ISSN: 0067-2904

Proposed Handwriting Arabic Words classification Based On Discrete Wavelet Transform and Support Vector Machine

Alia Karim Abdul Hassan, Mohammed Alawi*
Department of Computer Science, University of Technology, Baghdad, Iraq

Abstract

A proposed feature extraction algorithm for handwriting Arabic words. The proposed method uses a 4 levels discrete wavelet transform (DWT) on binary image. sliding window on wavelet space and computes the stander derivation for each window. The extracted features were classified with multiple Support Vector Machine (SVM) classifiers. The proposed method simulated with a proposed data set from different writers. The experimental results of the simulation show 94.44% recognition rate.

Keywords: Handwriting Word Recognition (HWR), Binarization, Feature Selection DWT, SVM.

نظام مقترح لتمييزبعض الكلمات العربية المكتوبة بخط اليد بالاعتماد على تقنية محول المويجات المتقطعة (DWT)وآلة داعم المتجهات SVM

علياء عبد الكريم ، محمد علاوي عباس*
علوم الحاسوب، الجامعة التكنولوجية، بغداد، العراق.

الخلاصة

تم اقتراح خوارزمية لاستخراج الصفات من الكلمات العربية المكتوبة بخط اليد. تستخدم الطريقة المقترحة التحويل الموجي (DWT) على الصورة ثنائية، بعد ذلك تم مسح الصورة ذات التحويل الموجي باستخدام نافذة ، ومن ثم يتم حساب قيمة الانحراف المعياري لكل نافذة . تم تصنيف الميزات المستخرجة بواسطة المصنفات SVMs . تم اقتراح قاعدة بيانات جديدة كتبت من قبل عدد مختلف من الكتاب .وهذة القاعدة تم استخدامها لاختبار العمل المقترح وان النتائج التجريبية للنظام اظهرت معدل تميز 94.44٪.

1.Introduction

The handwriting recognition process means converting the handwriting text images into understandable text by the computer and by many applications such as postal address reading for mail sorting purposes, cheques recognition and word spotting on a handwritten text page [1].Offline recognition to Arabic handwriting is still very challenging because the writing styles is defferted from person to other, Also for the same person at different times . Because of the coursive nature of the Arabic script and the similar appearance of some Arabic characters makes the handwriting recognition difficult task. the recognition system has several stages such as Image acquisition, preprocessing, feature extraction and classification / recognition.Achieving high recognition accuracy depends mainly on the used feature selection method [2].The first stage in any recognition system is preprocessing which tries to reduce the noise data and keep only the desired information and make the next operation (feature extraction process) easy to implement. Moreover, the second stage is feature extraction and selection; they are extracting useful information from the binary handwriting word

*Email:alawialawi3000@gmail.com

image to be used in recognition stage. The last stage is classification and recognition ,which classified all data in different classes, then recognize the unknown handwritten word image to desired class.In this paper, a proposed handwriting recognition system for Arabic words by discrete wavelet transform and Support Vector Machine (SVM).

2. Related Work

The most recently works have been done for the Arabic language. In [3] use Wavelet Transform and Genetic Algorithms (GA) for recognizing the printed Arabic words. The system consists of three stages, preprocessing is applied on the input word image and the next stage is a feature extraction by using 1-level linear wavelet decomposition technique. Besides that, the third stage is classification by using GA. The proposed system tested on 10 Arabic words, the recognition rate was 90%. In [4] the research use 4-level of the Discrete Wavelet Transform (DCT) to extract the features from the word image then fed into the K-nearest neighbor classifier (KNN). The recognition rate of this system was 50.83%.

