ACRS: Arabic Character Recognition System Based on Multi Features Extraction Methods

Mustafa S.Kadhm, Asst. Prof. Dr. Alia Karim Abdul Hassan

Abstract— this paper proposed a new architecture for Arabic Character Recognition System Based on Multi Features Extraction Methods and SVM Classifier (ACRS). An Arabic handwriting dataset proposed as well for training and testing the proposed system. Although half of the dataset used for training the 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.

Index Terms— Arabic Character, Preprocessing, Feature Extraction, Classification.
.

1 Introduction

Character recognition is the process of converting the handwriting text images into text file that understandable by the computer and used for many purposes. There are a lot of applications that depends on handwriting which are postal address reading for mail sorting purposes, cheque recognition and word spotting on a handwritten text page, etc. Naturally, handwriting is cursive and more difficult than printed recognition due to several factors which are the writer's style, quality of paper and geometric factors controlled by the writing condition its very unsteady in shape and quality of tracing. Several steps taking place in handwriting recognition system, starting with preprocessing, feature extraction and classification.

Preprocessing is the first step in Handwriting Recognition systems it is helpful to reduce the variability of handwriting by correct these factors and it will help to enhance the accuracy of segmentation and recognition methods. The second step in recognition system is the features extraction which extract a helpful information from the image character to distinguish it from the other characters. The last step of the recognitions is the classification which make the decision to sign the character to its desired class. [1]

2 Related Work

Number of researchers has been work with ACRS system and obtained different results. Many researchers used image thinning a chain code for preprocessing and to recognize the Arabic characters [2], [3], and [4]. For the features extraction, Clocksin[5] used moment functions to image and polar transform image. However, several researchers [6], [7] used structural feature like, loops, dots, intersection and endpoints to extract the required features. Where others using vertical and horizontal projection profile [8].

Moreover, in classification stage Majida and Hamid [9] proposed new architecture for Arabic character recognition based on Error Back Propagation Artificial Neural Network (EBPANN) as classifier and zoning technique for features extraction. Furthermore, Abdurazzag and Salem [10] proposed a system using Artificial Neural Network. Researchers 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. Characteristics Of The Arabic Characters

Arabic language is a widely used language by many countries around the world [11]. Arabic language is cursive and written from right to left. Arabic has 28 letters and eight diacritics [11]. Each character has different shape its position in the word. Moreover, there are different fonts that make Arabic character shape changed dramatically [12].

The table1 shows all the Arabic characters with its different shapes. Some characters has diacritics and some not. Character (ض) has upper diacritic .However, (ي) has lower diacritics.

TABLE 1
ARABIC LETTERS AND THEIR FORMS

Character Name		Isolated	Initial	Middle	Final	Character Name		Isolated	Initial	Middle	Final
Alif	ألف	ا	ا	ـا	ـا	Dhad	ضاد	ض	ضـ	ـضـ	ـض
Ba'	باء	ب	بـ	ـبـ	ـب	Tta'	طاء	ط	طـ	ـطـ	ـط
Ta'	تاء	ت	تـ	ـتـ	ـت	Dha'	ظاء	ظ	ظـ	ـظـ	ـظ
Tha'	ثاء	ث	ثـ	ـثـ	ـث	A'in	عين	ع	عـ	ـعـ	ـع
Jeem	جيم	ج	جـ	ـجـ	ـج	Ghain	غين	غ	غـ	ـغـ	ـغ
H'a'	حاء	ح	حـ	ـحـ	ـح	Fa'	فاء	ف	فـ	ـفـ	ـف
Kha'	خاء	خ	خـ	ـخـ	ـخ	Qaf	قاف	ق	قـ	ـقـ	ـق
Dal	دال	د	د	ـد	ـد	Kaf	كاف	ك	كـ	ـكـ	ـك
Thal	ذال	ذ	ذ	ـذ	ـذ	Lam	لام	ل	لـ	ـلـ	ـل
Rai	راي	ر	ر	ـر	ـر	Meem	ميم	م	مـ	ـمـ	ـم
Zai	زاي	ز	ز	ـز	ـز	Noon	نون	ن	نـ	ـنـ	ـن
Seen	سين	س	سـ	ـسـ	ـس	Ha'	هاء	ه	هـ	ـهـ	ـه
Sheen	شين	ش	شـ	ـشـ	ـش	Waw	واو	و	و	ـو	ـو
Sad	صاد	ص	صـ	ـصـ	ـص	Ya'	ياء	ي	يـ	ـيـ	ـي

4. Dataset

Here an Arabic character images dataset has been proposed. The dataset collected from different people with different ages and education background. All the participants received white papers and write down all the Arabic characters. The dataset has 560 character images. Each character has 20 images with different sizes and styles .Figure1 shows the collection of our dataset samples.

