IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 05, Issue 04 (April. 2015), ||V2|| PP 36-41

A Divide And Conquer Method For Arabic Character Recognition

Nehad H A Hammad[1], Mohammed Elhafiz[2]
[1](*Palestine Technical College, P.O.Box6037, Gaza, Palestine,*
[2](*Sudan University for Science and Technology, P.O.Box382,Mokren , Khartoum ,Sudan*

Abstract: - Optical Characters Recognition (OCR) has been an active research area since the early days of computer science. Despite the age of the subject, it remains one of the most challenging and exciting areas of research in computer science. In recent years it has grown into a mature discipline, producing a huge body of work. Arabic character recognition has been one of the last major languages to receive attention. In this paper a new divide and conquer algorithm is proposed to recognize isolated handwritten Arabic characters using character's curve tracing and number of dots information. A robust connected component labeling algorithm were used to divide the characters into four groups. To overcome the problem of size and scaling the tangent direction is used as feature (Divide-and-conquer). Four neural networks (NN) were used to classify each groups. The primary result is promising and outperform one neural network classifier.

Keywords: - *Arabic Character Recognition, Resizing, Feature Extraction, Neural Network.*

I. INTRODUCTION

Handwriting recognition refers to the identification of written characters. The problem can be viewed as a classification problem where we need to identify the most similarity character to the input character. The challenge of the recognition system increases with the increase of number of characters.
Actually ,the language contains more number of characters shape, the identification would be more difficult than the other language contains lesser number of characters.
Similarly we need looking to how the various characters are written and the differences between them.

The Arabic alphabet contains basically 28 letters are written from write to left. The Arabic isolated character recognition is more complex because the reason is the similarities among the different letters and the differences among the same latter. For example, letters Baa(ب), Taa(ت), and Thaa(ث) are three different letters, but they have similar body shape, they only differ in number and position of dots (one, two or three dots under or above the body of the character). Also Jeem(ج), Hhaa(ح) and Khaa(خ) they differ only in one dot, and same situation happens with the most of the rest[1].
In proposed method reduce similarity between each character by using number of character to four object(s).
The proposed method using "Divide and conquer" theory to divide Arabic character in four group depend on number of objects in characters.
The connected component labelling using to determine the character member of which group,works by scanning an image, pixel-by-pixel (from top to bottom and left to right) in order to identify connected pixel regions after that the dots using to make sub-group if is it available.
components and dot(s) level and number. The Arabic Isolated character contain one

II. RELATED WORK

Haraty and Ghaddar (2003) propose the use of two neural networks to classify previously segmented characters. Their method uses a skeleton representation and structural and quantitative features such as the number and density of black pixels and the numbers of endpoints, loops, corner points, and branch points. On 2,132 characters, the recognition rate is over 73%[2].

Amin (2003) presents an automatic technique to learn rules for isolated characters. Structural features including open curves in several directions are detected from the Freeman code representation of the skeleton of each character and the relationships are determined with Inductive Logic Programming (ILP). Test data consist of 40 samples of 120 different characters by different writers with 30 character samples used for training and 10 for testing for most experiments. A character recognition rate of 86.65% is obtained[3].

Alaei et al. (2010) proposed a two-stage approach for isolated handwritten Persian character recognition. They extracted features based on modified chain code directional frequencies and employed an SVM for classification. They obtained 98.1% and 96.6% recognition accuracy with 8-class and 32-class problems, respectively[4] .

Desai (2010) presented a technique for Gujarati handwritten numeral recognition. the author used features abstracted from four different profiles of digits with a multilayered feed forward neural network, and achieved an approximate 82% recognition accuracy for Gujarati handwritten digit identification[5] .

Sharma and Jhajj (2010) extracted zoning features for handwritten Gurmukhi character recognition. They employed two classifiers, namely k-NN and SVM. They achieved a maximum recognition accuracy of about 72.5% and 72.0% with k-NN and SVM, respectively [6].