Nemouchi1. S et al. 2012 [5] design a recognition system for Arabic words. In the feature extraction stage used a Freeman code method which determined by the contour of the image, Zernike moments, and structural features. In classification phase use multiple classifiers such as K-Means algorithm, Probabilistic Neural Network , Fuzzy C-Means algorithm and K Nearest Neighbor algorithm. Also introduced a handwriting database of Algerian city names. The recognition rate was 80% . El moubtahij et al. 2014 [2] proposed system to recognize Arabic word based on using DCT for feature extraction stage and support vector machine classifier for recognition stage . Recognition rate was 91.70% on the IFN/ENIT Arabic standard database. AlKhateeb 2011 [6] used DCT features which extracted from each word sample, then features are fed to train a neural network for classification. The proposed system tested on IFN/ENIT Database each time 80% of the samples in the database are used for training and the remaining 20% for testing. Recognition rate 82.5%

3.3. Basic Concepts and Definition

3.3.1. Gray Scale Conversision Process

In color image each pixel is a combination of three colors Red, Green and Blue (RGB).. So each RGB color pixel has a depth of 24 bits. To convert each color pixels with 24 bits into gray scale pixels that have 8 bits, using equation 1[7].

$$GRY_{XY} = \frac{R_{XY} + G_{XY} + B_{XY}}{3} \quad (1)$$

Where (0≤X ≤image width),(0≤Y≤image high) , R, G, B are the red, green, blue color intensity of a pixel (x, y) respectively.

3.3.2 Noise Removal process

After scanning the input image to the system, the original image not Match the scanned image, because it has some noise because the quality of obtical scanner. **Median filter** with 3*3 [8] apply to remove the noise and smooth the gray image

3.3 Binarization Method

The binarization method plays an important role in the recognition system for text extraction from images which is a prominent area in digital image processing [9]. Image binarization is the process of converting an image into a binary image. A binary image is the digital image which has two values, i.e. 0 and 1 .The pixel value 0 represents the black color for foreground and 1 represent the white color for background. Thresholding is a technique which is used for image binarization. The pixels of an image are distinguished as a background and foreground by comparing them to thresholding value. Thresholding is further classified as a global and local approach. The global threshold method is suitable for images with contrast foreground and background. The local threshold is used when the image contains large amount of noise, illumination and uneven lightning . Otsu binary method [10] has been used for the thresholding method.Which used to reduce the image dimension**.** Otsu's method works better where clear separation between foreground and background exists or where image illumination is not variable. for this reason used Otsu's binaryzation method [11]**.**

3.4 Clipping Method

For clipping the handwritten word used Bounding Box. The binary box is determined through finding out 4 points,one for each direction (up, left, down, right). These points are the first black point in each direction , and then determined the boundary box [11].

3.5 Normalization:

In order to make the recognition process has more accuracy must all images have the same size**.**

3.6 Feature Extraction

The main goal of feature extraction is to find the most relevant information from the input data and represent that information in a lower dimensional space. Since each word has its own different features, instead of the full size input word reduce the work to these features which play an important role in any system for handwriting recognition [12].

3.7 Discrete Wavelet Transform (DWT)

DWT is another technique used to extract the features of the words . One type of wavelet transforms, is Haar which gives the best result in Arabic handwriting recognition. The Haar wavelet was used to decomposed each word image at differants levels. Figure- 1 shows the decomposition of DWT at three levels. The Haar wavelet decomposes the input image into four sub-bands, three detail components (LH, HL, HH) and one average component (LL) by using two filters ,high pass filter and low pas filter. The sub-band (HL); which known as horizontal coefficients. it characterizes the high frequency and low frequency components of image . While, the sub-band (LH) has The low in horizontal and high in vertical frequency components of the word image which called the vertical coefficients . The sub-band HH known as diagonal coefficients which has vertical and horizontal high frequency components .and the horizontal and vertical low frequency components represent the sub band (LL) and known as approximation coefficients [4]. Figure- 1 showns this bands. In this work, each word image converted to DWT space through decomposed it by four levels

