Handwritten Arabic Character Recognition Based on Minimal Geometric Features

Majid Harouni, Dzulkifli Mohamad, Mohd Shafry Mohd Rahim, Sami M. Halawani, and Mahboubeh Afzali

***Abstract*— On-line handwriting recognition is one of the most successful applications in the area of pattern recognition. Though this field is quite matured, yet the research issues are still challenging, particularly in handwriting character recognition, where the problems are still wide open. The OCR system for printed characters is almost done, though it cannot guarantee for 100% accuracy. However, the research works in recognition of Arabic handwriting are still at the beginning and require more attention. This paper presents the novel on-line Arabic handwriting character recognition. An efficient approach is introduced here to divide it into some particular component. A set of features are extracted from these components, and then encoded for the classification stage. The system classification is implemented by using two processes, i.e. weight initialization in back propagation, and with multilayer perceptron neural network. Finally, the proposed system was tested on a database of Arabic handwritten samples.**

***Index Terms*—Feature extraction; on-line character recognition; classification.**

I. Introduction

On-line handwriting recognition is simply a process of determining whether a particular handwritten letter contains a language alphabet or otherwise. It is the task to present what letters or words are written without keyboard via an entrance device, such as optical pen or digital pen, on a surface or tablet pc. Various techniques have been proposed and reported different results by research groups across the world; they often are, the diverse results when comparing them. In fact, the inherent feature of on-line handwriting recognition is to have the dynamic data input based on the different style and speed of writing, therefore, almost of all these techniques could not be tested on the same data set and then apparently it is unfair to compare how they did against each other. Generally, the outline of on-line handwriting recognition system contains four stages, pre-processing, segmentation, feature extraction and reorganization. It may or may not involve the entire stages. Each of these stages can be divided into some sub-stages or has different kinds; such as pre-processing stage which can be divided into interpolation sub-stage and pre-segmentation sub-stage, or structural and statistical features for feature extraction stage as its different kinds. The sufficient input data for recognition stage can be collected from the first three stages so that they can directly affect on performance accuracy of the system as well.

Many different methods have been applied to this problem using neural network so far. The Multi-Layer Perceptron (MLP) with Back-Propagation (BP) classifier was used to classify handwritten katakana and reported recognition rate of 97.8% [1] and have sued by [2] for recognizing handwritten Persian\Arabic character using a deductive method, followed by [3] suggested a classification and done with using a multilayer perceptron network with back-propagation learning for Arabic handwritten character recognition, and also [4] with handwritten characters. In [5], it used the MLP with one hidden layer, for the connecting weights estimated by the error back-propagation (BP) algorithm minimizing the squared error criterion for recognition of handwritten Bangla and Farsi numeral characters. This paper is organized as follows. Section 2 briefly depicts Arabic alphabets. Section 3 presents an algorithm that how to provide a desired data from hand-drawn letter or a raw data set. Section 4 gives the definition of extracting features using geometrical function. Section 5 designs the structure of a BP and MLP classifier using two methods in its weight initialization. The experimental procedure and results are stated in section 6. Section 7 concludes the paper.

II. Overview of Arabic Script

The Arabic script is one of the most used scripts in the world; it is written in horizontal lines from right to left and mostly connected to the base line. The Arabic script consists of 28 letters. In fact, the Arabic letters can emerge in up to four different shapes; for instance, character “Ain “: isolated “ع”, initial “عـ” , middle “ـعـ” , and final “ـع” , it depends on whether it is located at the beginning, the middle, the end or alone on context into a word or its own. The vowel signs (hamza, fat-ha, kesra, sukkun, nunation diacritic with fat-ha, etc.) usually used in the Arabic words.

The following list shows some character attributes of

Manuscript received June 9, 2012; revised September 19, 2012. This research is supported by the Ministry of Higher Education (MOHE) and collaboration with Research Management Center (RMC) Universiti Teknologi Malaysia (UTM). This paper is financial supported by GUP Grant (NO. VOT: Q.J130000.7128.01J18).

Majid Harouni is a Ph.D. candidate at UTMViCube Lab, Faculty of Computer Science and Information Systems, Universiti Technologi Malaysia, P. O. Box 81310, Skudai, Johor, Malaysia and with the Department of Computer Science, Islamic Azad University, Dolatabad branch, Isfahan, Iran (e-mail: majid.harouni@ gmail.com).

