Proposed Multi Feature Extraction Method for Off-line Arabic Handwriting Word Recognition

Alia Karim Abdul Hassan*Ph.D(Asst.Prof.) Mohammed Alawi*

Abstract

This paper presents a proposed system for Arabic word recognition by using a set of techniques for feature extraction and KNN classifier. This system recognizes the Arabic word as one entity without segmentation approach .A proposed feature extraction method based on Discrete Wavelet Transform, upper/lower profile projection and the gradient (directional) feature. In order to evaluated the proposed system used IESK-arDB database and the experimental results showed the recognition rate of 89.05%.

Keywords: Feature Extraction, Gradient feature, DWT; KNN.

*University of Technology

1. Introduction

Offline Arabic handwriting recognition is the process of converting the Arabic text into its symbolic representation [1] . Many applications for text recognition that based on handwriting are postal address reading for mail sorting purposes, cheque recognition and word spotting on a handwritten text page, etc. [2] Handwritten word identification is a hard process because each writer has a unique style of writing and they have control over writing [3] . Two types for the recognition of Arabic scripts, are segmented based and segmentation free based. In the first type, each word divided into characters and recognizes each character in the word. This is called an analytical approach. While, the second type does not need segmentation and it is treating the word as a whole entity and recognize the word, based on the shape, structure and other features of the word [4] . Any recognition system has three stages, preprocessing, feature extraction and classification / recognition stags. The preprocessing stage tries to reduce the noise data while the feature extraction stage, It is a process of extracting useful information from the binary handwriting word image to be used in recognition stage. The last step is classification and recognition stage, which make the decision to assign each word to its desired class, then recognize the unknown handwriting word image to which class it denoted [5].

2. Related Work

The most recently works have been done for the Arabic word recognition without segment the word into its characters, in [6] suggest a recognition system based on proposed using DWT extract the features of the word image. And use Haar wavelet transform. Furthermore, the approximation coefficients (LL part) have been used to extract the features from the word image. The extracted features fed into the k-NN classifier. The recognition rate was 50.83% on IFN/ENIT databases .In [7] the outer based on the HMM (Hidden Markov Model) approach to build an explicit segmentation module. Shape representative based rather than a sliding window based features is extracted used to build a reference to a confirmation model for each letter in each handwritten form. IESK-arDB and IFN/ENIT databases are used for testing and evaluation of the proposed approach, respectively, and satisfactory results are achieved. The recognition rate reached 71% on the IESK-arDB and 58% for IFN/ENIT. In [8] an approach

to recognize the Arabic word by segment it into its basic letters the proposed approach based on using Gabor transform-based features that are extracted from each letter then passed to the SVM classifier for recognition. For training and testing, they used IESK-arDB database and the recognition rate was 74%. In [9] the input script segmented into words, after normalizing all words in size, each word is divided into overlapping blocks. The feature vector is computed by finding the absolute mean values for each block these feature vectors are fed to a K nearest neighbor classifier to classify the words. The system tested on IFN/ENIT database and the experimental results show a 76.04% recognition rate.

3-IESK-arDB database

IESK-arDB database consist of many Arabic word images exceed 4,000 word images and ‘6,000 segmented character images, each equipped with separate XML file includes a binary , thinned ,and the ground- truth information [10] . From this Database selected 70 words as classes. The no of images for training are 350 images and 210 images for testing. Figure1 shows samples of this database.

Figure1 samples of IESK-arDB database

4- Basic Concepts and Definitions: A set of techniques and methods were used to propose the new recognition method for Arabic word without segmentation are:

Median filter with 3*3 [11] apply to remove the noise and smooth the gray image.

Otsu binary methods [12] have been used for the thresholding method. Which used to reduce the image dimension?

Bounding Box the binary box is determined through finding out 4 points for each direction (up, left, down, right). These points are the first black point in each direction, and then determined the boundary box [13].

Normalization: in order to make the recognition process has more accuracy must all images have the same size.

Gradient direction feature: to model the rate of variation of the image / pixel the gradient direction feature is used. 3×3 Priwit filter mask used to find the horizontal & vertical gradient at each image pixel. The horizontal & vertical directions can be found by using two templates Priwit filter masks (shown in figure 2) [14]. Equations 1 and 2 are used to find the gradient strength and direction of each pixel I(x, y). Equation 2, θ (x, y) returns the direction of a vector (G_x, G_y). The θ(x, y) in the range [-90,90] [15].