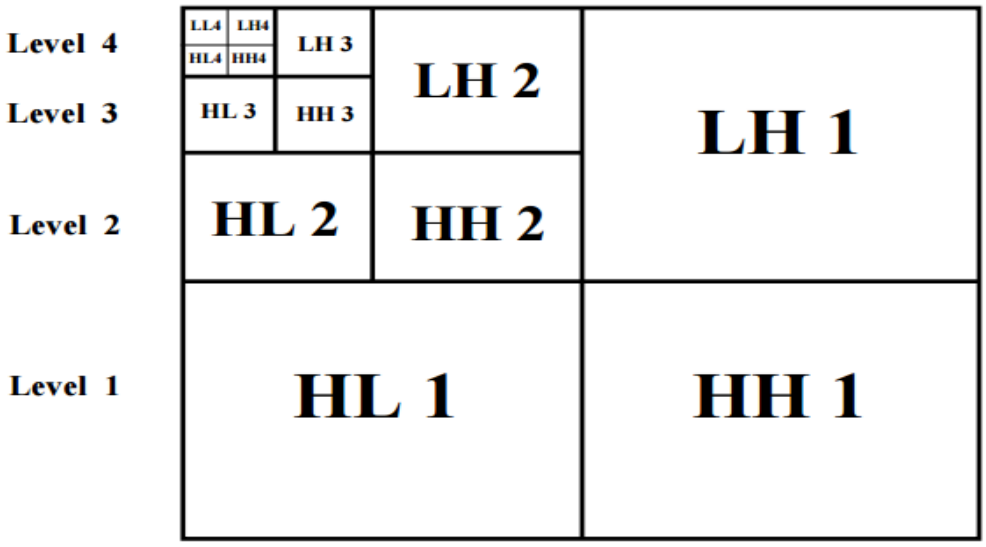


Figure 1- DWT at 4 levels

3.8 -Feature Selection (FS)

The process of choose subset of the input variables by eliminating features with little or no predictive information. There are several approaches in Feature selection [13,14] such as

1-*Forward selection*: begins with an empty set and features are added one by one, which decrease the error.

2-*Backward elimination:* begins with a feature set containing all features and features removed one at a time, at each step, removing the feature which increase or not effect on the error.

3.10 Support Vector Machine (SVM) Classifier

SVM is a statistical learning machine founded in the late 1990s, then became one of the most popular classification systems in recognition applications, due to their high classification rates [15].

SVMs were originally designed for binary classification, In order to extend SVM for multi-class classification, there are two types of approaches such as One- against-all and one-against-one approach [16]. SVM classifiers use kernels to give optimal decision boundary to separate between classes in higher dimensional feature spaces. By different kernel functions can transform a non-linear separable problem into a linear separable one by projecting data into the feature space, and then SVMs can find the optimal separating hyper plane [17].

4. The Proposed Word Recognition System

In this section, the proposed recognition system is described. This system consists of several stages such as Data collection, preprocessing, feature extraction, and classification / recognition stages. The detailed diagram of the proposed recognition The proposed system is shown in Figure- 2.

