Offline Isolated Arabic Handwriting Character Recognition System Based on SVM

Mustafa Salam[1] and Alia Abdul Hassan[2]
[1]Computer Engineering Techniques, Imam Ja'afar Al-Sadiq University, Iraq
[2]Computer Science Department, University of Technology, Iraq

Abstract: *This paper proposed a new architecture for Offline Isolated Arabic Handwriting Character Recognition System Based on SVM (OIAHCR). An Arabic handwriting dataset also proposed for training and testing the proposed system. Although half of the dataset used for training the Support Vector Machine (SVM) and the second half used for testing, the system achieved high performance with less training data. Besides, the system achieved best recognition accuracy 99.64% based on several feature extraction methods and SVM classifier. Experimental results show that the linear kernel of SVM is convergent and more accurate for recognition than other SVM kernels.*

Keywords: *Arabic character, pre-processing, feature extraction, classification.*

Received October 5, 2015; accepted February 2, 2017

1. Introduction

In the character recognition system thee handwriting character images is converting into text file in the computer system that could be understandable and used for many useful purposes. Varity of applications need a handwriting system for making our life easier. These applications includes: cheque recognition, postal address reading, text, word spotting, and etc. Generally, handwriting character is more difficult than printed recognition due to its cursive style and affected by variety factors such as the style of each writers, the paper quality and other factors that controlled by the writing condition called geometric. Three main stages are considered in any system in order to recognize the handwriting character including: image pre-processing, feature extraction and character classification.

The first stage in handwriting recognition systems is image pre-processing. It leads to improve the accuracy of segmentation and recognition results by reducing the variability of handwriting. Besides, features extraction is the second stage of the system which extract unique information from the image character to specify it from the other characters. However, classification is consider the last stage of the character recognitions system which makes the final decision of the system by signing each character to its desired class [22].

2. The Research Method

Number of researchers has been work with OIAHCR system and obtained different results. Many researchers used image thinning a chain code for pre-processing stage [16, 18, 21]. For the features extraction, Clocksin and Fernando [13] is used a moment method in order to extract the required features. However, several researchers [4, 5] have used a structural feature like loops, dots, intersection and endpoints as a features. Where others using vertical and horizontal projection profile for features extraction stage [10].

Moreover, in classification stage Abed and Alasad [2] proposed new recognition approach for Arabic character based on Error Back Propagation Artificial Neural Network (EBPANN) as classifier and zoning technique for features extraction. Furthermore, Abdurazzag and Salem [3] proposed a system using Artificial Neural Network. The authors proposed a new algorithm for feature extraction based on Wavelet Transform (DWT) to achieve high accuracy and less recognition time by compress the character images.

3. Dataset

In this research an Arabic hand writing character images dataset has been proposed. The dataset collected from different people within different ages and education background. All the participants received white papers and write down all the Arabic characters. The used dataset has 560 hand writing character images. Furthermore, 20 images for each character with various sizes and styles Figure 1 shows a sample of the proposed dataset.

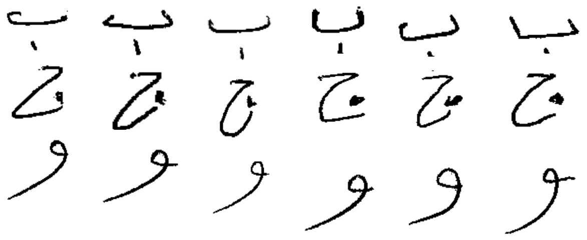

Figure 1. Samples of proposed dataset.

4. Proposed System

The proposed OIAHCR system has several major stages. Each of the step affect the accuracy and the performance of the recognition. First of all the input images converts into grayscale then it will pass through several process as shown in Figure 2.