Dzulkifli Mohamad, Mohd Shafry Mohd Rahim, and Mahboubeh Afzali are with the UTMViCube Lab, Faculty of Computer Science and Information Systems, Universiti Technologi Malaysia, P. O. Box 81310, Skudai, Johor, Malaysia (e-mail: dzulkifli@utm.my, shafry@utm.my, and afzali_mahboobeh@yahoo.com respectively).

Sami M. Halawani is with the Faculty of Computing and Information Technology, King Abdul Aziz University, Saudi Arabia (e-mail: halawani@kau.edu.sa).

Arabic alphabets, which are very helpful to promote better understanding the intricacy and elaboration of them using in the recognition system.

- The fixed number of dots is used for some of the letters in four different shapes and up to three.
- The Arabic script can be introduced in nine-group of similar body character, and adding one group for others.
- The letter maybe has dissimilar body in its own different shapes.

III. Pre-Processing and Dividing Letters

In this section we will introduce the base of an algorithm for pre-processing stage, so that the desired data will be collected by this algorithm from on-line hand-drawn letters (called the raw data). Interpolating or smoothing, resembling, normalizing and pre-segmenting of the raw data inputs are the sub stage of pre-processing stage. Some of these stages are utilized for implementing a research by decision making techniques. The following algorithm divides a hand-drawn letters into the some strokes (pre-segmentation), and dividing each stroke into some sub-stroke (tokens). The main idea of this algorithm is to find the local maximum and minimum points of this sequential list as strokes' tokens.

In algorithm 1, each stroke is easily obtained by assisting the pen-down and then the pen-up. A hand-drawn letter is arranged by a sequential list of (x,y) coordinate pairs as input data set for the above algorithm. Finally, based on the following two-point measurements, we conclude that;

- If the number of tokens become too much, then we face with increase of execution time and vice versa.
- If the number of tokens be too much, then we have high resolution and vice versa.

Algorithm 1. Dividing a Stroke into its Tokens

Begin
Strokes: *$S_{1...}S_k$, where k is the maximum number of strokes in a letter.*
For *j=1 to k* ***do***
Select minimum and maximum of Xs and Ys as X_{min}, X_{max}, Y_{min} and Y_{max}
End for
$X= X_{max} - X_{min}$ and $Y= Y_{max} - Y_{min}$
If *$X \geq Y$* ***Then***
HDLLenght= Horizontal
Else
HDLLenght= Vertical
End if
If *HDLLenght = Horizontal* ***Then***
Finding *Local Maximums of Fs (F(X,Y)).*
For *k=1 to m* ***do***
Assume that f_k (X,Y) is a local maximum point. So, to calculate of f_k the values of $f_{k\pm1}(y)$ on both sides of $f_k(y)$ are larger than it, and then there must be a local maximum of horizontal axis of the coordinates at the f_k.
End for
Else
For *k=1 to m* ***do***
Assume that f_k (X,Y) is a local maximum point. So, to calculate of f_k the values of $f_{k\pm1}(x)$ on both sides of $f_k(x)$ are smaller than it, and then there must be a local maximum of vertical axis of the coordinates at the f_k.
End for
End if
Saving *these local points as critical points; so that, they are the Start point and the End point of each token in its own stroke.*
End

IV. Proposed Feature Extraction Methods

This section briefly reviews and explains some of the prominent types of features, which are used regularly to recognize isolated letters. They are two main types, first structural features such as lines, curves, loops and arcs utilized by [6] and [7] , second statistical features such as massed pixels measured by [8] and or blend them together is utilized by [9]. In fact, feature extraction is still a challenging problem in handwriting recognition, and extracting these features is defined as the problem of obtaining an appropriate data set from the raw data, "which is most relevant for classification purposes, in the sense of minimizing within-class pattern variability while enhancing the between-class pattern variability" [10]. The following common features are obtained by geometrical functions and they use all acquired critical points in previous section. So, locating these features in each token and stroke signify different properties of letter for classifying or recognizing purposes.

The first feature (Equation 1) is the direction of the straight line of each token from both the start point and end point. Fig. 1 shows the direction of a token. In sum, this feature is measured by the following geometric function and used by some researchers [2], [7], [11] and [12]. In this research work, this feature is used for both the tokens and the length of main body letter.

$$X_{direction}^{Token_n} = \tan^{-1}\left(\frac{Y_{end} - Y_{start}}{X_{end} - X_{start}}\right) \quad (1)$$

where the Yend ,Ystart, Xend and Xstart are coordinates of starting and ending points of each token (Fig. 1).