Strength f(x,y)=squrt($G_x^2 + G_Y^2$) ...1

Direction θ (x,y)=$\tan^{-1}(G_x/G_y)$ …2

-1	-1	-1
0	0	0
1	1	1

Gx

-1	0	1
-1	0	1
-1	0	1

Gy

Figure 2 Priwit filter masks for Extracting Gradient Features

Profile projection is a type of structural features which provide the geometrical & topological properties of a pattern.In this work used a profile for upper & lower directions that captures the outline shape of connected parts of the word. For each columns calculate the distance

(number of points) from upper/lower points of the boundary box of the connected part to the first ink pixel of object [16] .

Discrete Wavelet Transform (DWT)

DWT is another technique used to extract the features of the words were, at each decomposition level. One type of wavelet transforms is Haar which gives the best result in Arabic handwriting recognition. The Haar wavelet was used to decompose each word image into 4- levels. Figure 3 shows the decomposition of DWT at one level. The horizontal sub-band (HL); it characterizes the high frequency in horizontal and low frequency in vertical, while. The low in horizontal and high frequency in vertical characterize the vertical sub-band (LH). The high frequency diagonal coefficients characterize the sub-band HH [6] .

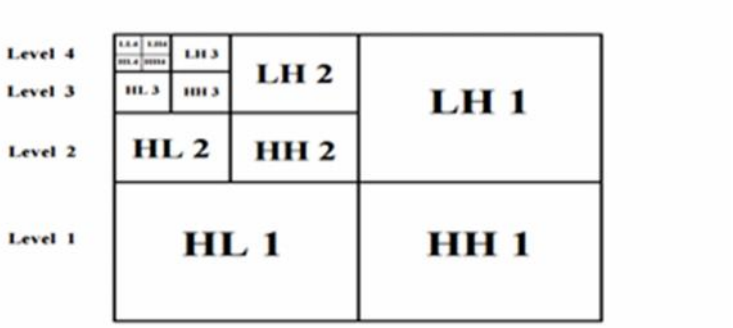

Figure .3 the DWT at 4 levels

K-Nearest Neighbor classifier (K-NN) is a simple, non-parametric instance-based learning, classifier use vectors in a multidimensional feature space as training samples, each with a class label. The training phase consists only of storing the feature vectors and class labels of the training samples. In the classification phase, k is defined by the user. The more frequent label among k train samples used to label unlabeled [17] . It calculates the majority vote of the closest neighbors of each K training sample.

Manhattan Distance Metric*:* equation 3 is used to find the distance between a points P= (p1, p2...) and a points Q= (q1, q2 …) [18] .

$(x)=\Sigma|pi-qi|$ ... 3

5. The Proposed Word Recognition System

The proposed system consists of three main stages are:

A- Preprocessing,
B- Feature extraction,
C- And, classification.

A- Preprocessing stage

A proposed algorithm for Preprocessing described in Algorithm 1.Step1 and step2 produce a gray image without noise .step3,4 convert the gray image into binary image the image by using Otsu method, then removing the white space around the word object by find out 4 points for each direction .In step 5 normalize the image into three different sizes. Finally step 6 use the Sobel and Priwit detectors to extract the image edges.

Algorithm1: Proposed Preprocessing Stage Input: word image Output: preprocessed word image
Step1: Convert the input word image to gray image. Step1: Convert the input word image to gray image. Step2: Apply median filter with 3*3 apply to remove the noise and smooth the gray image. Step3: use Otsu binary method to reduce the image dimension. Step4: remove the white space by using the boundary box determined through finding out 4 points for each direction (up, left, down, right). Step5: normalization, each image was normalized to three different sizes, 200*100 128*128 and 256*256 pixels. Step6: Use the Sobel edge detector and Priwit detection to extract the image edges.

B- Feature Extraction stage

The stage based on combination two groups of feature vectors .the first group extracted by using Algorithm 2 which considered a proposed algorithm used to extracted features from the preprocessed word image. In algorithm-2: step1,2 read the clipping binary image with 265*265 size. Then DWT Harr decomposition applied at 3 level. step 3,4,5 the decomposition image scanned with window 2^level*2^level without overlaps, then Stander division computed for each window to produce 64 feature vector. Step 6,7,8 cut the LL3 part (image 32*32),apply the edge detection process on this part (sub image)by using Sobel filter, then apply upper/lower to generate 64 feature vector. Finally combine the two features to constructed one vector with 128 features.

