Journal of Theoretical and Applied Information Technology
31st May 2017. Vol.95. No 10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OFFLINE ARABIC HANDWRITTEN ISOLATED CHARACTER RECOGNITION SYSTEM USING SUPPORT VECTOR MACHINE AND NEURAL NETWORK

[1]MOHAMED AL-JUBOURI, [2]HESHAM ABUSAIMEH

[1] Master Degree in Computer Science, Middle East University, Amman, Jordan

[2]Associate Professor of Computer Science, Middle East University, Amman, Jordan

E-mail: [1]dr.mohamedanas@yahoo.com, [2]habusaimeh@meu.edu.jo

ABSTRACT

The Arabic Language had a little attention in this field compared with other languages due to the high cursive nature of the handwritten Arabic language, especially with their dots. The difficulty lies in the complexity of locating the wavy shape in the characters, which solved by the combination of certain features extraction methods that work in separate way. The proposed of Isolated Arabic off-line handwritten recognition system based on two stages classifiers (Hybrid). First stage is a linear Support Vector Machine (SVM) for splitting the dataset characters into two groups - Characters with dots and Characters without dots, by giving certain extraction features to each group. This division can reduce the error rate of characters recognition which has similar looking shape. Second stage supplies the first stage result to Neural Network (NN) stage which granted one of the best correctness and accuracy by training. Finally, a fully recognized character is acquired successfully. This work is implemented (IFN/ENIT) dataset, the system significantly reduce the load of NN process by SVM classifier, which can be used for real-time applications. A total accuracy of this proposed work reaches 92.2%

Keywords: *Arabic Handwritten, Optical Character Recognition. Support Vector Machine, Feature Extraction, Neural Network, IFN/ENIT.*

1. INTRODUCTION

Instead of storing images that contains text as "just images", an OCR system makes "understanding" what's in those images possible, by integrating the techniques and algorithms of three of the hottest areas of research in computer science: machine learning, pattern recognition, and the new "computer vision" area. Simply, an OCR system takes a scanned image that contains text, identifies the text in it and produces another electronic form of the file that is searchable and editable by a text editor. Arabic characters are considered highly cursive, which makes the recognition of Arabic characters an open and active research problem. The proposed OCR system uses the IFN/ENIT dataset (the isolated characters only) for training the system, and a same dataset to test the accuracy. A sequence of operations will be performed in order to measure the accuracy of the recognition system. Preparing the data includes transforming the scanned images from the training dataset into black and white, filtering, and then centralization and skew-correction. Then, these pre-processed datasets are divided into two main groups by SVM classifier: letters with dots, and letters without dots. These are then used to extract the main features to be used in classification with a curvelet/wavelet feature extraction algorithm to later fed into the BPNN to measure the accuracy of the classification process, the output of the Arabic handwritten Character Recognition (AHCR) system is a real-valued number that expresses the measurement of accuracy.

2. LITERATURE REVIEW

The following related work is based on IFN-ENIT with no characters segmentation applied.

Alkhateeb, et.al., (2011) proposed a combination of classifiers, the first one is Hidden
Markov Model (HMM) than re ranking to improve the accuracy, the re ranking used with topological feature extraction, the features are collected from 500 word, the accuracy reach 84.09%. The hybrid classifier approached by Bouchareb, et.al., (2008) has a satisfying result in the accuracy of recognition which reach 96% by using SVM and Principal

