IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Two stage classifier for Arabic Handwritten Character Recognition

Omer Balola Ali[1], Adnan Shaout[2], Mohammed Elhafiz[3]

Sudan University for Sciences and Technology[1, 3]

The University Michigan – Dearborn[2]

Abstract: In this paper we will present a two phase method for isolated Arabic handwritten character recognition system. The new method combines two levels based on two classifiers, a public and a private according to the similar features among characters. In the first level, we built a public classifier to deal with all character groups, each group contains characters with overlapped feature. The public classifier classifies the characters in the SUST-ARG dataset (Sudan University for Sciences and Technology Arabic Recognition Group) to specified groups. In the second level, we created a private classifier for each group to recognize and classify the characters within a group. The system was applied to 34 Arabic characters and achieved 78.79% recognition rate for the tested dataset within the first level of the grouping model and achieved 93% recognition rate for the tested dataset using the two level models.

Keywords: Isolated Handwritten Arabic character recognition, Back Propagation, feature extraction, classifiers combination, Artificial Neural Network.

I. INTRODUCTION

Handwritten Arabic Character Recognition system (HACR) is the system that attempts to recognize a text that has been written by a person (not a machine). Character recognition systems can contribute tremendously to the advancement of the automation process and can improve the interaction between man and machine in many applications, including office automation, check verification and a large variety of banking, business and data entry applications. Character recognition in Arabic has many researchers who are interested in this field [1].

The goal of a character recognition system is to transform a character handwritten on paper into a digital format that can be manipulated by characters processor software.

Neural Network approach is an emerging technique in the area of handwritten character recognition through the use of Artificial Neural Network (ANN) implementations were ANN employs specific learning rules to update the links (weights) among their nodes. Such networks can be fed with data from an input picture and trained to output characters in one or more forms [2].

Several classification techniques with many stages have been applied using neural network and others classifiers. Elanwar, Rashwan and Mashali [3] have proposed an offline character recognition system for isolated Arabic alphabet written by a single writer. They proposed multiple classifier system for handwritten Arabic alphabet recognition which has achieved an increase of about 27% in the recognition accuracy compared to a single classifier system. They used a five stage classification system to end up with an average recognition accuracy of 97% of isolated Arabic handwritten alphabet and a maximum accuracy of 98.6% with an increase of about 27% from the recognition accuracy achieved by a single classifier system. But these results were achieved for a single writer data base only. Alijla and Kwaik[4] introduced an online isolated Arabic handwritten character recognition system. Feed forward back propagation neural networks were used in the classification process. Features extraction selected were Density, Aspect Ratio and Character Alignment Ratio Clustering. Characters were organized into four groups that lead to the design of a system with four Neural Networks (NN) with small number of features in each to decrease the system complexity and increase the accuracy. Al-Jawfi [5] proposed a handwriting Arabic character recognition method using Le Net NN. He designed a neural network with two main stages to recognize character shape and have used pixel matrix of 16×16 as feature inputs in the first stage. In the second stage, he used back propagation algorithm to recognize the number of dots, position, and where it is a dot or zigzag.

Alaei, Nagabhushan and Umapada [6] described a technique for the recognition of Persian handwritten isolated characters with two-stage SVM based scheme. In the first stage, they categorized similar shaped characters into eight groups and obtained a recognition results. In the second stage, they selected groups containing more than one similar shape characters and were considered further for the final recognition. They used feature techniques that are based on under sampled bitmap techniques and modified chain-code direction frequencies. They computed 49 dimension features based on under sampled bitmaps and 196 dimension chain-code direction frequencies. The system has achieved an accuracy rate of 96.68%. Shaout and Balola [7] survey stated that" There are many similarities between Arabic characters in terms of structural and morphology". There are many number and position of dots that differentiate among the otherwise similar characters like (ج ,ح and خ) and (ب, ن, ت). This type of similarity among the Arabic letters that makes the Arabic character recognition difficult.

In this paper, we are proposing a two phase approach for

DOI 10.17148/IJARCCE.2015.412154

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

manipulate hand-written Arabic characters. In the first phase, different dataset size groups are used as general classification for all similar characters. The second phase deals with each group at a time. For both modeling levels (general and singular) we have adopted the Back-Propagation Neural Network learning algorithm. The paper will present and discuss results of the approach for the recognition rate of the classifiers used in this research. The paper is organized as follows: Section 2 describes our dataset, section 3 describes the approach architecture, experiments and results, section 4 will introduce a new isolated Arabic handwritten character recognition system and the conclusion is presented in section 5.