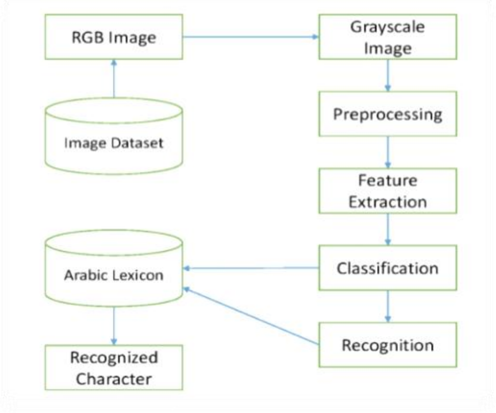

Figure 2. Flowchart of the proposed system.

The proposed system involves several stages which are; pre-processing, features extraction, classification and recognition. Besides that, each step has it benefits for the recognition process. The main stages of the proposed system are described in following:

5. Pre-Processing

Pre-processing is an essential stage in the OIAHCR system due to the effectiveness of this process on the recognition accuracy. Several steps has been taken place in the pre-processing stage that make the proposed system obtains a high accuracy. Figure 3 illustrate the main steps of the pre-processing stage.

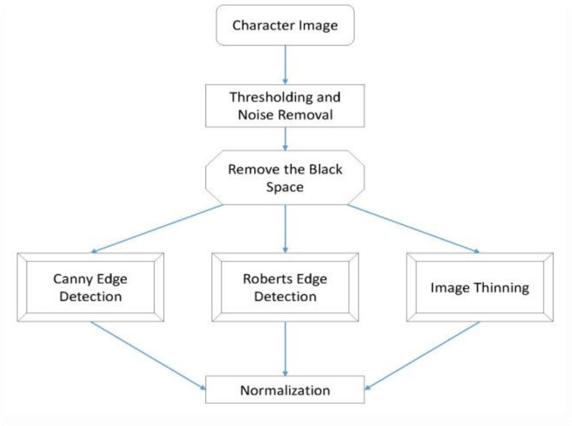

Figure 3. Main steps of the pre-processing.

- *Image thresholding and noise removal*: The input to the OIAHCR system is a RGB handwriting character image. The RGB image first converts to grayscale then goes to the next steps of the pre-processing. The image then converted to binary by thresholding method. The benefit of the thresholding is reducing the image diamantine to make it simple for processing.

In the proposed system Fuzzy C-Means clustering (FCM) and proposed noise removal algorithm [8] have been used to for thresholding and removing the undesired noise.

- *Remove the black space*: The second step of the pre-processing is removing the unwanted black space in the image background. A presented algorithm in [6] is used for removing the black space. First, the number of (0) values are calculated from the all image borders until the character which is representing by (1) value as in Figure 4.

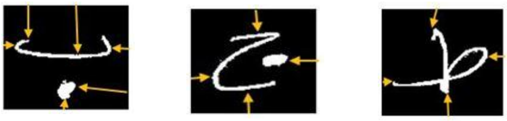

Figure 4. Calculating of (0) value.

Bounding Box is used to eliminate the space around the character and crop the desired space only as in Figure 5.

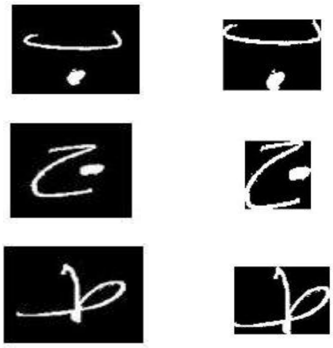

Figure 5. The output of applying the black space removal algorithm.

- *Image thinning*: is the process of reducing image size by remove the redundant pixels without losing the representation of the original image. 3*3 mask used to scan the whole image and find the 4 connected pixels. After that the unaffected pixels are eliminate from the image. This process must save the geometry and the connections between the characters and the location of original character based on border pixels removing recursively taking into account saving the geometry, location and connections. Image thinning method in [7] has been used as shown in Figure (6).

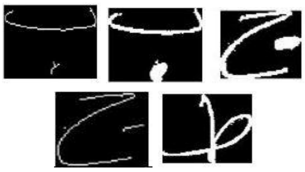

Figure 6. Image thinning.