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

component analysis (PCA) for 1000 isolated character principal shape (main body of the characters without dots). Shanbehzadeh, et.al., (2007) proposed a method to choose the best feature extraction among others that boosted the accuracy by algorithm, the system achieved 85.59% accuracy for 3000 Farsi characters that normalize to (50x50) grid than classified by Vector Quantization. Al-HAJJ, et.al., (2007) offline recognition of cursive Arabic handwritten words is proposed using combination of multi-stage classifier generated from fusing three HMM classifiers, the system reach 90.96% accuracy rate for 6,709 images. Dupre, (2003) proposed a system that extracted from the graphemes and a hybrid HMM/ANN is used for classification. HMM is used to represent each character- categorization, while NN make the calculations of the remaining process. This system was tested and achieved an 87.40% recognition rate. Mowlaei, et.al., (2002) proposed a very fast system for recognition of handwritten Farsi characters and numerals, Haar wavelet transform and Multi-Resolution Analysis (MRA) for feature extraction has been used. This system achieved 97.24 % for recognition the isolated handwritten postal addresses. Maddouri, et.al., (2002) present a system that take advantage from feature extraction which use combination of local and global vision modelling, the system reach an accuracy 97% for 70 word. Dehghani, et.al., (2001) proposed a system that use HMM as a single classifiers with one dimensional process based on model with 6 states and 8 mixtures for 17000 Persian characters, the system reach a recognition accuracy 65% (Lawgali, 2015).

3. DATASET

The using of IFN-ENT dataset because it consist of more than 2900 various characters that have low overall accuracy, the Arabic language contain 28 isolated characters plus the Hamza, the use of this dataset because it is reliable, realistic and the most common used in such researches (Pechwitz, et.al., 2003), the dataset is designed to cover specific shapes of Arabic characters. it contains 100 sample image for each isolated character, plus the Hamza which some researchers consider it a character, the forms will be scanned in black and white mode with low resolution of dots per inch (dpi), The images of the characters shall be converted into a binary format with small objects considered as noise and removed, the dots and marks, such as " همزة ", are removed from the characters since they can affect the classification. The dots will be transfer to original location after character segmented. In this case the dots are considered as primary features to be extracted. In order to keep the originality of character image. The images of characters will not be resized and it will be used as 128x128 grid for normalization purpose. The dataset will be used in this study for training and testing. The implemented recognition system will use a free segmented dataset, which requested from IFNENIT.com, (Lawgali, et.al., 2013).

4. PREPROCESSING

The images in the dataset are clear and without noise, either way the system will be going to include Thinning as noise reduction. Thinning: this is a process used to simplify the shape of text and reduces the data amount required for handling in order to come up with connected character Skelton of the image input. In a real time system the need to remove noise from the images is required, even if the noise is barely observed. Median filter will be included in the system, since images in the original dataset have different sizes. As a preprocessing step in this system. The normalization of the size is required for all the images by changing their block size. The given dataset image come in one size of grid, which is 128×128 grid. The normalize technique changes the available block size into another by cutting the edges (Blob). This process compares the level of similarity and differences in the processing data (mostly images or kinds of visual data). Blob detection was used to obtain zones of interest for additional processing. The process of rotating the normalized data to enhance the alignment of character and facilitate the capture of visual data is called skew correction. Binarization is the process of changing available data before normalization into binary matrix which can be understood by the computers (machine language). In general, noise filtering, smoothing, binarization, skew-correction and normalization should be done in this pre-processing step. (Sadri, et.al., 2003). The Optical Character recognition (OCR) objective is to convert the text into digital image using computer vision. There are two categories of automatic handwriting recognition, these are the online and offline recognition. Handwriting recognition using the online method is easier than the offline because the script temporal information is available (AL-Zawaideh, 2012). One of the most challenging task is recognition of Arabic handwriting character (Plamondon, et.al., 2000).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. SEGMENTATION

Reaching the best recognition depends on finding the appropriate segmentation method/model. The key models are estimated baseline, line segmentation, thinning foreground, dots extraction and no character segmentation. Segment the character from the background is done by ref (x) in this steps. Ideally, the character segmentation depends on identifying the ending and starting of sub-words/words and identifying the point of segmentation among the characters in the sub words/words (Aouadi, et.al., 2016). The segmented dataset are downloaded from IFN-ENIT. The segmentation process is no longer required, because there are many kinds of researches about IFN-ENIT segmentation that already reach a high accuracy segmentation result, the choice of using pre-segmented dataset is to focus on recognition accuracy instead of segmentation process, the used Isolated Arabic dataset achieved 90.8% accuracy of segmentation.