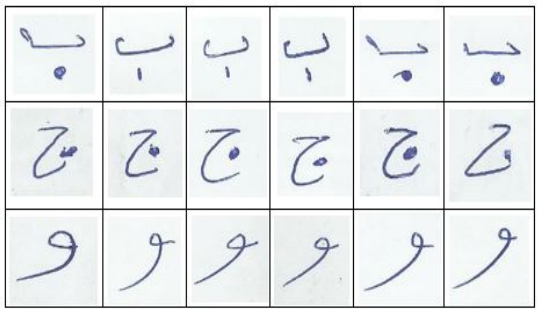

Fig. 1. Samples of proposed dataset.

5. Proposed Handwriting Recognition System

The proposed method for ACRS has several major steps. Each of the recognition step affect the accuracy and the performance of the recognition. First of all the input images converted into grayscale it pass through several process as shown in figure2.

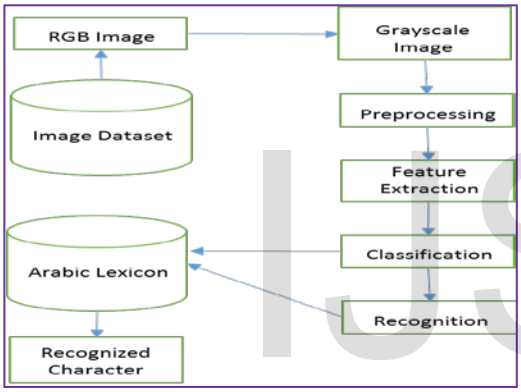

Fig. 2. Flowchart of the proposed system.

The proposed system involves several steps which are; preprocessing, feature extraction, classification and recognition. Besides that, each step has it benefits for the recognition process. Here the proposed method steps described in details:

5.1 Preprocessing

Preprocessing is an essential step in the ACRS due to the effectiveness of this process on the recognition rate. Several steps has been taken place in the preprocessing phase that make the proposed method obtain a high accuracy. Figure3 illustrate the main phases of the preprocessing step.

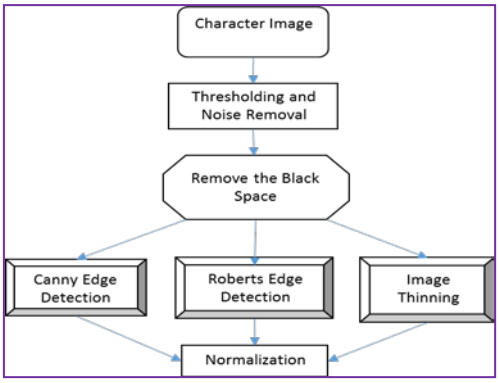

Fig. 3. Main steps of the preprocessing.

5.1.1. Image Thresholding and Noise Removal

The input to the ACRS is a RGB text image which has the Arabic word. RGB images convert to grayscale first then the preprocessing taking place. The image then converted to binary by thresholding method. The benefit of the thresholding is reducing the image diamantine to make it easy to process.
In the proposed system Fuzzy C-Means clustering (FCM) in [13] has been used to for thresholding purpose. After that, some noise appear due to the thresholding. Median used to remove undesired information from the binary image as shown in figure4.

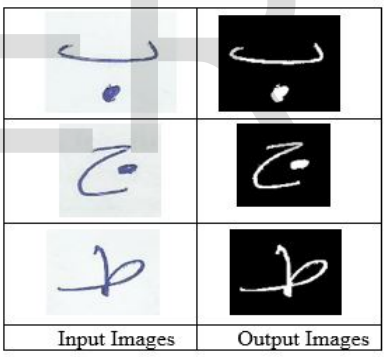

Fig. 4. Image thresholding.