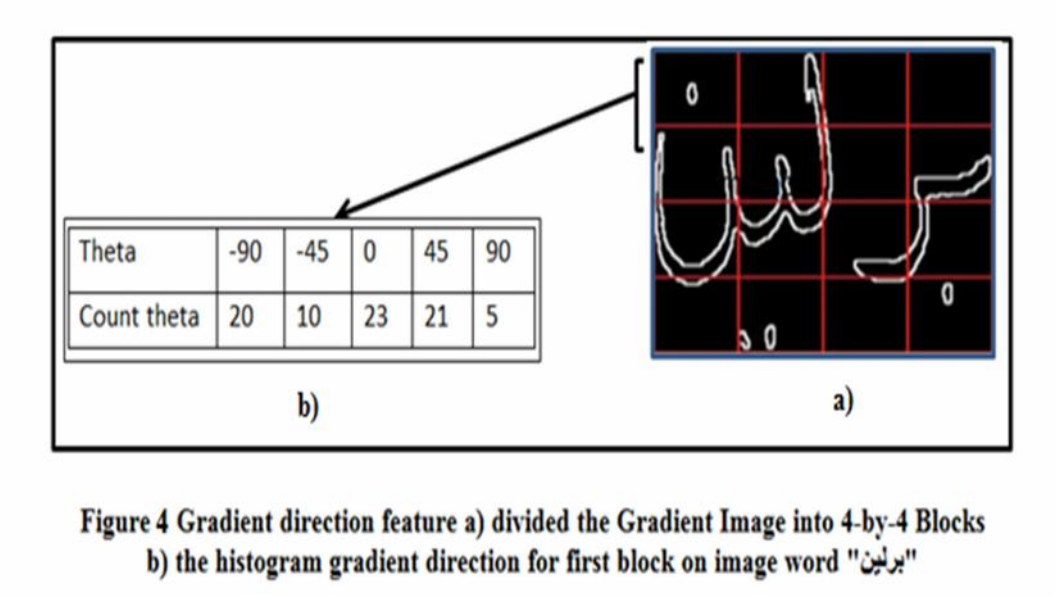

Figure 4 Gradient direction feature a) divided the Gradient Image into 4-by-4 Blocks b) the histogram gradient direction for first block on image word "برلين"

Algorithm-2: Proposed Feature Extraction Stage Input: clipping binary image Output: Features vector2
Step1: Read the clipping binary image with 256*256 sizes. Step2: apply the DWT Haar decomposition at 3 levels Step3: make window with 2^level *2^level size Step4: scan this window from top left to the bottom right direction of the input image Step5: calculate the stander derivation value for each window, then put this value in the feature vector list, called vector 21. Step6: repeat step 4 on the full image Step7: cut the LL3 part from the DWT image, called it sub _image its size (32*32) Step8: apply edge detection process on sub_image (LL3 part) by using a Sobel filter. Step9: apply profile projection method from the top and down direction on Sub _image ,the feature from this method is vector22 (64 values) Step10: combine these feature vectors to make a new vector called V ector2= (vector21+vector22). Step11: return vector 2

The second group of vectors is extracted by using A gradient direction method .When applying the gradient operation on image we get two components for each point in the image: strength and direction operation using equation (1 and 2) .This method used Priwit filter .The gradient components are vertical and horizontal components. In this method used the direction operation only by finding θ value .And the range of this θ is [-90 to 90].Which is divided into 5 values are [-90,-45, 0, 45, 90], if the value >-90 and <-67.5 make theta =-90, If theta < -22.5 Theta >= -67.5 then theta= -45, If theta < 22.5 and theta >= -22.5 then Theta= 0, if the value <90 and >67.5 make theta= 90, and if theta >22.5 And Theta <= 67.5 make Theta= 45. In this method divided the binary image into 4*4 blocks. For each block calculate the histogram Gradient direction for five values (-90,-45, 0, 45, 90). The feature vector length from this method is (5 angles *16 blocks) = 80 values. Algorithm 2 illustrated this method. The next step combination group 1 and group2 feature vectors to constructed one feature with 208 features which fed to classification stage.

Algorithm 2 : Gradient feature Input: clipped Binary image(image1) Output: gradient Features(vector1)
Step1: input the clipped binary image (image1) Step2: Apply the Priwit edge detection operator on the Binary image. Step3: Compute horizontal gradient component I_gx and vertical gradient component I_gy(Equations 1 and 2),then calculate gradient direction of each pixel in image1 by equations 2.4. Step4: Divide the gradient image into of 4*4 blocks Step5: for each block find the histogram gradient directions, for 5 angles which are [-90,-45, 0, 45, 90]. Step6: put the histogram values into 1D vector called it vector1. Step7: Return vector1 (80 values)