6. FEATURE SELECTION AND EXTRACTION

After the data has been prepared, the proposed system will extract designated features for each of the character groups defined by the feature extraction techniques. Since Arabic letters are considered cursive letters, so the most appropriate method to use in order to extract the main features of the training dataset would be curvelets and wavelets feature. The determination of the importance of each feature by considering how the performance is influenced without that feature is needed. If removing a feature deteriorates the classification performance, the feature is considered important (Chang, et.al., 2008), with curvelet and wavelet based feature extraction (Separated) is the best way to achieve a high overall accuracy rate. Based on object-counting which is basically a pixels boundary detection and extraction in morphological operation. Considering that the number of feature extractions is not important, but the quality of retrieving of each one is, because it may not correspond with the classifiers, and for that a training of the system in which feature extraction is suitable is in demand for the scanned character in this system. Both DWT and curvelet are used to detect the feature components in images. Wavelet features are used due to its performance of finding low and high pixels density frequency in the letter and detect in which region of area the dots going to be. It is performed on rows first then on columns, H and L denote high-pass and low-pass channels respectively. The DWT divided the image logically into four parts LL, LH, HL and HH, each one represents a quarter part of the image, to locate the features of dots (Hiremath, et.al., 2015). On the other hand, curvelet is used to study the shape, main body and the direction of each continuity of the character. The structuring element consists of a pattern specified as the coordinates of a number of discrete points relative to some origin. The following steps clear the feature classification:

- Grid coordinates are used to represent the element as a small image on a square grid in IFN-ENIT, the grid is represented in 128×128 pixels.
- In each case (character image) the origin is marked by a ring around that point, the origin does not have to be in the center of the structuring element (character foreground pixels).
- One of the reasons that the accuracy is lowered that the grid is not modified, so the structuring is done by translating the structuring element to various points in the input image, and examining the intersection between the translated kernels coordinates and the input image coordinates. The basic effect of the operator on a binary image is to isolate the boundaries of regions foreground pixels.

The operator takes two pieces of data as inputs, the first is the image which is to be trained, The second is a set of coordinate points known as a structuring element which is the testing image (also called the desire output image).

The SVM takes these two pieces and calculate the difference between them, then put the predicted value according to the calculation.

The selection of features from The wavelet has a limitation in this case because it will split the image into several parts (theoretical), somehow the Siin (س) look just like the noon (ن) from respective of view except the top right angle as shown in Figure 1, however the problem can be reduced by curvelet feature extraction, and for that the use of dual extraction method are required so it can reduce the defects while extracting the characters, dual stage of classifiers-feature extraction that reduce the error rate which is generated from the difference between them.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

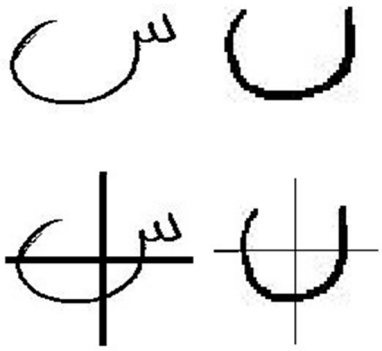

Figure 1 Seen and Noon Characters

7. CLASSIFIERS

A. SVM Classification

Classification is a general process related to categorization, the process in which main body and objects are recognized, differentiated, and understood. In the case of isolated Arabic handwritten characters, SVM classification object is to split the characters with dots and characters without dots from each other. This makes the NN process easier to recognize them as shown later on in chapter 4. The difference between characters with dots from characters without dots, lead to a major reduce in the error rate on some characters within the dual stage classifiers. By other means the probability of characters similarity with the same shape will be reduced, when increasing the categorization by object-counting feature of those letters according to their main body shape and its location, as well as to their object (dots).