Kumar et al. (2011) extracted intersection and open end points features for offline handwritten Gurmukhi character recognition. They used SVM for classification, taking 90% of the dataset as a training set and 10% of the dataset as a testing set. They achieved a maximum recognition accuracy of about 94.3%[7].

III. PREPROCESSING

Character recognition system perform steps before recognition. The first is image acquisition where the images are being inputted to the computer using an optical device usually this input image colored or grey scale image. Therefore most of recognition system convert this image to binary (0-black,1-white) this step call binarization [8] .Other important steps are normalization, noise removal ,thinning[9].

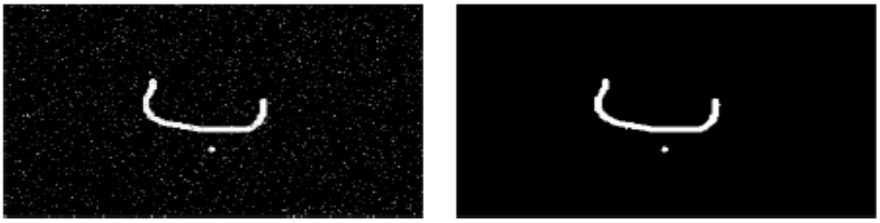


Fig. 1. Noise Removing using Median Filter.

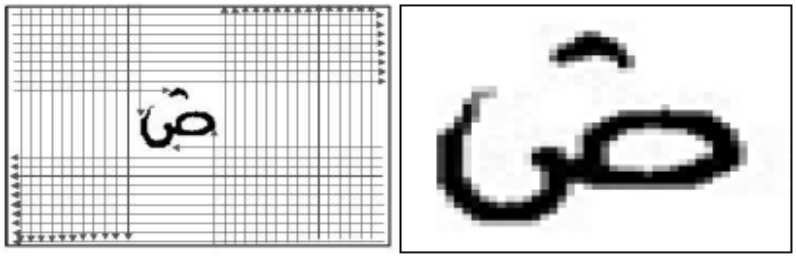


Fig. 2. Binary Image Normalization(Cropping).

Fig. 3. Arabic Isolated Character Thinning.

IV. THE PROPOSED ALGORITHM

The proposed Arabic isolated characters recognition method divides all Arabic characters into four groups. This group based on the number of connected component in the character.

$$\text{Angle}(Q) = \arctan(p_n(y) - p_{n+1}(y) / p_n(x) - p_{n+1}(x)) * 180 / PI \qquad (1)$$

The method feature vector contains 17 values 15 represent characters contours and 2 represent dot(s) count and position level .

After thinning ,15 points are taken from character trace, the angles between two adjacent points are is calculated using Equation (1).Figure 4 illustrates the directions which has been chosen for experiments of the proposed method.