5.1.2. Remove the Black SpaceFinal Stage

The second step of the preprocessing is removing the unwanted black space in the image background.
BoundingBox tool in Matlab 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 figure5.

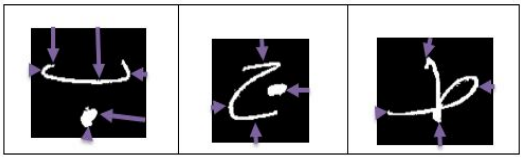

Fig. 5. Calculating of (0) values.

BoundingBox eliminate the space around the character and crop the desired space only in figure6.

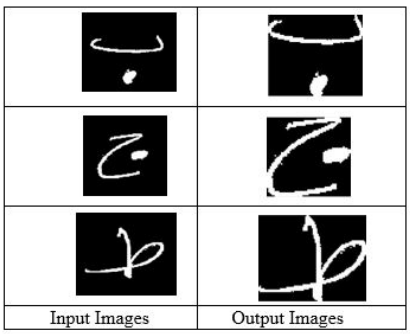

Fig. 6. The output of remove the black space method

5.1.3. 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 [14, 15], based on border pixels removing recursively taking into account saving the geometry, location and connections. Image thinning method in [15] has been used as shown in figure (7).

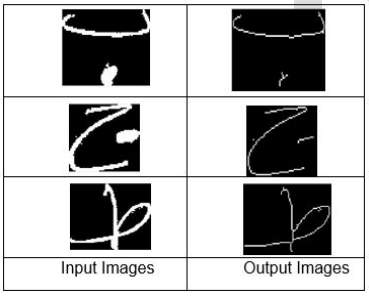

Fig. 7. Image thinning.

5.1.4. Canny Edge Detection

In the proposed system canny edge detection used to produce edge image to detect Discrete Cosine Transform (DCT) and zoning features in the features extraction phase of the system.

5.1.5. Roberts Edge Detection

After many testing of the system , roberts edge detection is best choose for producing edge image to detect Histogram Oriented Gradient (HOG) features in the features extraction phase of the system.

5.1.6. 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.

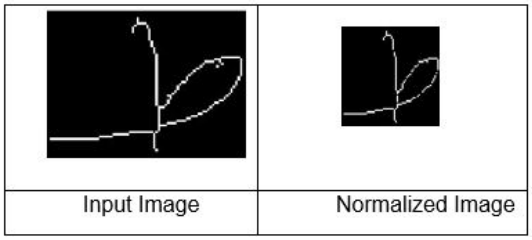

Fig. 8. Image normalization.

5.2 Feature Extraction

The most important process in ACRS. The best recognition depends on a successful feature extractions methods. A lot of feature extractions methods has been proposed for recognition purpose. However, there are three main types of features that can obtained from the character images.

5.2.1. 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.

Preprocessing phase produce three type of images. One of this type is the thinned image which is used to extract the structural features. In OIARC, several structural features has been extracted which are: dots, loops, end points, intersection points as shown in figure9.

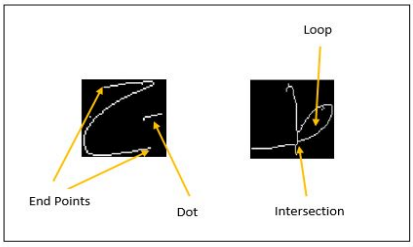

Fig. 9. Structural features.

5.2.2. Statistical Features

Statistical features are numerical measures computed over images or regions of images. They include, but are not limited to, histograms of chain code directions, pixel densities, moments, and Fourier descriptors [17] .Statistical features are easy to compute and text independent. In the proposed system two types of statistical feature has been used which are:

a. 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 as shown in figure10.

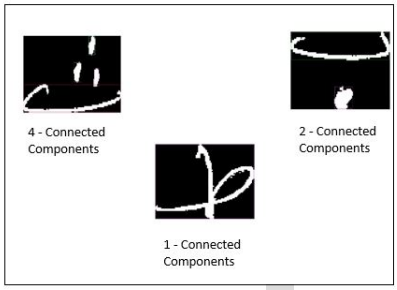

Fig. 10. Connected Components Feature.

b. 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 figure11 then for each zone summation of the diagonal pixels has been calculated as a feature for that zone.