- *Canny edge detector*: In the proposed system canny edge *detector*is used to produce edge image that will used for extracting Discrete Cosine Transform (DCT) and zoning features in the features extraction stage of the proposed system.
- *Roberts's edge detector*: After many testing on the handwriting images of the proposed dataset, Roberts edge detector was the best choose for producing edge image in order to be used for extracting Histogram Oriented Gradient (HOG) features in the features extraction stage. Figure 7 shows the results of applying Roberts's edge detector.

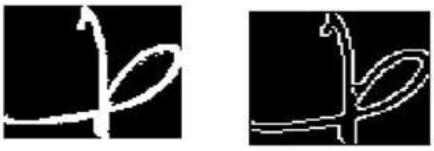

Figure 7. Applying roberts's edge detector.

- *Size normalization*: The proposed Arabic dataset has various image sizes. It important to make all the image in the dataset in the same size and make the recognition process fast. After testing several sizes the 64*64 gave best recognition rate. All the dataset images normalize into size 64*64 an example in Figure8 for this normalization.

Figure 8. Image normalization.

6. Features Extraction

The most important stage in OIAHCR system is the features extraction. The best recognition depends on a successful feature extractions methods. A lot of feature extractions methods has been proposed in this research. Furthermore, three main types of features are obtained from the handwriting character images which are:

6.1. Structural Features

Structural features describe the geometrical and topological characteristics of a pattern by describing its global and local properties. The structural features depend on the kind of pattern to be classified [16].

For Arabic characters, the features consist of zigzag, dots, loops, end points, intersection points and strokes in many directions.

Pre-processing stage produce three type of images. One of this type is the thinned image which is used to extract the structural features. In OIAHCR, several structural features has been extracted which are: dots, loops, end points, and intersection points.

6.2. Statistical Features

In statistical features several measures numerically are computed of the images over images or regions. Fourier descriptors, Histograms of chain code directions, moments, and pixel densities are an example of this features [17]. It can be computed easily and its text independent. Two statistical feature are used in the proposed system:

- *Connected components feature*: The idea behind of the connected component is to scan the whole image from left to right to find the groups of connected pixels (8-connected neighbors). After that, each group of the connected pixels will get a label number. Therefore, the feature that obtained from this method is the number of connected components. This method is useful in Arabic characters, since there are several characters has different number of connected components. The connected components feature extracted from the binary image that obtained from the previous phase and there will be different rectangle color drawing around each component in the binary image.
- *Zoning features*: In zoning features the image divided into number of zones and a particular features extracted from each zone. Several features extracted in this method which increased the recognition accuracy. In this method an image with canny edge detection is used from the precious phase.

First the image divided into four zones, then for each zone summation of the diagonal pixels has been calculated as a feature for that zone.

Second, the image divided into sixteen (16) vertical and horizontal blocks, then the summation of each block pixels will be the feature of that block [20].

6.3. Global Transformation

The transformation schemes convert the pixels transformation of the pattern to a more compact form which reduces the dimensionality of features [25].

- *The Discrete Cosine Transform Features (DCT)*: The DCT converts the pixel values of an image in the spatial domain into its elementary frequency components in the frequency domain. Given an image *f* (*x*, *y*), its 2D DCT transform is defined as follows:

$$F(u,v) = \frac{2}{N}\ C(u)C(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)\cos\left[\frac{(2x+1)u\pi}{2N}\right]\cos\left[\frac{(2y+1)v\pi}{2N}\right] \quad (1)$$

The inverse transform is defined by:

$$F(i,j) = \frac{2}{N}\sum_{u=0}^{N-1}\sum_{v=0}^{N-1} C(u)C(v)\ f(u,v)\cos\left[\frac{(2x+1)u\pi}{2N}\right]\cos\left[\frac{(2y+1)v\pi}{2N}\right] \quad (2)$$

Where

$$C(u), C(v) = \begin{cases} \frac{2}{\sqrt{2}} & for\, u, v = 0 \\ \frac{2}{\sqrt{2}} & for\;\, u, v\; \neq 0 \end{cases} \tag{3}$$

Due to its strong capability to compress energy, the DCT is a useful tool for pattern recognition applications. DCT could be employ successfully in various types of pattern recognition systems with SVM and ANN classification techniques [19].