II. DATASET

SUST-ARG dataset which stands for Sudan University for Sciences and Technology Arabic Recognition Group (SUST-ARG) is used in this research paper. The dataset includes one hundred and forty one forms filled by different subjects. The form as shown in fig. 1 has been designed to collect the required handwritten letters. These forms are scanned by a scanner with accuracy of 300 dpi, saved as color images. Fig. 2 shows the letter extraction process. This process extracts each specific letter from all forms and put it in a separate folder as a gray scale image. All letters are composed from 1410 images written by hand.

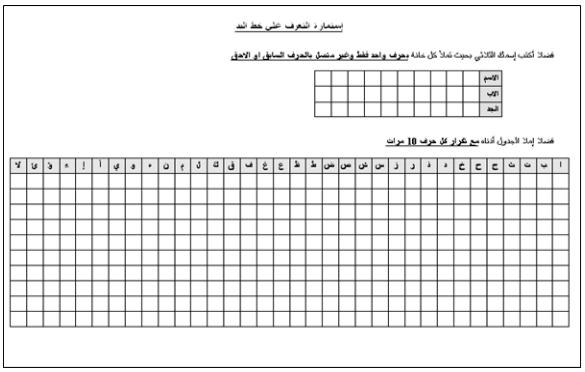

Fig. 1.Illustrates alphabet Arabic letters document form.

III. THE RECOGNITION PROCESS

Three main processes are included, preprocessing, feature extraction, and classification in our proposed system for multi-classifier system classification of handwritten isolated Arabic character as shown in fig.3.

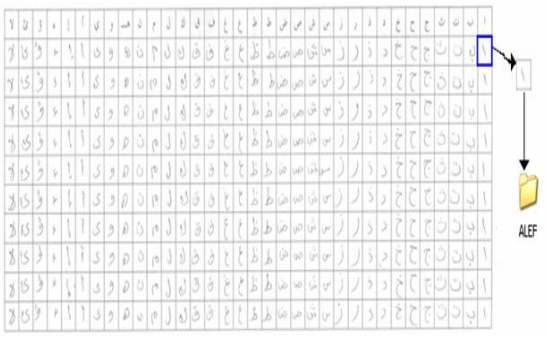

Fig.2.Illustrates letters extraction.

The preprocessing is utilized to remove noise then features are extracted from each character. These features are then used for classification to identify the character, using back-propagation algorithm.

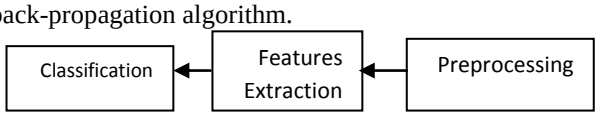

Fig.3.Illustrates classification system steps

III.1 Preprocessing

A handwriting character was sampled on A4 size paper. The characters were scanned using a scanner with a resolution of 300dpi. These characters then will be segregated according to their own character group and stored as grey scale images. We used the following preprocessing techniques: spatial filters, resized, noise removal and median filter. Fig.4 shows a sample of the handwritten characters ba and kha.

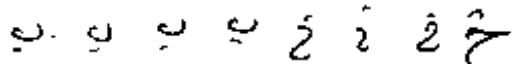

Fig.4.A sample of characters (ba) and (kha) scanned and collected together.

III.2 Features Extraction

This process extracts the features of the characters that are most relevant for classifying at the recognition stage. This is an important stage as it can help avoid misclassification, thus increasing recognition rate. Principal Components Analysis (PCA) is a very popular technique for dimensionality reduction. Given a set of data in n dimensions, PCA aims to find a linear subspace of dimension d lower than n such that the data points lie mainly on this linear subspace. Such a reduced subspace attempts to maintain most of the variability of the data [8]. Abandah, Youn is and Khedher [9] proposed five Arabic handwritten character recognition classifier system were they applied the PCA. They used 95 feature vectors, secondary components features, main body features, skeleton features, and boundary feature. In this paper we are using PCA as a feature extraction technique.

III.3 Classification

The classification process is carried out at the final stage to recognize the characters. It assigns an input character to one of many pre-specified classes which are based on the extracted features.