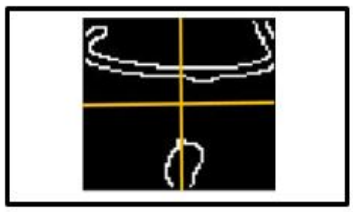

Fig. 11. Divide image into 4 zones.

Second, the image divided into sixteen (16) vertical and horizontal blocks figure (12) then the summation of each block pixels will be the feature of that block.

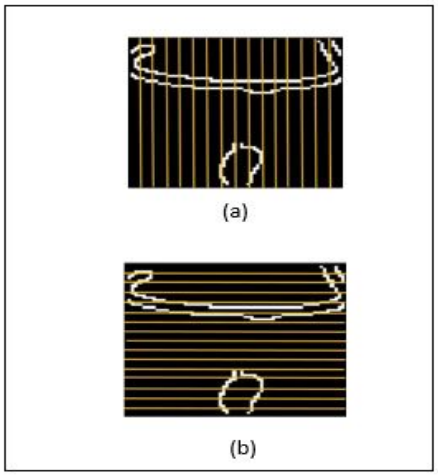

Fig. 12. (a) 16 vertical zones (b) 16horizontal zones.

5.2.3. Global Transformation

The transformation schemes convert the pixels transformation of the pattern to a more compact form which reduces the dimensionality of features [18].

a.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 (i, j), its 2D DCT transform is defined as follows:

$$f(u,v) = \alpha(u)\alpha(v)\sum_{i=0}^{I-1}\sum_{j=0}^{J-1} f(i,j)\cos\left[\frac{(2i+1)}{2I}\right]\cos\left[\frac{(2j+1)}{2J}\right] \quad (1)$$

The inverse transform is defined by:

$$f(i,j) = \sum_{U=0}^{N-1}\sum_{V=0}^{N-1} \alpha(u)\alpha(v)f(u,v))\cos\left[\frac{(2i+1)u\pi}{2N}\right]\cos\left[\frac{(2j+1)v\pi}{2J}\right] \quad (2)$$

Where

$$\alpha(u) = \alpha(v) = \begin{cases} \frac{1}{\sqrt{2}} & for \quad u,v = 0 \\ \frac{2}{\sqrt{2}} & for \quad u,v \neq 0 \end{cases} \quad (3)$$

Due to its strong capability to compress energy, the DCT is a useful tool for pattern recognition applications. The DCT can contribute to a successful pattern recognition system with classification techniques such as Support Vector Machine and Neural Network [19].

In the proposed system the DCT applied for the whole can-

ny edge detection image that produced from the previous phase. The output of the DCT is an array of DCT coefficients.

The features are extracted in a vector sequence by arranging the DCT coefficient in zigzag order, so that most of the DCT coefficients away from the beginning are small or zero. After testing the coefficients it found that the best number of DCT coefficients to represent the character as feature vector is vector is 20.

b. Histogram of Oriented Gradient (HOG)

Histogram of Oriented Gradient (HOG) was first proposed by Dalal and Triggs [20] 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 SIFT and 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.

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:

$$A' = \frac{A - \text{Min}(A)}{\text{Max}(A) - \text{Min}(A)} \quad (4)$$

Where A is an original value, A′ is the normalized value.

5.3 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).

5.3.1. SVM Classifier

Vapnik and Cortes developed SVMs [21, 22] as a statistical learning machine in the late 1990s. Within a short time, they became one of the most popular classification systems in data mining and pattern recognition applications, due to their high classification rates. Researchers successfully applied SVMs in many modern learning applications such as Optical Character Recognition (OCR), bioinformatics, document analysis, and image classification.

SVM commonly used with linear, polynomial, RBF and sigmoid kernels. A multiclass SVM classification (libsvm) has been used in the proposed system [23] with different kernels of 1) linear, 2) polynomial, 3) RBF, 4) sigmoid and it achieves a very high recognition accuracy.

The final step is the recognition which is matching the selected class by the SVM with the character ASCII and find the desired character in the Arabic lexicon.

6. Experimental Results and Discussions

The proposed method is implemented using Matlab R2015a version, under windows7 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 ACRS 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 table2.