B. Backward Propagation of Errors BPNN

BPNN algorithm can be explained in the following steps:

- The error term that is concluded from the output can be measured by the difference of nodes weights that generate the error.
- The error term will be saved as a value and return backward from the output layer to the first step after input (first hidden layer).
- Within each layer, the system will compare and calculate the current layer weights with the desired output weights, trying to reduce the gap between them.

After using SVM, the output of SVM comes in two groups, each group contains mini-small groups of letters, after that, each group is entered to BPNN in order to apply and compute accuracy. Training the system by initializing the network with random weights, the random weights are assigned from the NN to break the symmetry and this makes the neural network learn faster (Rojas, 2013). Symmetry is a type of invariance: the property that something does not change under a set of transformations. The process starts from the last layer which is called the output layer heading to the first one (input layer), comparing the network random weights with the actual output weights this process is called (error function), then modify or adjust the layers one by one heading to the source and calculate real value to random weights (weights updating) until the input weights are nearly the same of desire weights, mixing ANN with SVM reduce the probability of classification error. The optimization is not a single or standalone step, it is combined with several parts of the ANN process. Optimization ensures that the input pattern that is taken from the background has the best quality, in optimization the feature extraction deals with each character in a different way in order for better classification.

8. HARDWARE SPECIFICATION & IMPLEMENTATION

For this system an ASUS laptop computer has been used as a working area for installation and implementation with the following specification:

- Windows 8.1 64 bit.
- Processor: Intel Core I7-4720 HQ @ 2.6 GHz.
- RAM: 16 GB Ram DDR4.
- GPU: NIVIDIA 960 GTXM.

Arabic letter recognition system model is constructed using MATLAB2014A, using SVM and neural networks. In order to guarantee the main objective of this thesis by analyzing the main concepts and the performance of Arabic letter recognition system to obtain results and compare them. Data Base loading: the dataset character images used as a total 2928 for both training and testing is loaded. The images used for testing is 439 image, the images used for validation is 439 and the images used for training is 2049. The flow chart (Figure 2) shows the numbers of images used, all the characters are loaded within one database, and no real separation is applied.

- Data Set Classification: classify dataset each letter in a specific table cell.
- Image Storing: both the training and testing images is being saved as different variables
- Feature Extraction: curvelet and wavelet feature extraction are applied to recognize

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

isolated characters shapes and stored in the temporary database.

- SVM: the images variables are being entered into a binary SVM, to classify them into two groups, with and without dots to decrease the number of classified groups and though enhance the accuracy.
- Neural Network: Apply training images to get neural network recognition learning using back propagation method in order to get optimum weights at specific mean square error.

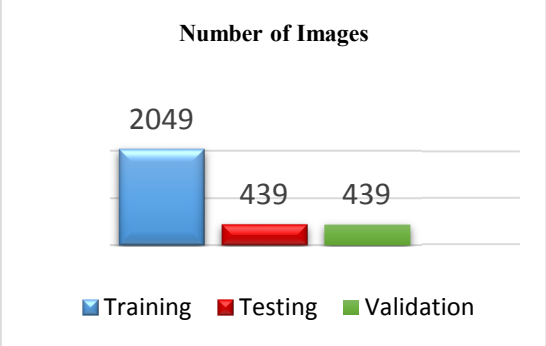

Figure 2 Number of Images Used in This System

9. EXPERIMENTAL RESULT

I. Without SVM stage

The single Neural Network classifier has a low overall accuracy, the reason of that is the NN cannot deal with the proposed features-extraction together, as well the categorization of the characters take more time and resource to assign the character to its group, also there is no specified categorization that limit the characters similarity of shape like seen (س) and Sheen (ش), Taa (ت) and Thaa (ث) and so on that difference generates a lack of precision in recognition process. and adjusted according to the desired output one time at least and all of the training samples pass through the learning algorithm. The NN performance in this code is equal to 0.0527.