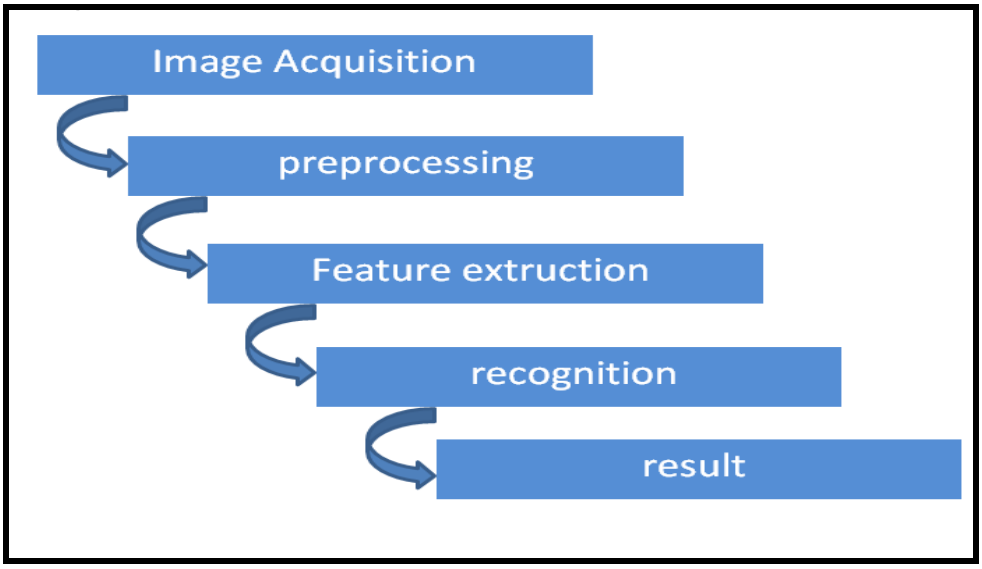


Figure 2- stages of proposed system

4.1. Database

A new database for Iraqi cities' names was developed. The number of words (city names) is 40 words which written by 30 writers of different age and education, on blank papers and used two pens with black and blue colors. The no of images for training are 800 images and 360 images for testing. For evaluated the proposed system builds this database because we need a real database for Iraqi cities' names and noted in the literature works several databases for their cities' names. And it is the first step to build a big database for Iraqi cities' names which used for any recognition system. Figure- 3 shows samples of developed database.

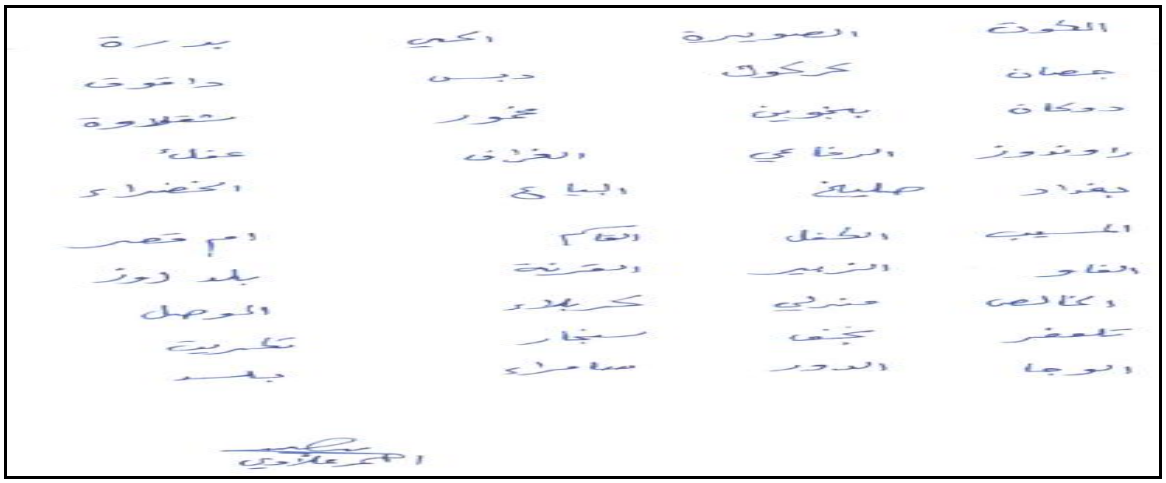


Figure 3- eamples of proposed database

4.2 Image Acquisition

In this stage convert the input image to digitally image by using a camera or a scanner with JPG format, which is suitable for a digital computer. The scanner which used in this system is Cannon I_SENSYS FM3010. After this process cut every word as an image by snipping tool program that associated with windows 7 operating system, then stored in a separated file as JPG format at 300 dpi (*dpi (dots per inch)*) with *(8 bits/pixels)*. Figure- 4 show The developed data base after cutting process.

Figure 4- samples of Iraqi database

4.3 Preprocessing Stage

All operations that apply to input image is called preprocessing operation, in order to reduce or eliminate noise data. The pre-processing stage consists of many operations Such as: converting the input image to gray image, noise removing, image binarization, Clipping and normalization Figure- 5 shows the proposed preprocessing method. Algorithm-1 illustrated the proposed preprocessing stage. In step 1,2 the input image converted to a gray scale image by using equation-1, then used the median filter with 3*3 mask to remove the noise and smooth the gray image. Figure- 5 Shows this operation.