TABLE2
RECOGNITION ACCURACY FOR ACRS 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 $\gamma = 0.25$.

Furthermore, different SVM kernels has been tested and the best achievement was by using SVM linear kernel.

TABLE 3
COMPARISON BETWEEN DIFFERENT KERNELS OF SVM

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

7. Conclusion

In this a paper, a proposed a high accurate Arabic Character Recognition system. A dataset for Arabic handwriting characters proposed as well. The system use 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 preprocessing stage with the use of FCM the with efficient feature extraction methods and finally with more accurate recognition classifier .Experiments, our proposed system gave best recognition accuracy than the existing systems.

REFERENCES

[1] Mori S., Nishida H. & Yamada H. (1999) Optical Character Recognition. (JohnWiley & Sons, NY).

[2] Khorsheed M.S. (2003) Recognising Handwritten Arabic Manuscripts Using a Single Hidden Markov Model. Pattern Recognition Letters, 24, 2235-2242.

[3] Haraty R. & Hamid A. (2002) Segmenting Handwritten Arabic Text. Proc. Int'l Conf. Computer Science, Software Eng., Information Technology, e-Business, and Applications, 2002.

[4] Haraty R. & Ghaddar C. (2004) Arabic Text Recognition. Int Arab J.Information Technology, 1, 156-163, 2004.

[5] Clocksin W.F. & Fernando P.P.J. (2003) Towards Automatic Transcription of Syriac Handwriting. Proc. Int'l Conf. Image Analysis and Processing,

[6] Abuhaiba I.S.I. & Ahmed P. (1993) Restoration of Temporal Information in Off-Line Arabic Handwriting. Pattern Recognition, 26, 1009-1017.

[7] Abuhaiba I.S.I., Mahmoud S.A. & Green R.J. (1994) Recognition of Handwritten Cursive Arabic Characters. IEEE Trans. Pattern Analysis and Machine Intelligence, 16, 664-672.

[8] Al-Yousefi H. & Udpa S.S. (1992) Recognition of Arabic Characters. IEEE Trans. Pattern Analysis and Machine Intelligence, 14, 853-857.

[9] Majida Ali Abed & Hamid Ali Abed Alasad.(2015) 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

[10] Abdurazzag Ali ABURAS and Salem M. A. REHIEL.(2007) Off-line Omni-style Handwriting Arabic Character Recognition System Based on Wavelet Compression. ARISER Vol. 3 No. 4 123-135

[11] Cox, E.: Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration. Elsevier, Amsterdam (2005)

[12] Klir, G.J., Folger, T.A.: Fuzzy Sets, Uncertainty and Information. Prentice Hall, Englewood Cliffs (1988)

[13] Elzobi, Moftah, Ayoub Al-Hamadi, Zaher Al Aghbari, and Dinges, Laslo (2012) http://www.iesk-ardb.ovgu.de/

[14] Sabri A. Mahmoud : KHATT (KFUPM Handwritten Arabic TexT) database, King Fahd University (2014) http://khatt.ideas2serve.net/

[15] Niall, Gunter, Innsbruck: Optical Character Recognition. Informatics Research Institute (IRIS) at University of Salford(2011)

[16] V.Govindan and A.shevaprasad."Character Recognition – a review" Pattern Recognition, vol. 23, no.7 , pp 671-683 , 1990.

[17] Issam Bazzi ,Richard Schwartz and John Makhoul. An "Omnifont Open-Vocabulary OCR System for English and Arabic". IEE Trans. On pattern analysis and machine intteligance. Vol 21 , no. 6 , pp 495-504,1999

[18] F.zaki, S. Elkonyaly, A. A. Elfattah, and Y. Enab. "A new technique for Arabic handwriting recognition". In processing of the 11th international conference for statistics and computer science ,(cairo , Egupt), pp. 171-180,1986

[19] JIANG, J., WENG, Y. & LI, P. (2006) Dominant colour extraction in DCT domain. Image and Vision Computing, 24, 1269-1277.

[20] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005

[21] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–297.

[22] Vapnik V. The nature of statistical learning theory. Springer; 1999.

[23] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. ACM T. Intell. Syst. Technol., 2(3):1–27, 2011