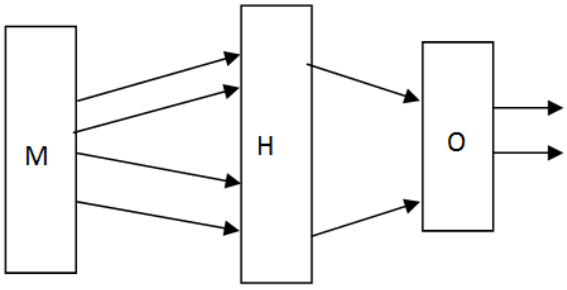

Fig.5. Network topology with M= 50, 100 or 350, H=10 – 150 and O= 34.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

For the classification process, ANN is used in this paper. Back propagation (BP) neural network is used for training and classification of Arabic handwritten characters.
The proposed network topology is shown in fig. 5many experiments have been conducted to find the best values for the number input units (M) and hidden layer units (H).

IV. EXPERIMENTS AND RESULT

A set of Arabic handwritten characters was selected concentrating on the basic Arabic letters, which are 28 characters and the related characters such as أ، إ ئ ء ،لا، ؤ. The dataset consists of 30600 images, divided into 23800 images as training dataset and 6800 image as testing dataset. A set of different experiments were conducted on the training dataset and the testing dataset as shown in table 1. The highest recognition rate for the testing dataset is 54.3%. This rate occurred with 350 features as input and 100 units for the hidden layer.

Table1: Different Hidden Layer nodes for 34 classes refer to the Recognition Rate.

Feature	Hidden nodes	Time Minutes	Accuracy rate %	
			Training dataset	Testing dataset
50	40	15,23	30	30.11
	70	24,46	41.89	37
	100	54,49	68.85	51.36
100	70	27,52	44.09	36.06
	100	59,27	77	51.33
350	70	1,11,38	67.03	41.33
	100	**1,32,9**	**83.87**	**54.30**

When the confusion matrix was created, the classifier read some letters as other letters, for example the letter"ت" with200 samples was correctly classified for53 samples and incorrect for the rest as shown in table 3. The more overlapped the characters are such as "ن" and "ث" as shown in table 2 the more the characters are misclassified.

Table 2: Illustrates the "ت" Ta letter confusion matrix

	ا	ب	ت	ث	ج	ح	خ
ا	**155**	0	0	2	1	0	0
ب	0	**126**	2	2	3	1	1
ت	0	2	**53**	48	6	2	1
ث	0	0	48	**56**	2	1	5
ج	1	1	1	3	**88**	18	11
ح	0	3	2	1	28	**80**	7
خ	0	0	0	2	5	9	**92**
د	1	1	0	0	5	0	0
ذ	1	1	1	0	1	0	3

Table 3: confusion matrix for 200 samples of the(ta) letter

character	ت	ن	ث
ت	53	19	48

V. NEW PROPOSED METHOD TO IMPROVE CHARACTER RECOGNITION

Two stages of classifier are proposed as shown in Fig. 7it is a multistage classifier that has two stages. The first stage is based on features extracted from groups of similar characters. The last stage has sub-classifiers. A sub-classifier is a multiple neural network (BPNN) classifier that is used to recognize only one group characters.

V.1 General Level

Some Arabic letters are similar in their shapes. When written by hand some letters are similar in dots and rings as shown in fig.6the system accepts the features of the letters that we need to recognize and classify to the correct group. At this level we have conducted many experiments that have used two datasets, which are different in number of classes. The datasets are as follows:

Fig.6. Example of similarly characters (Ta, Tha, Noon and Hamza on ya).

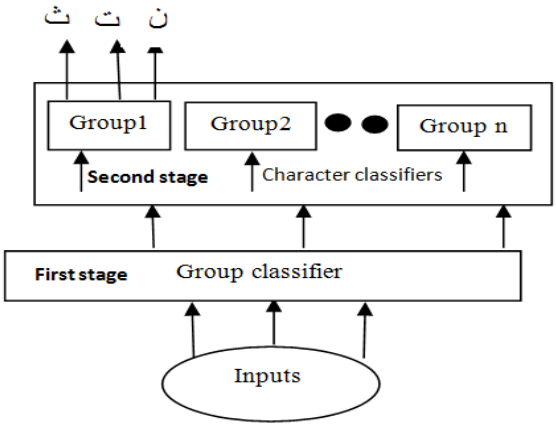

Fig7. Two stages classification system.