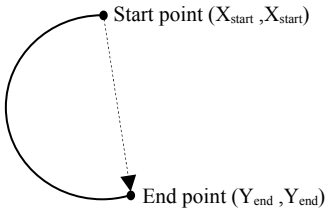

Fig. 1. The starting and the ending points

The second feature is the orientation of each token which is either clockwise (CW) or counter clockwise (CCW). [12], [13], [2] and [7] used this feature too, (As it shown in Fig. 2).

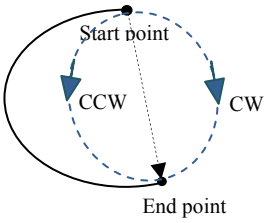

Fig. 2. Orientation side of a token

Finally, this feature (see Equation 2) is the ratio of length of each token in terms of the total length of stroke.

$$X_{tokenRatio}^{token(n)} = \frac{lengthT(n,i)}{lengthS(i)} \quad (2)$$

Where n and i are the token number and stroke number respectively; and length T(n,i) stands for the length of *n-th* token in *i-th* stroke.

V. Classification

The proposed classification stage is based on the features extracted in binary input and using a Back Propagation with a Multilayer Perceptron (BP/MLP) neural network for recognizing the Arabic letters. The output value is a decimal code that relates each letter. The activation function in the BP/MLP neural network is a sigmoid function. In order to evaluate and to compare the total program execution time as well as training time, two weight initialization methods [14] were applied as follows:

- Method 1: It is based on this equation; $|w_o|<|w_i|$, where w_o is the threshold value and w_i is the rest of weights. The weight initialization was obtained from generating random numbers in the range of [-0.5,0.5], and threshold values were assigned by the maximum random numbers found.
- Method 2: We only initialized the weights at random with a uniform distribution inside the interval of [-0.05, 0.05].

The BP/MLP is implemented in three layers as shown in Fig. 3.

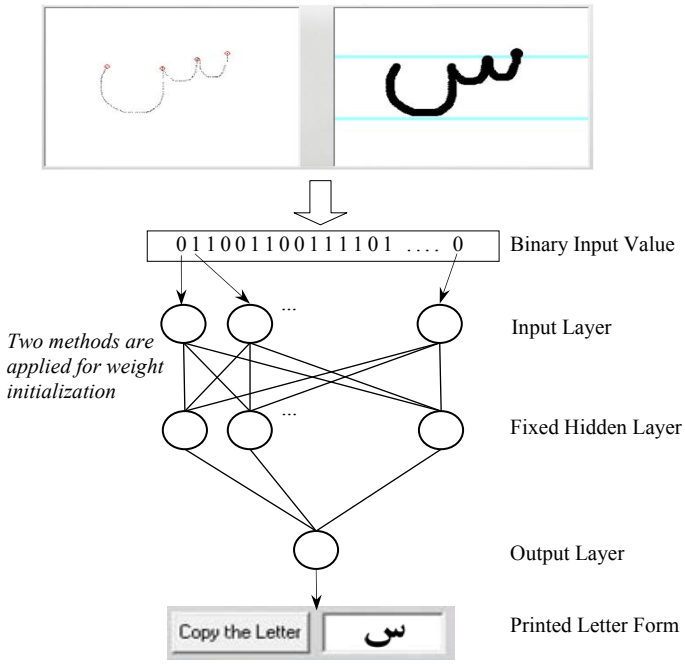

Fig. 3. The structure of the proposed classifier

The purpose of using BP and MLP in our proposed system is to increase high accuracy of classification in a main group of body letter. For training and testing the classifier, all features should be swap to binary input value of neural network architecture as same coding format. So, the result of this architecture is based on classifying of the training and testing data sets. After the training level is done completely, the testing level comprises of presenting the sample query characters to the BP/MLP and then introduces which class is belonged to the sample query characters. The matching class includes to which main group of body letter is similar to the query character, from the standpoint of the feature extracting.

As the BP/MLP neural network is a supervised learning method and thus the number of its inputs and outputs should be fixed. According to the dividing stroke section, the maximum number of tokens in all character is ten, and it is

clear that there are three features which belong to each token and one for main body letter, which are Direction, Rotation and Token-length ratio. Three bits are adequate to represent the feature of the Direction; one bit and two bits are sufficient for signifying the Rotation and the Token-length ratio respectively. Logically, it is a formal to guess that input numbers of the BP/MLP neural network can be roughly selected from one and a half of the maximum token number, which is one hundred. On the other hand, the sole output of the system is in terms of decimal values. Each decimal value introduces one Arabic character.