In the proposed system the DCT applied for the whole canny edge detection image that produced from the previous phase. The output of the DCT is an array of DCT coefficients.

Zigzag order is used to arranging the DCT coefficient as a features, so small or zero coefficients will be away from the beginning. Experimental showed that, the first 20 coefficients is considered the best number of DCT coefficients to represent the character as feature vector.

The Algorithm DCT_EXT is used to extract the required features as following:

```
Algorithm: DCT_EXT
Step1: Read input image (with canny detection)
Step2:FindDiscrete Cosine Transform (DCT) for the input image
Step3: Convert the output of DCT into 1D array by zigzag order
Step4: Choose the first 20 DCT coefficients (features)
Step5: Use 1D array to save the obtained features
End
```

- *Histogram of Oriented Gradient (HOG)*: Histogram of Oriented Gradient (HOG) was first proposed by Dalal and Triggs [15] for human body detection but it is now one of the successful and popular used descriptors in computer vision and pattern recognition. HOG counts occurrences of gradient orientation in part of an image hence it is an appearance descriptor.

HOG divides the input image into small square cells (here 32×32 has been used) and then computes the histogram of gradient directions or edge directions based on the central differences. For improve accuracy, the local histograms have been normalized based on the contrast and this is the reason that HOG is stable on illumination variation. It is a fast descriptor in compare to the Scale Invariant Feature Transform (SIFT) and Local Binary Pattern (LBP) due to the simple computations, it has been also shown that HOG features are successful descriptor for detection. The HOG applied for the Roberts edge detection images from the previous phase.

Moreover, a thirty seven (37) features has been obtained by applying HOG for each Arabic character.

Thus, the number of overall extracted features is ninety seven (97) features[(20) from DCT(37) from HOG, (4) structural features and (36) statistical features] for each character which then saved in array of features vectors.

- *Features normalization*: An important step to make the mathematical computing simple and fast a feature normalization (scaling) has been used to make the features ranges [0 1] by applying the following formula:

$$A' = \frac{A - Min(A)}{Max(A) - Min(a)} \tag{4}$$

Where *A* is an original value, *A'* is the normalized value.

7. Classification and Recognition

After the feature extraction, the major task is the make decision to classify the character to which class it belongs. There are various classifiers that can applied in character recognition. The most important and more effective classifier is Support Vector Machine (SVM).

7.1. SVM Classifier

In the late 1990s Vapnik and Cortes developed a statistical learning machine called SVMs [14, 24]. Due to SVMs high classification rates, they are considered a best common classifier various pattern recognition and data mining applications. Besides that, several applications successfully used SVMs such as, bioinformatics, Object tracking, document analysis, Optical Character Recognition (OCR) and image classification.

Four common kernels are included in SVM which are linear, polynomial, RBF and sigmoid. An important multiclass SVM library called (libsvm) is used in the proposed system [12].An accurate results are obtained using these four kernels in the proposed system

After the classification is accomplished, the recognition is performed. Each recognized class is matched with its character ASCII then the Arabic lexicon used to retrieve the desired character of the chosen ASCII.

8. Experimental Result And Discussions

The proposed methods are implemented using Matlab R2015a, under windows10 64-bit Operating System, with RAM 6GB, CPU 2.50GHz core i5 and it achieved fast and effective results.

The proposed dataset has 560 handwriting character images. Each character has 20 images written in different style. In the OIAHCR system 50% of the dataset used for training purpose and 50% for testing and it achieved 99.64% recognition accuracy.

By testing all the 50% testing images, all the character images gave 100% recognition accuracy except the character (و) which gave 99% recognition accuracy as shown in Table 1.