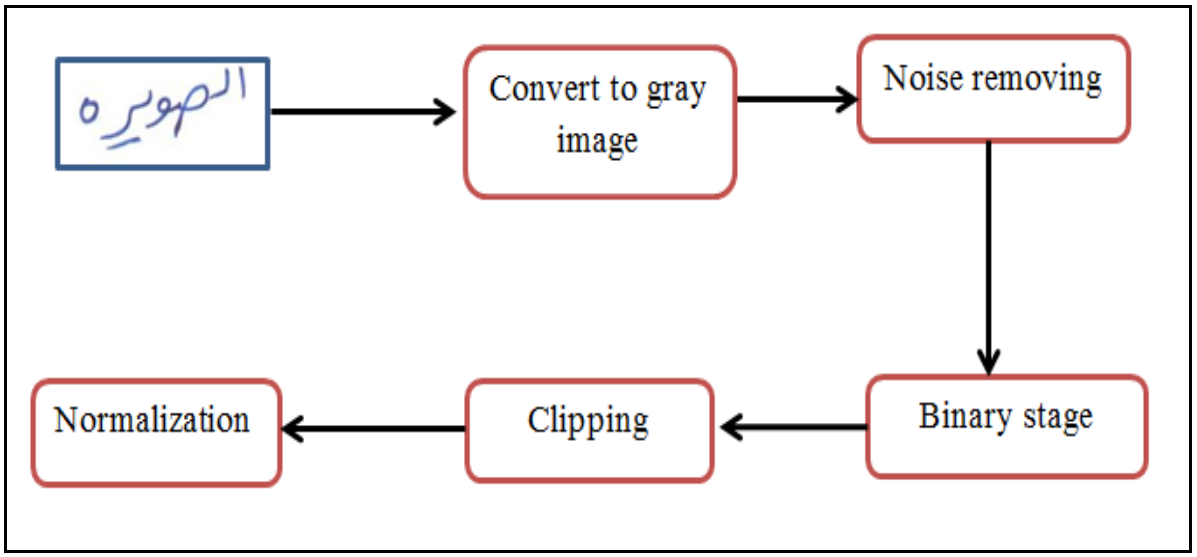


Figure 5- preprosessing stage

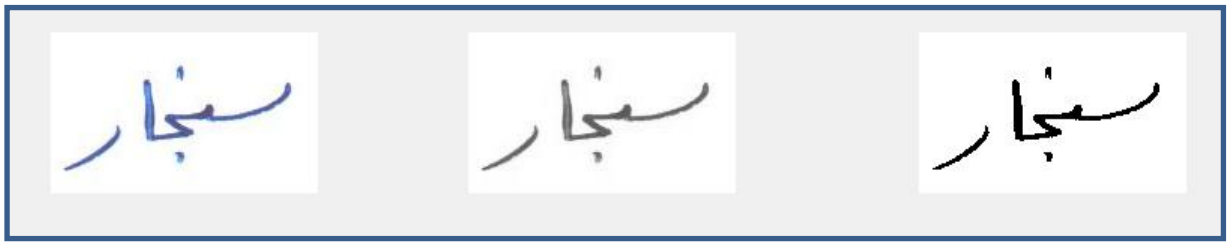
Algorithm -1: Preprocessing **Input:** word image **Output:** preprocessed word image
Step1: Convert the input word image to gray image **Step2:** Apply Median filter with 3*3 apply to remove the noise and smooth the gray image**. Step3:** use Otsu binary method to reduce the image dimension. **Step4:** clipping the word from its image 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 256*256 pixels.

Step 3 used Otsu binarization method [10] because its easy and fast binaryzation method. Figure- 7 shows Otsu binarization method which used in the proposed work. In Step 4 using the bounding box for clipping the word from its image. Figure- 6 Shows the clipping process. In step 5 using the normalization process with 255*255 size

.