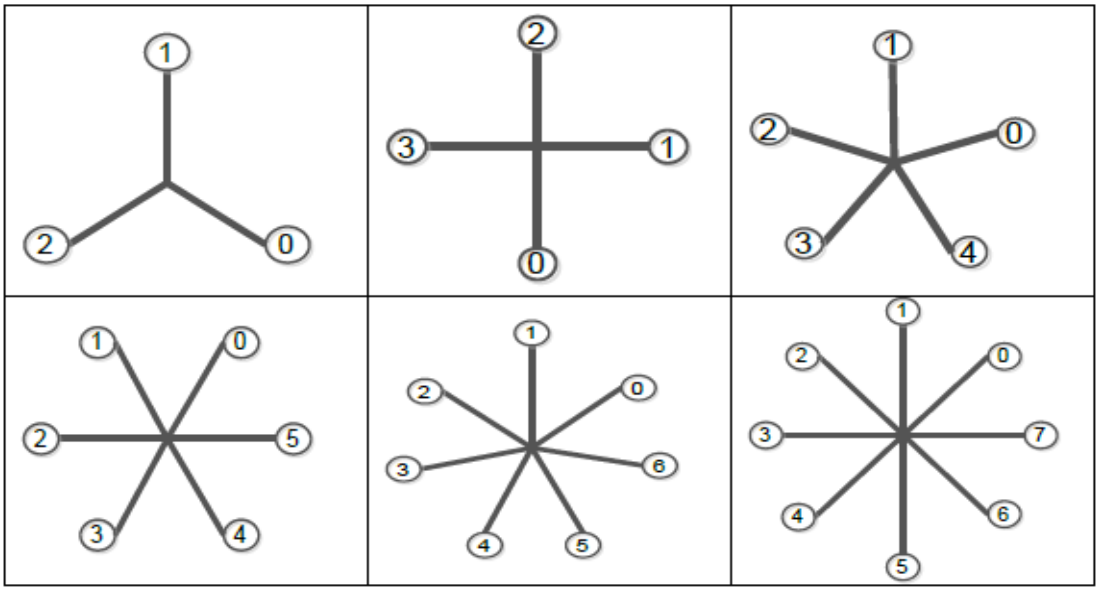


Fig. 4. The Selected Directions to be used in Proposed Methods

For Arabic character dot(s) appear in 3 different positions upper level , middle level and lower level represented in one value from 1 to 3 , one for upper , two for middle and three for lower.

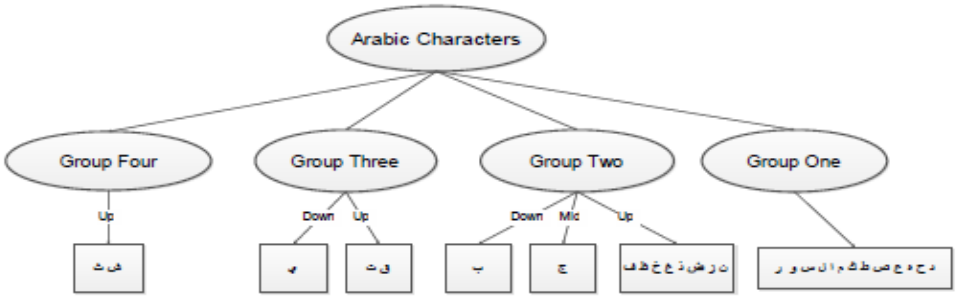


Fig. 5. Arabic Character belongs Groups

Therefore we have the following categories:

- Category 1: Characters have one Dot in the upper(11) like (Gain غ, Fehف, khahخ , Noon ن , Zen ز, Theh ذ , Theh ظ,Zah ض),two Dots in the upper(12)like (Tehت), three in the upper(13) like (Thehث , Sheen ش)
- Category 2 :characters have just one Dot in the middle(21) like (Jeemج).
- Category 3: characters have one Dot in the bottom(31) like (Beh ب,),two Dots in the bottom(32) like (Yeh ي).

Table1. Illustrates these Categories.

Number of Dots	One Dot	Two Dots	Three Dots
No Dots	[00]	[00]	[00]
Level 1(Up)	[11]	[12]	[13]
Level 2(Mid)	[21]	-	-
Level 3(Bottom)	[31]	[32]	-

Dot Detection process:

1. Calculate the number isolated objects in the image .
2. If number of objects = 1 then pattern be belong to group one.
3. If number of objects >= 2 then
4. The Longest object size is character body and other is dot(s).
5. The character dots recognized using other process to calculate number of dots and position level.
6. The number of dots depend on dot length using thresholds and dots shape.
7. The dots level extracted when remove all space around character after divide the character height into 3 area on Y axis.