VI. Experimental Results

The proposed Arabic recognition system was implemented using Visual Basic. It has been evaluated using a sample test data set which was obtained by 25 different writers, they who wrote each letter in 10 times. It's important to note that whereas all processes of on-line handwriting recognition are dynamic and thus there may be no possibility of comparison with a same test data set for all researchers.

The pre-processing stage is for dividing each strokes of letter into the some tokens. We were used the proposed algorithm for finding critical points of each stroke; these points helped us to know the minimum tokens in the stroke, hence extracting the defined features are basically based on these points. All features were extracted in terms of these tokens; therefore they could convert into special binary data as inputs of the BP/MLP neural networks system. As mentioned before, the two methods were applied into the BP/MLP neural network as weight initialization; the first method is the base of weight initialization one and just used random values between some pre-defined range and it should be some numbers between minimum weight and maximum weights. The second methods is asked to produce the same random numbers between minimum weight and maximum weights with the difference that you are going to change process threshold dynamically, so in this case, it should get the minimum number of weights for its new threshold and then update them. Table 1 (next page) shows the results of the character recognition system. In the BP/MLP architecture for both methods, 70% of samples are selected for training and 30% testing samples, threshold: the value of target mean square error and the training stops once it is achieved, the learning rate and momentum were 0.0, 0.1 and 0.5 respectively, and we applied 5, 10 and 20 number nodes of hidden layer in which the best result was acquired by 20 nodes.

VII. Conclusion

Here we have probed a dynamic supervised back propagation with multilayer perceptron neural network as a classifier to be applied to on-line Arabic character handwriting recognition. The simple classifier is utilized the two methods of the weight initialization, we have obtained 96.50% and 93.01% rate recognition respectively for test data set (see Fig. 4). As results shows, this system has been confirmed quite successful results in recognizing isolated Arabic letters, and in comparison with the two mentioned methods: method 1 was obtained better result than method 2. We are now working on a segmentation of the handwritten cursive words within its isolated letters to acquire a better result for Arabic cursive words in various handwriting styles.

TABLE I: The Recognition Rate for Each Character

Character	Printed shape	Character Rec. Rate		Character	Printed shape	Character Rec. Rate	
		Method 1	Method 2			Method 1	Method 2
Alef	ا	100	100	Zad	ض	94.8	95.6
	آ	99.6	95.2	Tah	ط	95.2	91.6
Beh	ب	97.2	93.6	Zah	ظ	94.4	91.6
Teh	ت	96	90	Ain	ع	98	96
Theh	ث	94.4	91.6	Ghain	غ	96.4	92.8
Jeem	ج	97.6	91.2	Feh	ف	95.6	89.6
Heh	ح	98.4	88.8	Ghaf	ق	97.2	91.2
Kheh	خ	97.2	90.6	Kaf	ک	96.8	94
Dal	د	99.6	96.4	Lam	ل	93.6	93.6
Thal	ذ	98.4	94	Meem	م	96	93.2
Reh	ر	100	98	Noon	ن	94.8	95.2
Zeh	ز	98	95.6	Waw	و	92	94
Seen	س	93.6	90	Heh	ه	94	88.4
Sheen	ش	96	91.2	Yeh	ی	95.6	88.8
Sad	ص	98	95.6				