(a) (b) (c)

Figure 6- Nosie Removing operation a) original image b) gray image with noise c) gray image after applying a median filter



(a) (b) (c)

Figure 7- binary method a: original image b : gray image c: Otsu binary method

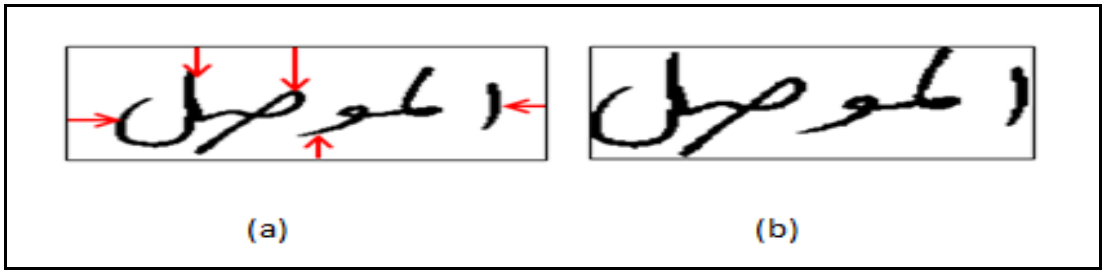


Figure 8- clipping image process a) original image b) clipping image

4.3 Feature Extraction

The feature extraction method extracted the important information from word image.In this paper used, a new method which based on DWT and sliding window methods are adapted to extract the features of the word images.

4.3.1 Proposed Feature Extraction Method

A proposed method for feature extraction is used based on DWT (Discrete wavelet transforms) and zoning methods. Algorithm 2 describes the main steps of the method. The steps 1,2 converted the normalized binary image to DWT decomposed at 4 levels. Step 3,4,5 scan the decomposed image with window 2^level*2^level without overlaps. Then find the stander derivation for this window, then puts the value in list feature vector.Step 7,8 repeat step 4,5 on the full image ,then return the feature vector (255 feature). Figure- 9 shows the proposed method.

Algorithm 2: (proposed Feature Extraction Method) **Input**: clipping binary image **Output**: Features vector1
Step1: Read the normalized binary image **Step2**: apply the DWT Haar decomposition at 4 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 vector1. **Step6**: repeat step 4 on the full image **Step7**: return vector1

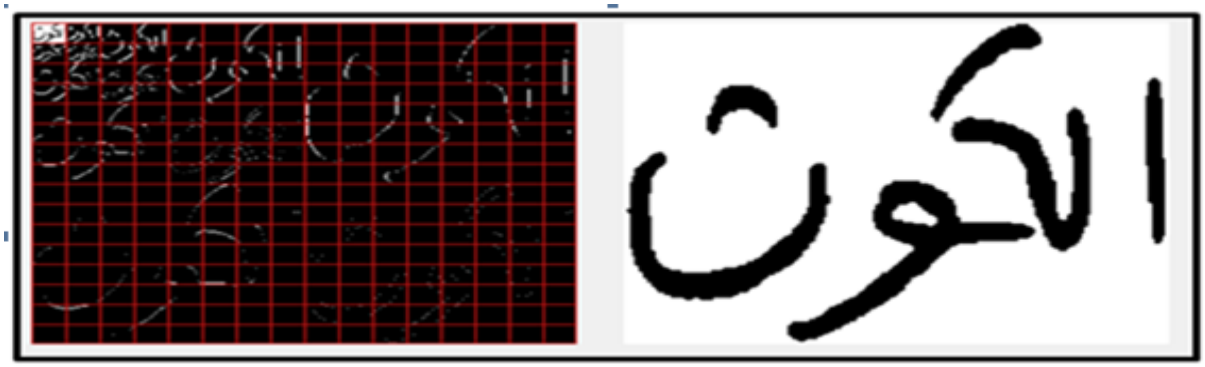


Figure 9- Decomposition word "الكون" at level 4 and scanning window (16*16) on word image

3.4 Feature Selection (FS)

In order to reduce the feature vector and choose a subset feature attributes to increase The accuracy rate used feature selection method.in this work used Backward elimination method for this proposed .Algorithm 3 illustrated the feature selection method.