A. dataset with 11 classes

In this dataset we havegrouped the letters that have similar features in eleven groups as shown in table 4. Many experiments have been conducted as shown in table 5. The highest recognition rate is **71.73%** for the testing dataset.

Table 4: the similar characters groups.

Groups number	Eleven Character Groups	Fifteen Character Groups
1	ا إ أ	ا إ أ
2	ب ت ث ن ش	ب ت ث ن
3	ج ح خ م	ج ح خ
4	د ذ ر ز ء	د ذ ر ز ء
5	س ص ض	س ص ض
6	ط ظ	ش
7	ع غ	ط ظ
8	ف ق ؤ	ع غ
9	لا ل ك	ف ق ؤ
10	و هـ	ل ك
11	ئ ي	م
12		هـ
13		و
14		ئ ي
15		لا

DOI 10.17148/IJARCCE.2015.412154

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Table 5**:** Recognition accuracy testing dataset for 11 groups

Feature	Hidden	Time Minutes	Train accuracy %	Test accuracy %
50	70	14,3	87.83	61.23
	100	18,11	91.56	63.5
	200	35,26	97.02	67
100	100	22,29	94.76	62.37
	200	40,38	98.08	68.28
	600	**2,13,1**	**89.77**	**71.73**
350	70	28,34	93.52	59.73
	100	39,39	95.10	62.55
	200	**1,12,37**	**94.93**	**64.64**

B. Dataset with 15 classes

In the confusion matrix shown in table 6, the experiment which is has the highest recognition rate for the testing dataset is shown in table 5. Some letter groups have low classification rate. For example, the letters of the second group was classified 17 times as eighth group and 17 times was read as ninth group. The letters that leads to the similarity between the two groups was the letter "ش", so it was set in separate class. The algorithm will work with the rest of similar letter in the same way. The new grouped dataset are trained again with 15 classes that contain letter groups in separate classes. Many experiments were conducted with various number of feature sets as shown in table 7. The highest recognition rate is **78.77%** for the testing dataset for the fifteen groups of characters as shown in table 7. The resulted plains the experiment performance with the highest recognition rate.

The output resulted from this stage is not final output, but it is fed as input pattern to the final stage.

V.2 Singular Level

After the character has been classified into its group using the public classifier, then the character is inputted to the classifier of the individual group. The result of this singular level is a match with one of the particular group's element. The experiments are applied to one group of the 15 dataset classes, the letter group that contains the letters "إ، أ ، ا" as shown in table 8. The highest recognition rate is 92.77% to the testing dataset.

Table 6: confusion matrix on testing dataset for 11 groups.

accuracy	6	5	4	3	2	1	GROUP
86	7	0	3	0	3	172	1
40.5	5	20	13	9	81	0	2
49.5	7	12	.5	99	1	7	3
74.5	9	0	149	2	2	4	4
87	3	174	0	5	1	0	5
66	132	6	3	1	2	1	6
74.5	5	0	2	30	1	5	7
66.5	11	12	5	2	14	0	8
64.5	8	1	1	7	24	1	9
78.5	16	5	9	3	0	0	10
63.5	4	12	2	5	16	2	11

Followed Table 6: confusion matrix on testing dataset for 11 groups

Table 7: Testing the dataset for the 15 groups of characters

Accuracy rate	11	10	9	8	7	GROUP
86	1	0	3	2	9	1
40.5	26	3	17	17	9	2
49.5	7	11	11	7	33	3
74.5	3	17	3	9	2	4
87	3	8	1	5	0	5
66	6	17	10	22	0	6
74.5	3	2	1	2	149	7
66.5	4	9	5	133	5	8
64.5	7	14	129	7	1	9
78.5	2	157	4	3	1	10
63.5	127	5	10	17	0	11

Table 8**:** for the letter group (1) that contains the letters

Feature	Hidden	Time Minutes	Train accuracy %	Test accuracy %
100	200	11,41	99.84	91.50
	600	**32,10**	**99.76**	**92.77**
350	**200**	**21,30**	**99.94**	**90.5**
	300	28,40	98.70	87.67

Feature	Hidden	Time Minutes	Recognition rate %	
			Training dataset	Testing dataset
100	70	23,10	88.31	61.60
	100	31,3	92.51	65
	200	**59,5**	**97.75**	**71.04**
	300	1,23,21	98.18	74.27
	400	1,55,23	98.63	76.64
	500	**2,38,21**	**99**	**77.10**
	600	**2,49,39**	**99.04**	**78.77**
	700	3,11,17	92.47	74
	600	3,0,10	98.63	77.37