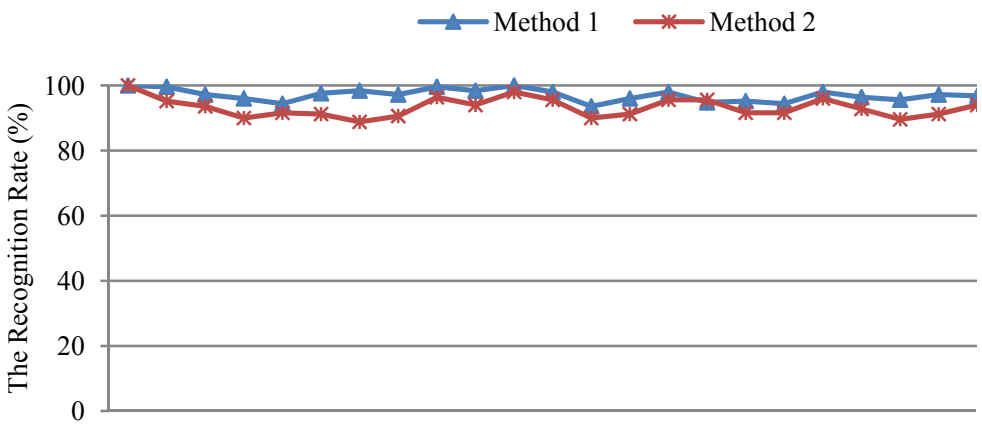

Fig. 4. The performance of the system using two different initializations methods

ACKNOWLEDGEMENTS

This research is supported by the Ministry of Higher Education (MOHE) and collaboration with Research Management Center (RMC) Universiti Teknologi Malaysia (UTM). This paper is financial supported by GUP GRANT (NO. VOT: Q.J130000.7128.01J18).

REFERENCES

[1] Takeshi A., Hiroki T., and Hiroshi N., "Recognition of Handwritten Katakana in a Frame using Moment Invariants Based on Neural Network," *IEEE International Joint Conference on Neural Network*. pp. 659-664J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73, 1991.

[2] Harouni M., Mohamad D., and Rasouli A., "Deductive method for recognition of on-line handwritten Persian/Arabic characters," *Computer and Automation Engineering* (ICCAE), 2010 The 2nd International Conference on vol.5, no., pp.791-795, 26-28 Feb. 2010.

[3] Altuwaijri M. M. and Bayoumi M. A., "Arabic Text Recognition Using Neural Networks," *International Symposium on Circuits and Systems* (ISCAS 1994). London. IEEE: pp.415–418, 1994.

[4] Chung Y. Y., Wong M. T., and Bennamoun M., "Handwritten Character Recognition by Contour Sequence Moments and Neural Network," *IEEE International Conference on System, Man and Cybernatics*. pp. 4184 – 4188, 1998.

[5] Cheng Lin and Ching Y. Suen, "a New Benchmark on the Recognition of Handwritten Bangla and Farsi Numeral Characters," *ICFHR*, 2008. Montreal, Canada. Aug. 2008. pp: 278-283.

[6] H. Almuallim and S. Yamaguchi, "A method of recognition of Arabic cursive handwriting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Washington, DC, USA, 1987, 715-722.

[7] M. Soleymani Baghshah, S. Bagheri Shouraki, and S. Kasaei, "A Novel Fuzzy Approach to Recognition of Online Persian Handwriting," *5th International Conference on Intelligent Systems Design and Applications* (ISDA'05), 2005. pp. 268-273.

[8] El-Hajj R., Likforman-Sulem L., and Mokbel C., "Arabic handwriting recognition using baseline dependent features and hidden Markov modeling," in: *ICDAR*'05, Seoul, South Korea, vol. 2, pp. 893–897, 2005.

[9] J. Shanbehzadeh, H. Pezashki, and A. Sarrafzadeh, "Features Extraction from Farsi Hand Written Letters," in *Proceedings of Image and Vision Computing New Zealand 2007*, Hamilton, New Zealand, December 2007. pp. 35–40, 2007.

[10] Devijver P. A. and Kittler J., "Pattern Recognition: A Statistical Approach," Prentice Hall, Englewood Cliffs, NJ. 1982.

[11] J. Sternby, "An additive single character recognition method," in *Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition*, 2006, pp.417–422.

[12] N. Mezghani, A. Mitiche, and M. cheriet, "On-line recognition of handwritten arabic characters using a kohonen neural network," in *Proc. 8th International workshop on frontiers in handwriting recognition*: *'IWFHR'02*, pp. 490-495, Niagara-on-the-Lake, Canada, 2002.

[13] Mozaffari S., Faez K., and Rashidy-Kanan H., "Recognition of Isolated Handwritten Farsi/Arabic Alphanumeric Using Fractal Codes," in *IEEE Proceedings of Southwest Symposium on Image Analysis and Interpretation*, pp. 104–108, 2004.

[14] M. Fernandez-Redondo and Carlos Hernandez-Espinosa, "A Comparison among Weight Initialization Methods for Multilayer Feed forward Networks," *IEEE-INNS-ENNS International Joint Conference on Neural Networks* (IJCNN'00), vol. 4, pp. 4543, 4, 2000.