Algorithm 3: **Features selection (backward method)**
Input: Feature Vector with n length
Output: Feature Vector with m length m<=n

Step1: Read the feature vector F
Step2: Calculate the test accuracy for this vector called old_acuracy
Step3: i=0
Step4: Remove I attribute from the vector
Step5: Calculate the test accuracy called new_accuracy
Step6: If old_acuracy<=new_accuracy then remove this attribute
Else remind this attribute in its location
step7:Increment I go to step3
step8: return feature vector with m length

4.5 Classification and Recognition

In the proposed system used Support Vector Machines (SVM) for classification and recognition because it is the best and efficient classifier. In the proposed system used (Accord.NET) library Ver 3.02[RE] which supports multi-class problem to classify images of handwritten word classes. There are three types of kernels which used in this library such as1-polynomial 2-Gaussian (RBF) 3-linear. And Gaussian kernel achieved a best recognition rate. Algorithm 4 illustrated the classification and recognition stage in this work.

Algorithm 4: **Classification and Recognition stage**
Input: Feature Vectors for all training images called **train_data** with length N attributes, Feature vector for testing image called **test_image**, NO of classes called **no_classes**,
Output: **class_test_label** for test image

Step1: Read **train_data**, **test_image**, **no_classs**es, N,
Step2: recalls Arecord.net functions **for SVM.**
Step3: choose the SVM IKernel type (Gaussian, Linear, Polynomial)
Step4: using Sequential Minimal Optimization algorithm to learn the SVM
Step5: train SVM to classify the training images by using the function
Step6: in order to find out the class of **test_image** use Compute() function. **class_test_label** = Compute (**test_image**).
Step7: return class_test_label

4. Experimental Results and Discussion

The visual basic.net language is used to programming all system parts except
The classifier stage, which uses the Arecord.net library which contain SVM classifier functions for multi classes. The recognition system evaluated by using a development database which contains 40 words of Iraqi city names were written by 29 writers from different ages and educational backgrounds. In this work Used 800 images for training and 360 images for testing (70% of data for training and 30% for testing). The next stage is preprocessing operations, which applied to an input image, After that the 4 level DWT have been used to extract the wavelet coefficients, then scanned window (16*16) size of DWT space. Then find stander derivation for each window. Due to the number of windows in DWT image is 255 .therfore the length of feature evector is 255 values. After that, apply feature selection method to reducing feature vector length by using the backward elimination method.
The database image has different styles and font sizes because it's written by different writers therefore most make all the images with the same size. Different experiments were performed with different normalization images.
Table- 1 shows the comparison of various normalization sizes. The 256-by-256 size gave the best accuracy.

Table 1- Accuracy for different normalization types

Normalization size	Recognition rate	Feature size
64*64	80.42	255
128*128	85%	255
256*256	90	255

Table- 2 illustrated the different levels For DWT decomposition. In 4 levels the feature vector has the longest length, but higher recognition rate of decomposition at 3 levels. Therefore the DWT at 4 levels choosed for the proposed system.

Table 2- Recognition rate for DWT at 3 and 4 levels

Normalization size	DWT level	Recognition rate	FL
256*256	3 levels	84.72	64
256*256	4 levels	90	255

Refering to Table 3 , In order to illustrated the effect the feature slection method on recognition rate used two experemants.these experemants illustrated the regnition rate and feature length before and after using this process. The first experemant appled on DWT at 3 levels , Its recognition accuracy was changed from 84.72 to 86.39 and feature length changed fron 64 to 53.second experemant applied on DWT at 4 levels ,its recognition rate was changed from 90 to 92.5 and feture length changed from 255 to 193.

Table 3- Recognition rate after using feature selection

Normalization size	DWT level	Recognition Rate	FL	Recognition Rate	FL
256*256	3 levels	84.72	64	86.39	53
256*256	**4 levels**	**90**	**255**	**92.5**	**193**

In recognition stage, apply different types of SVM kernels. Table- 4 illustrated compression the recognition rates for three types of SVM kernels. The Gaussian kernel has higher accuracy than other types which is 94.44%.