The gradient algorithm is an optimization algorithm used to find a local minimum of that function, the detector goes to the negative of the gradient of the function at the present point, and the number one represents the local maximum of that function. In addition to that, the min-gradient is equal to 0.00501.NN-tool shows the values for the optimized learning that occurred at epochs 54. Figure 3 shows Neural Network Tool

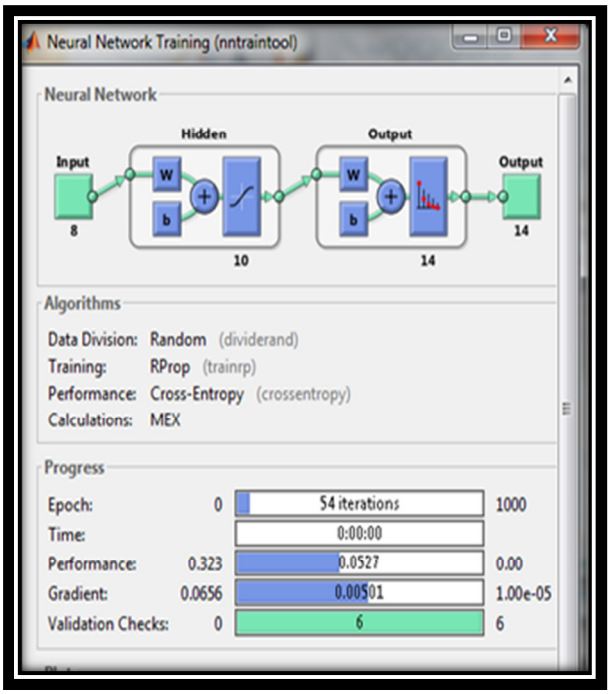

Figure 3 Neural Network Tool

The result of NN-tool is shown in Fig 4, clearly the optimized learning occurred at epochs 54, cross entropy around 0.1 which is the point that connect training, testing and validation.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

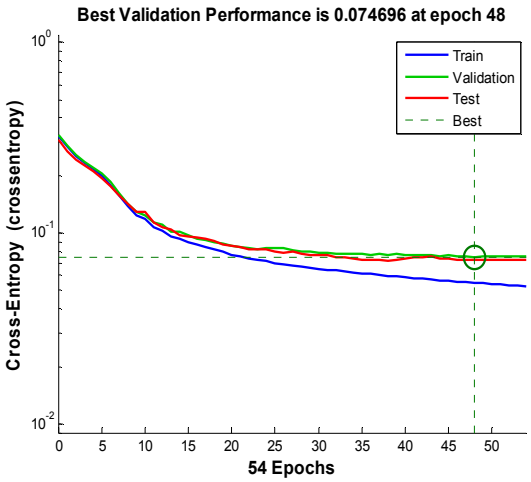

Figure 4 Best Validation Performance

II. Hybrid Stage

The Hybrid approached solve this particular problem by using the SVM classifier as categorization process which remarkably increased the recognition rate chart (Figure 5) shows the accuracy percentage of the hybrid system.