VI. DISCUSSION

The new proposed method for Arabic recognition presented in this paper has divided the problem into multiple sub-problems. The method is made of two stage classification system. The two-stage classifier separates the classification problem into different modules. To improve the performance of recognition result, the characters are divided into multi groups using some knowledge about similarities among the characters. The design of the classifier is made based on both the similarities among character structures and a multistage classification. The design of the handwritten Arabic character classifiers is a multistage classifier that has two stages. The first stage is based on features extracted from each group of similar characters that were placed in

separate classes. The last stage has sub-classifiers each sub-classifier is a multiple parallel neural network (BPNN) classifier. The final decision is made by calculation the largest value of the averaged outputs from the two network stages.

Experiments have shown that the proposed method is an effective technique for improving character recognition accuracy.

VII. CONCLUSION AND FUTURE WORK

This paper has presented a system for recognizing handwritten Isolated Arabic characters. The dataset that was used in the paper experiments was from isolated Arabic character set, the SUST-ARG dataset (Sudan University for Sciences and Technology Arabic Recognition Group). It was collected from 141 writers were each person writing each letter ten times and scanned with 300 dpi. The characters, totally 30600 characters, are divided into a training (23800 characters) and testing sets (6800 characters). The division of the characters into training and testing was achieved during the first phase and was obtained from several experiments. We have designed other datasets that contains groups of similar letters to improve the recognition rate. The method that was described in this paper for Arabic handwritten character recognition can be extended for other Arabic character position by including few other preprocessing activities. The gray scales of pixels from letter images were used as inputs for the BP network. In our future research work, we would like to improve the recognition accuracy of network for handwritten Arabic character by using more training samples written by one person and by using a good feature extraction system.

Many experiments have been conducted during this research. The experiments conducted have used 34, 11, 15 and 3 character classes. Features were extracted from images using structural feature extraction algorithms. The highest result accuracy for testing dataset with 15 classes with one stage was 78.77% and for one group with two stages was 92.77%.

REFERENCES

[1] S. Impedovo, "Frontiers in Handwriting Recognition", in "Fundamentals in Handwriting Recognition", S. Impedovo (ed.), NATO-ASI Series, Springer- Verlag Publ., Berlin, 1994, G. Dimauro, S. Impedovo, G. Pirlo, A. Salzo Handwriting Recognition: State of the Art and Future Trends.

[2] "Neural Networks for Unicode Optical Character Recognition", www.projectsatbangalore.com/IEEE.pdf.

[3] R. I. M. Elanwar, M. A. A. Rashwan and S. Mashali, "A Multiple Classifiers System For Solving The Character Recognition Problem In Arabic Alphabet", Conference Paper · December 2006.

[4] Basem Alijla and Kathrein Kwaik, "OIAHCR: Online Isolated Arabic Handwritten Character Recognition Using Neural Network", The International Arab Journal of Information Technology, Vol. 9, No. 4, July 2012.

[5] Al-Jawfi R., "Off Handwriting Arabic Character Recognition LeNet Using Neural Network," The International Arab Journal of Information Technology, vol. 6, no. 3, pp. 304-309, 2009.

[6] A. Alaei, P. Nagabhushan and Umapada Pal, "A New Two-stage Scheme for the Recognition of Persian Handwritten Characters", 12th International Conference on Frontiers in Handwriting Recognition 2010,978-0-7695-4221-8/10 $26.00 © 2010 IEEE DOI 10.1109/ICFHR.2010.27

[7] Adnan shaout and Omar balola, "Isolated Arabic Handwritten Character Recognition: Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 10, October 2014 ISSN: 2277 128X.

[8] S.F. Bahgat, S.Ghomiemy, S. Aljahdali and M. Alotaibi, "A Proposed Hybrid Technique for Recognizing Arabic Characters", (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 1, No. 4, 2012.

[9] G. A. Abandah, K. S. Younis and M. Z. Khedher, "Handwritten Arabic Character Recognition Using Multiple Classifiers Based On Letter Form", In Proc. 5th IASTED Int'l Conf. on Signal Processing, Pattern Recognition, & Applications (SPPRA 2008), Feb 13-15, Innsbruck, Austria.