C-Classification/recognition stage

This stage based on KNN classifiers. This stage consists of two phases: training and testing phases. The training phase of KNN classifier considers only storing the feature vectors and class labels of the training samples. In the classification phase, k is a constant that define previously (k=1). In order to recognition the feature vector of testing image, Selected the K training vectors which closed to the testing vector by Calculating the distances by using Manhattan measure between each training vector and test vector. Test vector is assigned the class label which more frequently in K training data by using the majority vote

6. Experimental Results and Discussion

Visual basic.net language is used for programming the system stages. The IESK-arDB database is used to test the proposed word recognition system. From this database selected 650 word image for70 classes. Used 350 word images for training phase and 210 word images for testing phase (70% of data for training and 30% for testing). When entered the trained or tested images to the preprocessing stage applied several operations applied to reduce the noise and dimensional space on input images, the output from this stage a binary clipping image. Take 2 clipping images one which normalized 200*100 then applies gradient feature extraction, after that divided the image into 4*4 blocks to find the

feature vector. The second image is normalized into 256*256 then applies the proposed feature extraction method. By combining two feature vectors to construct a single vector from these methods. Then all vectors (training images) fed to train KNN classifier. A set of experiments was performed with, different feature extractions individually and in combination, different filters for gradient feature extraction and with more than one image size. Table 1 shows the comparison of different feature vectors the best accuracy when combines the proposed method with gradient direction method (Fv3). Table 2 illustrates some experiments with various sizes in normalization. The best result when the input image to the Gradient feature method is normalized with 200*100 and 256*256 with the proposed method.

Table 1 Recognition rate, according to used feature extraction method

Feature Vector	Proposed	Priwit mask	Recognition Accuracy%
Fv1	Y		80.48
Fv2		Y	85.24
Fv3	Y	Y	89.05

Table 2 : Recognition rates with different sizes.

The Proposed Method features	Gradient Feature (Priwit mask)	Recognition Accuracy
256*256	200*100	89.05%
256*256	128*128	83.81%
256*256	256*256	81.90%
256*256	64*64	80.47%
128*128	200*100	84.76%

Table 3 showed the experiment to extract the gradient features one times by using a Sobel filter and the other time by Priwit filter mask. Higher accuracy achieved with Priwit filter 89.02%, while 87.62 % of Sobel filter the same feature set.

Table 3: Recognition rate with different gradient filters

Gradient Feature types for Fv4	Feature length	Recognition Accuracy
Priwit Filter	208	89.05%
Sobel Filter	208	87.62%

Table 4 illustrated the recognition rate by using KNN classifier when using distance measurement such as Euclidean distance and Manhattan distance. The recognition rate is best for Manhattan distance.

Table 4: Recognition rate with different using distance metrics

Gradient Feature types for Fv4	Feature length	Recognition Accuracy
Manhattan distance	208	89.05%
Euclidean distance	208	85.24

7. Comparative Study

In Table 5 a comparative study of the proposed method with related work using the same data set, from this table can be seen the proposed technique for feature extraction with KNN classifier improve the recognition rate of Arabic word recognition systems.

Table 5 comparison results

Author	**Data base**	**Feature extraction method**	**Classifier**	**Accuracy**
Moftah Elzobi [7]	IESK-arDB databases	Sliding window based features	HMM	71%
Moftah Elzobi [8]	IESK-arDB database,	Gabor transform-based features	SVM	74%
Proposed system	IESK-arDB database,	Gradient direction feature and new feature extraction	KNN	89.05%

8. Conclusion

An Arabic word handwritten recognition system was proposed, which combines the strengths of both statistical and structural feature extraction. The proposed system used a combination of gradient based features profile projection and DWT. KNN classifier is used for classification and gets responsible results. The obtained results showed that using a combination of different types of features achieve a higher recognition rate. The accuracy of this system is 89.05%. The type and size of databases have an influence on handwritten Arabic word recognition systems, so may be used another database on this system. Another classifier such as neural network or SVM can be used in future works. The proposed system has the capability to process off-line handwriting printed Arabic numbers or letters images (for both English and Arabic characters).

References

[1] A. Benouareth, A. Ennaji and M. Sellami,“Arabic handwritten word recognition using HMMs with explicit state duration”, EURASIP Journal on Advances in Signal Processing, Article ID 247354, Vo l ,2008.

[2] M. S.Kadhm, A. K. Abdul Hassan,‘ACRS: Arabic Character Recognition System Based on Multi Features Extraction Methods’, International Journal of Scientific & Engineering Research, Volume 6, Issue 10, October-2015.

[3] Shruthi A , M. S. Patel , "Offline Handwritten Word Recognition using Multiple Features with SVM Classifier for Holistic Approach", International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, Issue 6, June 2015.