Table 1. Recognition Accuracy for OIACR system.

No.	**Character**	**Recognition Accuracy**
1	**أ, ب, ت, ث, ج, ح, خ, د, ذ, ر, ز, س, ش ,ص, ض, ط, ظ, ع, غ, ف, ق, ك, ل, م, ن , ه, ي**	100%
2	**و**	99%

In the proposed system SVM classification work with different kernels and each kernel achieved different accuracy. Besides that, there are an important parameters which make the SVM work perfectly.

The most important parameters in SVM are: cost (c) and gamma (γ). After many testing of the system the best values of the parameters was c=4 and γ =0.25.

Furthermore, different SVM kernels has been tested and the best achievement was by using SVM linear kernel as shown in Table 2.

Table 2. Comparison between different kernels of SVM.

SVM Kernels	**Linear**	**Polynomial**	**RBF**	**Sigmoid**
Recognition Accuracy	99.64%	85%	95%	94%

Finally, several researchers work on OIAHCR system and they achieved good accuracy. However, after compering the previous systems with the proposed one, the proposed system gives the best accuracy recognition among the previous systems as shown in Table 3.

Table 3. Comparison accuracy between the proposed system and existing systems.

Author	**Classifier**	**Recognition Accuracy**
Sahlol and Suen[23]	Neural Network	88%
Abed and Alasad [2]	Error Back Propagation NN	93.61%
Ben Amor and Essoukri [11]	Hidden Markov	Varies from 94% to 98%
Farah [9]	Model	73.33%
Abdullah and Al-Harigy [1]	similarity percent	99.64%

9. Conclusions

In this a paper, an accurate Offline Isolated Arabic Handwriting Character Recognition system has been proposed. The paper proposed a dataset for evaluating the Arabic handwriting characters recognition systems. The proposed system used 50% of the dataset for training and 50% for testing and obtained high accuracy with SVM linear kernel. The high accuracy achieved by several factors starting from the efficient pre-processing stage with the use of FCM method, with efficient feature extraction methods and finally with an accurate classifier. Experiments, the proposed system has achieved a best recognition accuracy than the existing systems.

References

[1] Abdullah M., Al-Harigy L., and Al-Fraidi H., "Off-Line Arabic Handwriting Character Recognition Using Word Segmentation," *Journal of Computing*, vol. 4, no. 3, pp. 40-44, 2012.

[2] Abed M. and Alasad H., "High Accuracy Arabic Handwritten Characters Recognition Using Error Back Propagation Artificial Neural Networks," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 2, pp. 145-152, 2015.

[3] Abdurazzag A. and Rehie S., "Off-Line Omni-style Handwriting Arabic Character Recognition System Based on Wavelet Compression," *ARISER*, vol. 3, no. 4, pp. 123-135, 2007.

[4] Abuhaiba I. and Ahmed P., "Restoration of Temporal Information in Off-Line Arabic Handwriting," *Pattern Recognition*, vol. 26, no. 7, pp. 1009-1017, 1993.

[5] Abuhaiba I., Mahmoud S., and Green R., "Recognition of Handwritten Cursive Arabic Characters," *IEEE Transactionon Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 664-672, 1994.

[6] Abdul Hassan A. and Kadhm M., "An Efficient Preprocessing Framework for Arabic Handwriting Recognition System," *Diyala Journal For Pure Sciences*, vol. 12, no. 3, pp. 147-163, 2016.

[7] Abdul Hassan A. and Kadhm M., "Arabic Handwriting Text Recognition Based on Efficient Segmentation, DCT and HOG Features," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 11, no. 10, pp. 83-92, 2016.

[8] Abdul Hassan A. and Kadhm M., "An Efficient Image Thresholding Method for Arabic Handwriting Recognition System," *Engineering and Technology Journal*, vol. 34, no. 1, pp. 26-34, 2016.

[9] AL-Shareefi F., "A Haar Wavelet-Based Zoning for Offline Arabic Handwritten Character Recognition," *Journal of Babylon University/Pure and Applied Sciences*, vol. 23, no. 2, pp. 575-585, 2015.