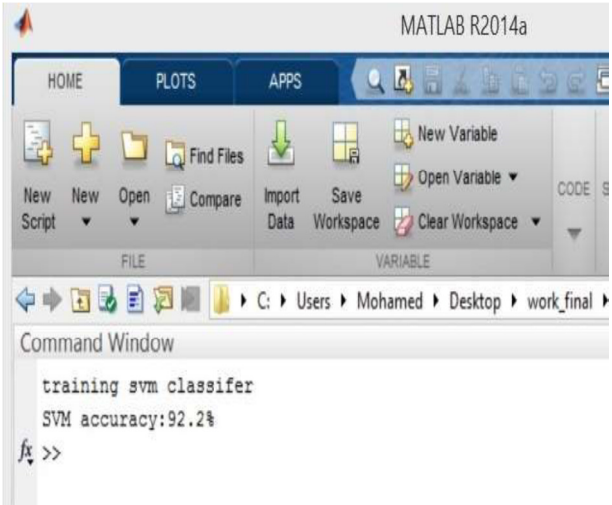

Figure 5 System Accuracy Percentage

There are two main groups in this work, and each group contains many sub-groups. The group that contain the characters with dots are divided according to the number of object and directions of that object. The group that contain the characters without dots are randomly divided, because the system depend on object-counting technique. Which mean only characters with dots will be highly categorize. Maximum accuracy of letters can be reached up to 99% while some letters are difficult to recognize, because it recognizes in the wrong group in SVM-stage previously. The character (ق) has the highest recognition rate, because the two dots, character continuity and domestic long-bending (Curved inwards) make it the most recognizable character, where the character (ه) has the lowest one. Table 1, shows the worst cases in recognition of the isolated characters in this system, the worst case among them is Haa **(ه)** in both single and Hybrid classifiers, the reason for that is the lack of feature extraction that extracts and detect the holes, Zaay (ز**)** is second rated case as worst character, because it has similarly structure element as character Laam (ل **)** the instance represent the character image that have classification error form the 439 test set.

Table 1 Worst Cases Classification Errors in Isolated Character

No	Letter	Instance/439-Error rate%	Often mistaken for
1	Haa (ه)	24 -1.7%	و
2	Zaay (ز)	20 -1.4%	ل
3	Alif (ا)	17-1.2%	و
4	Thaa (ث)	17-1.2%	ش
5	Sheen (ش)	15-1.0%	ث
6	Faa (ف)	14-0.9%	ز

The chosen characters above are the characters that reach 0.9 error rate, since the error rate percentage equals (100%-recognition percentage) for the hybrid proposed system the overall error rate reach as (7.8%), and there is much more characters in the neural network as a single classifier that reach up to (28.05%) error rate.

10. CONCLUSIONS AND FUTURE WORK

In this Paper, the process of Arabic Letters recognition based on using both the neural network with support vector machine technique is proposed. This process starts with dividing known Arabic Letters images dataset into two databases, applying the feature extraction technique on each dataset. The developed algorithm can be used in order to overcome the main restrictions of using the traditional NN algorithms, which depends on one classifier only. SVM-NN can provide high accuracy with low time processing especially with a huge dataset. SVM classifier provides accuracy up to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

92.2% by dividing the whole dataset into two groups for decreasing processing time for NN stage. The probability of error tends to be zero as the categorization of characters increases. Results show that there is a match between both the recognition rate and the success values of the algorithm and the resultant maximum recognition rate is 99% using confusion matrix scheme. Despite of the computational complexity of this system, the classifier is suitable for real time applications because the run time is acceptable.

For this Paper, the researchers can recommend the following ideas: In future, many solutions can be used in order to overcome the disadvantages that face this project, for example, some enhancement steps can be done on the algorithm to decrease its complexity as well as develop the accuracy of the model in addition to the efficiency of the retrieval.

- Another environment of the database also can be used which is the online database environment that permits the addition and removal of images in order to use the database in easy and simple way than that used with the offline database.
- Further theoretical analysis is needed to find further optimality in choosing the number of layers, in addition to the number of neurons per layer using the back propagation neural network.
- Since the handwritten Arabic characters word and sentences are required, the recommendation is to use the proposed system in the recognition of such characters, the capability of achieving a higher recognition rate could be done.
- The expectation of a higher recognition accuracy from this system for printed characters since it is smoother and well defined.
- The real-time system has a limitation in such system, building a pre-extracted features database from the characters, which is provided via cloud could accelerate this particular process.

11. ACKNOLEDGMENT

The authors are grateful to the Middle East University, Amman, Jordan for the financial support granted to cover the publication fee of this research article.