[4] J. H AIKhateeb, J. Ren, J. Jiang, S. S Ipson, H. El Abed, "Word-based Handwritten Arabic Scripts Recognition using DCT Features and Neural Network Classifier", IEEE ,5th International Multi-Conference on Systems, Signals and Devices, Jordan ,2008.

[5] M. S. Kadhm, A. K. Abdul Hassan,‘Arabic Handwriting Text Recognition Based on Segmentation and DCT And HOG Features ’, International Journal of Multimedia and Ubiquitous Engineering ,Volume 11, NO.101, pp83-92, 2016.

[6] J. Alkhateeb ," Word Based Off-line Handwritten Arabic Classification and Recognition", Ph.D. thesis, School of Computing, Informatics and Media, University of Bradford, 2010.

[7] M. Elzobi, A. Al-Hamadi, L. Dings, M.Elmezain, and A. Saeed ,"A Hidden Markov Model-based Approach with an Adaptive Threshold Model for Off-line Arabic Handwriting Recognition " International Conference on Document Analysis and Recognition, Publisher: IEEE 2013.

[8] M. Elzobi, A. Al-Hamadi, A. Saeed , and L. Dings, "Arabic Handwriting Recognition Using Gabor Wavelet Transform and SVM" (Volume 3) Page(s): 2154 – 2158 Publisher: IEEE, 2012.
[9] J. H. AIKhateebl, F. Khelifil, J. Jiani, S.S Ipsonl ," A New Approach for Off-Line Handwritten Arabic Word Recognition Using KNN Classifier", IEEE, International Conference on Signal and Image Processing Applications, Malaysian, 2009.
[10] M. Elzobi, A. Al-Hamadi ·Z. Al Aghbari · L. Ding, "IESK-ArDB: a database for handwritten Arabic and an optimized topological segmentation approach", International Journal on Document Analysis and Recognition (IJDAR) , Volume 16, Issue 3, pp 295–308, 2013.
[11] N. Sakthivel, L. Prabhu," Mean – Median Filtering For Impulsive Noise Removal", International Journal of Basic and Applied Science, Vol. 02, No. 04, pp. 47-57, April 2014.

[12] P. Grdiet, N. Kumar Garg, "Binarization Techniques used for Grey Scale Images"
International Journal of Computer Applications (0975 – 8887) Volume 71– No.1, June 2013
[13] A. K. Abdul Hassan, M. S. Kadhm, "An Efficient Preprocessing Framework for Arabic Handwriting Recognition System", Diyala journal of pure sciences, Vol: 12 No:3, ISSN: 2222-8373. July 2016.
[14] K. Kaur, S. Malhotra , "A Survey on Edge Detection Using Different Techniques",(IJAIEM) Volume 2, Issue 4, April 2013.
[15] N.Yadav, M. Ashraf, M. Yadav, " improvement of Edge Detection Technique in Images Using Soft Computing, IJIACS, ISSN 2347 – 8616 Volume 4, Issue 6 June 2015.
[16] Ahmed Sahlol , Cheng Suen, off-line system for the recognition of handwritten arabic character, Computer Science & Information Technology (CS & IT). pp. 227–244,CS & IT-CSCP 2014.
[17] A. Hirwani, N. Verma, S. Gonnade," Efficient Handwritten Alphabet Recognition Using LBP based Feature Extraction and Nearest Neighbor Classifier", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 11, November 2014
[18] M. H. Ramappa1 and S. Krishnamurthy," A Comparative Study of Different Feature Extraction and Classification Methods for Recognition of Handwritten Kannada Numerals", International Journal of Database Theory and Application Vol. 6, No. 4, August, 2013.

اقتراح طريقة لاستخلاص صفات متعددة لتميز الكلمات العربية المكتوبة بخط اليد

محمد علاوي عباس * أم.د. علياء عبد الكريم *

المستخلص

تقدم هذه المقاله نظاما مقترحا للتعرف على الكلمات العربية باستخدام مجموعة من التقنيات لاستخلاص الخصائص ومصنف الجار الاقرب . ويميز هذا النظام الكلمة العربية ككيان واحد بدون أسلوب تجزئة. الطريقة المقترحة لاستخراج الصفات تعتمد على تحويل المويجات المنفصلة،الاسقاط العلوي والسفلي وخاصية الأنحدار (الاتجاهي). لتقييم النظام المقترح استُخدمت قاعدة البيانات IESK-arDB ، حيث أظهرت النتائج التجريبية معدل اعتراف 89.05٪.

*الجامعة التكنولوجية