Ben Amara, "Neuro-Fuzzy approach in the recognition of Arabic Characters ", IEEE, 2006.

[3] A. M. ALIMI, "An Evolutionary Neuro-Fuzzy Approach, to Recognize On-Line Arabic Handwriting”, IEEE, 1997.

[4] M. Kefl, L.Chergui and S. Chikhi, “A novel fuzzy approach for handwritten Arabic character recognition”, Springer-Verlag London 2015.

[5] Ch. Leila, K. Maâmar and Ch. Salim, “Combining Neural Networks for Arabic Handwriting Recognition”, IEEE 978-1-4577-0908-1/11, 2011.

[6] H. A. Abed, “Fuzzy Logic approach to Recognition of Isolated Arabic Characters,” International Journal of Computer Theory and Engineering, Vol. 2, No. 1, pp.9-17, February. 2010.

[7] M. Z. Khedher. G. Al-Talib," Recognition of secondary characters in handwritten Arabic using Fuzzy Logic ", International Conference on Machine Intelligence (ICMI’05), Tozeur, Tunisia, 2005.

[8] M. A. Abdullah, L. M. Al-Harigy, and H. H. Al-Fraidi," Off-Line Arabic Handwriting Character Recognition Using Word Segmentation ", Journal of Computing, Vol. 4, Issue 3, MARCH 2012.

[9] H. Aljuaid, Z. Muhammad and M. Sarfraz, "A Tool to Develop Arabic Handwriting Recognition System Using Genetic Approach", Journal of Computer Science, vol. 6, 2010.

[10] S. M. Ismail, S. Norul, and H. Sheikh," Online Arabic Handwritten Character Recognition Based On A rule Based Approach", Journal of Computer Science, pp.1859-1868, 8 September 2012.

[11] M. Albakoor, K. Saeed and F. Sukkar, "Intelligent System for Arabic Character Recognition ", IEEE, 2009.

[12] R. I. Zaghloul, E. F. Alrawashdeh, D. Mohammad, and K. Bader, " Multilevel Classifier in Recognition of Handwritten Arabic Characters", Journal of Computer Science 7 (4), pp.512-518, 2011

[13] G. A. Abandah, K. S. Younis and M. Z. Khedher, "Handwritten Arabic Character Recognition Using Multiple Classifiers Based ON Letter Form”, 5th IASTED Int'l Conf. on Signal Processing, Pattern Recognition, & Applications (SPPRA, Innsbruck, Austria, 2008.

[14] M. Namaz, and K. Faez,"Recognition of Multifont Farsi / Arabic Characters Using a Fuzzy Neural Network", IEEE, 1996.

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

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

[17] R.T. Olszewski, “Generalized feature extraction for structural pattern recognition in time-series data”, Doctor of Philosophy, Carnegie-Mellon Univ, Pittsburgh PA, school of computer science, 2001.

[18] G. A. Abandah, and M. Z. Khedher. "Analysis of handwritten Arabic letters using selected feature extraction techniques", International Journal of Computer Processing of Languages, vol. 22, pp. 49-73, 2009.

[19] A. Nijim, A. El Shenawy, M. T. Mostafa and R. Abo Alez. "A Novel Approach for Recognizing Text in Arabic Ancient Manuscripts", International Journal on Natural Language Computing (IJNLC), Vol. 4, No.6, December 2015.

[20] J.T. Jolliffe. Principal component analysis 2nd. Springer series in statistics, 2002.

[21] O. B. Ali, and A. Shaout, “Hybrid Arabic Handwritten Character Recognition Using PCA and ANFIS”, In International Arab Conference on Information Technology (ACIT‟2016), 2016.

[22] F. Al-Saqqar , M. Al-Diabat , M. Aloun, M. AL-Shatnawi, "Handwritten Arabic Text Recognition using Principal Component Analysis and Support Vector Machines", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 12, 2019.

[23] Jang R, “Neuro-Fuzzy Modeling: Architectures, Analyses and Applications”, PhD Thesis, University of California, Berkeley, July 1992.

[24] Omar Balola, Adnan Shaout and Mohammed Elhafiz, “Two stage classifier for Arabic Handwritten Character Recognition”, International Journal of Advanced Research in Computer and Communication (***IJARCCE),*** Vol. 4, Issue 12, December 2015.