REFERENCES

[1] Al-Hajj R, Mokbel C, Likforman-Sulem L. Combination of HMM-based classifiers for the recognition of Arabic handwritten words. InDocument Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on 2007 Sep 23 (Vol. 2, pp. 959-963). IEEE.

[2] AlKhateeb JH, Ren J, Jiang J, Al-Muhtaseb H. Offline handwritten Arabic cursive text recognition using Hidden Markov Models and re-ranking. Pattern Recognition Letters. 2011 Jun 1;32(8):1081-8.

[3] AL-Zawaideh, F. H. A, (2012). A Three Stages Segmentation Model for a Higher Accurate off-line Arabic Handwriting Recognition, *WCSIT 2 Vol. 3,* 2012.

[4] Aouadi N, Echi AK. Word Extraction and Recognition in Arabic Handwritten Text. International Journal of Computing & Information Sciences. 2016 Sep;12(1):17.

[5] Bouchareb F, Hamdi R, Bedda M. Handwritten Arabic character recognition based on SVM Classifier. InInformation and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on 2008 Apr 7 (pp. 1-4). IEEE.

[6] Chang YW, Lin CJ. Feature ranking using linear SVM. InWCCI causation and prediction challenge 2008 Jun 4 (pp. 53-64).

[7] Dehghani A, Shabini F, Nava P. Off-line recognition of isolated Persian handwritten characters using multiple hidden Markov models. InInformation Technology: Coding and Computing, 2001. Proceedings. International Conference on 2001 Apr (pp. 506-510). IEEE.

[8] Dupre, X. *Reconnaissance de l'ecriture manuscrite* (Doctoral dissertation, PhD thesis, Univ Rene Descartes-Paris V), 2003.

[9] Hiremath TR, Patil SM, Malemath VS. Detection and Extraction of Text in Images using DWT. Int. J. Adv. Res. Comput. Commun. Eng.. 2015;4(6):533-7.

[10] [IFN/ENIT - database Arabic OCR handwritten arabic word recognition, Arabic database. *(n.d.). Retrieved January 24, 2017, from* http://www.ifnenit.com/

[11] Lawgali A. A survey on arabic character recognition. International Journal of Signal Processing, Image Processing and Pattern Recognition. 2015;8(2):401-26.

[12] Lawgali A, Angelova M, Bouridane A. HACDB: Handwritten Arabic characters database for automatic character recognition. InVisual Information Processing (EUVIP), 2013

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4th European Workshop on 2013 Jun 10 (pp. 255-259). IEEE.

[13] Maddouri SS, Amiri H. Combination of local and global vision modelling for arabic handwritten words recognition. InFrontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on 2002 (pp. 128-135). IEEE.

[14] Mowlaei A, Faez K, Haghighat AT. Feature extraction with wavelet transform for recognition of isolated handwritten Farsi/Arabic characters and numerals. InDigital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on 2002 (Vol. 2, pp. 923-926). IEEE.

[15] Pechwitz M, Maergner V. HMM Based Approach for Handwritten Arabic Word Recognition Using the IFN/ENIT-Database. InICDAR 2003 Aug 3 (Vol. 3, p. 890).

[16] Plamondon R, Srihari SN. Online and off-line handwriting recognition: a comprehensive survey. IEEE Transactions on pattern analysis and machine intelligence. 2000 Jan;22(1):63-84.

[17] Rojas R. Neural networks: a systematic introduction. Springer Science & Business Media; 2013 Jun 29.

[18] Sadri J, Suen CY, Bui TD. Application of support vector machines for recognition of handwritten Arabic/Persian digits. InProceedings of Second Iranian Conference on Machine Vision and Image Processing 2003 Feb (Vol. 1, pp. 300-307).

[19] Shanbehzadeh J, Pezashki H, Sarrafzadeh A. Features extraction from farsi hand written letters. Proceedings of Image and Vision Computing. 2007 Dec:35-40.