Table 4- Recognition rate for SVM types

SVM type	Recognition rate	FL	FS	FL
Linear	90.56	255	92.5	168
Gaussian	**89.17**	**255**	**94.44**	**179**
Polynomial (2)	90	255	92.5	190

5. Conclusion

A proposed system to off-line handwriting Arabic word recognition was developed based on DWT transforms and SVM with **Gaussian kernel**. The accuracy of the developed system is 94.44% . The type and size of databases have an influence on handwritten Arabic word recognition systems, so may be used another database on this system.and another classifier such as neural network or KNN can be used in future works. This system can use another feature extraction method to improve the recognition accuracy. 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. Abdul Hassan, A. K. and Kadhm, M. S. **2015.** Handwriting Word Recognition Based on SVM Classifier. *International Journal of Advanced Computer Science and Applications*, 6(11): 64-68**.**

2. El moubtahij, H., Halli, A. and Satori, K. **2014.** Review of feature extraction techniques for offline handwriting arabic text recognition. *International Journal of Advances in Engineering & Technology*, **7**(1): 50-58.

3. Ali, R. H. **2016.** Printed Arabic Words Recognition Using Genetic Algorithm. University of Baghdad, Computer Department, Education College for Women, *AL-Ustath Journal*, Number extension 217–volume 2.

4. AlKhatee, J. H. **2010.** Word Based Off-line Handwritten Arabic Classification and Recognition. MSc. Thesis , University of Bradford Open Access repository.

5. Nemouchi, S. Souici, M. L. and Farah, N. **2012**. Classifiers Combination for Arabic Words Recognition Application to Handwritten Algerian City Names. International Conference on Image and Signal Processing, 7340, Pages 562-570, Agadir Morocco**.**

6. AlKhateeb, J. H. **2011.** Word-Based Handwritten Arabic Scripts Recognition Using Dynamic Bayesian Network. The 5th International Conference on Information Technology.

7. Jalil, L. F. and Mohammed, M. M. **2016.** A Modified Back Propagation Algorithm for Assyrian Optical Character Recognition Based on Moments. *Eng. &Tech.Journal,* **34**,Part (B), No.2.

8. Sakthivel, N. and Prabhu, L. **2014.** Mean – Median Filtering For Impulsive Noise Removal. *International Journal of Basic and Applied Science*, **02**(04): 47-56.

9. Sharma, D. and Singh, S. **2015.** Binarisation Algorithms Analysis on Document and Natural Scene Images. *International Journal on Recent and Innovation Trends in Computing and Communication,* **3**(8): 5235-5245.

10. Gusain, D. Pandey, B. K. and Mandoria, H. L. **2016**. Performance Analysis of Filters on Complex Images for Text Extraction through Binarization. *International Journal for Research in Applied Science & Engineering Technology (IJRASET),* 4(I): 41-52.

11. Abdul Hassan, A. K.and Kadhm, M. S. **2016.** An Efficient Preprocessing Framework for Arabic Handwriting Recognition System. *Diyala Journal for Pure Sciences*, 12(3): 147-163.

12. Kadum, M.K., KAbdul Baqi, B. and Naser, A.G. **2014.** Offline HandWritten Letter Recognition Using Neural Networks Based on Multiple Feature Extraction Algorithms. *Iraqi Journal of Science ,* 55(4A): 1625-1637 .

13. Aghdam, M. H. Aghaee, N. G and Basiri, M. E. **2009.** Text feature selection using ant colony optimization. *Expert Systems with Applications journal homepage,* Elsevier, 36(3): 6844-6852.

14. Ladha, L. **2011.** Feature Selection Methods and Algorithms. *International Journal on Computer Science and Engineering (IJCSE),* 3(5): 1787-1797.

15. Abdul Hassan, A. K. and Kadhm, M. S. **2015.** ACRS: Arabic Character Recognition System Based on Multi Features Extraction Methods. *International Journal of Scientific & Engineering Research*, 6(10): 656-661**.**

16. Shubhangi, D. C. **2010.** A Study on English Handwritten Character Recognition Using Multiclass SVM Classifier. PhD Thesis,Gulbarbarga University,India.

17. Sagheer, M.W. He, C., Lei, Nobile, N. and Suen, C. Y. **2010.** Holistic Urdu Handwritten Word Recognition Using Support Vector Machine. International Conference on Pattern Recognition, Istanbul,Turkey,**:** IEEE, PP:1900-1903.