[10] Al-Yousefi H. and Udpa S., "Recognition of Arabic Characters," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 14, no. 8, pp. 853-857, 1992.

[11] Ben Amor N. and Essoukri B., "Hidden Markov Models and Wavelet Transform in Multifont Arabic Characters Recognition," *in Proceedings of International Conference on Computing, Communications and Control Technologies*, United State, pp. 50-54, 2005.

[12] Chang C. and Lin C., "Libsvm: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1-27, 2011.

[13] Clocksin W. and Fernando P., "Towards Automatic Transcription of Syriac Handwriting," *in Proceedings of the 12th*

International Conference on Image Analysis and Processing, Mantova, pp. 664-669, 2003.

[14] Cortes C., "Support-Vector Networks," *Machine Learning*, vol. 20, no 3, pp 273-297, 1995.

[15] Dalal N. and Triggs B., "Histograms of Oriented Gradients for Human Detection," *in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, pp. 886-893, 2005.

[16] Govindan V. and Shevaprasad A., "Character Recognition-A Review," *Pattern Recognition*, vol. 23, no.7, pp. 671-683, 1990.

[17] Haraty R. and Ghaddar C., "Arabic Text Recognition," *The International Arab Journal of Information Technology*, vo. 1, no. 2, pp. 156-163, 2004.

[18] Haraty R. and Hamid A., "Segmenting Handwritten Arabic Text," *in Proceedings of the International Conference on Computer Science, Software Engineering, Information Technology, E-Business, and Applications*, Foz do Iguazu, pp. 95-101, 2002.

[19] Jiang J., Weng Y., and LI P., "Dominant Color Extraction in DCT Domain," *Image and Vision Computing*, vol. 24, no. 12, pp. 1269-1277, 2006.

[20] Kadhm M. and Abdul Hassan A., "ACRS: Arabic Character Recognition System Based on Multi Features Extraction Methods," *International Journal of Scientific and Engineering Research* vol. 6, no.10, pp. 665-661, 2015.

[21] Khorsheed M., "Recognizing Handwritten Arabic Manuscripts Using A Single Hidden Markov Model," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2235-2242, 2003.

[22] Mori S., Nishida H., and Yamada H., *Optical Character Recognition*, John Wiley, 1999.

[23] Sahlol A. and Suen C., "A Novel Method for the Recognition of Isolated Handwritten Arabic Characters," Technical Report, Cornell University, 2013.

[24] Vapnik V., *the Nature of Statistical Learning Theory*, Springer, 1999.

[25] Zaki F., Elkonyaly S., Elfattah A., and Enab Y., "A New Technique for Arabic Handwriting Recognition," *in Proceedings of the 11th International Conference for Statistics and Computer Science*, Cairo, pp. 171-180, 1986.

Mustafa Kadhm is a lecturer at the Department of Computer Engineering Techniques, Imam Ja'afar Al-Sadiq University. He received his B.S. degrees in Software Engineering from Al-Mansour University College, Baghdad, Iraq in 2009 and M.S. in Information Technology from University of Tun Abdulrazak, Malaysia in 2012. His research interests include Artificial Intelligence, Image Processing, Computer Vision, Pattern Recognition, and Data Mining.

Alia Abdul Hassan, Date of Birth: 28-3- 1971. Computer Science Dep. /University of Technology. B.Sc. Computer Science/ University of Technology/1993, MSc. Computer Science/ University of Technology/ 1999, Ph.D. Computer Science/ University of Technology /2004. Assistant professor since 20/3/2008. Position Deputy Dean of Computer Science Department since Feb 2015 till now. Supervised on 22 MSc. &PhD thesis in Computer Science since 2007. Publications Published more than (45) papers in International Conferences and Journals. Current Research Interests Soft computing, Green computing, AI, Data Mining, Software engineering, Document Recognition, Electronic Management, Computer security.