Fig. 6. Character Dots Levels

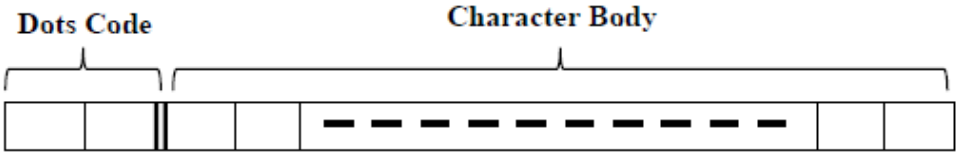


Fig. 7. Feature Vector Structure Design

Fig. 8. Arabic Character Representation

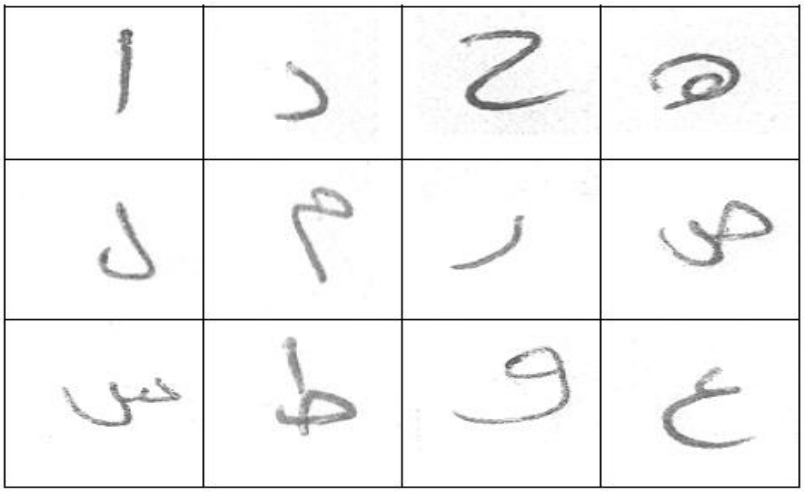
V. DATASET

The off-line Arabic isolated character Dataset is Sudan University For Science and Technology Arabic Recognition Group (SUST ARG) used in proposed methods.

VI. EXPERIMENTAL AND RESULT

Group one contains all isolated Arabic characters with one part, so the character body written using connected line, This group contains 12 characters (ا,د,ح,ع,هـ,ص,ط,م,ل,س,و,ر) .See Table 2 and Fig 9.

Table 2. Group One Isolated Arabic Characters Handwritten.



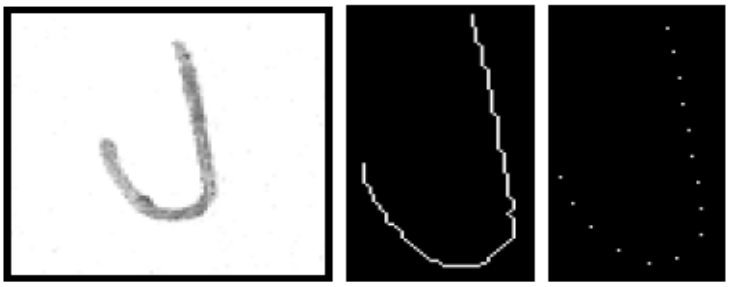


Fig. 9. Arabic Isolated Thinned Character.

As this is biggest category ,neural network has been designed to classify the member of this group.6 experiments has been performed to test the six chosen directions . Table 3 shows the result of these six experiments. These results show that eight direction outperform the other directions. That mean experiments on more direction should be performed[10].

Table 3. Cross Validation for Group One Recognition.

Direction No	Three	Four	Five	Six	Seven	Eight
1	75.3	86.4	85.5	81.2	81.3	88.6
2	85.2	89.3	85.2	89.2	84.1	89.2
3	82.6	81.7	79.0	87.9	84.1	90.1
4	78.1	86.5	85.4	85.6	83.6	88.9
5	82.7	85.6	86.2	84.2	84.0	89.3
6	81.8	86.0	85.2	84.7	80.3	88.9
7	78.8	84.5	84.4	86.4	85.6	89.2
8	79.7	86.0	84.4	82.5	83.7	85.6
9	84.7	83.1	82.8	82.1	85.8	83.0
10	84.2	88.4	84.4	85.1	79.7	88.3
Average	81.31	85.75	84.25	84.89	83.22	88.11
Standard Deviation	2.668	1.62	1.34	1.95	1.672	1.524

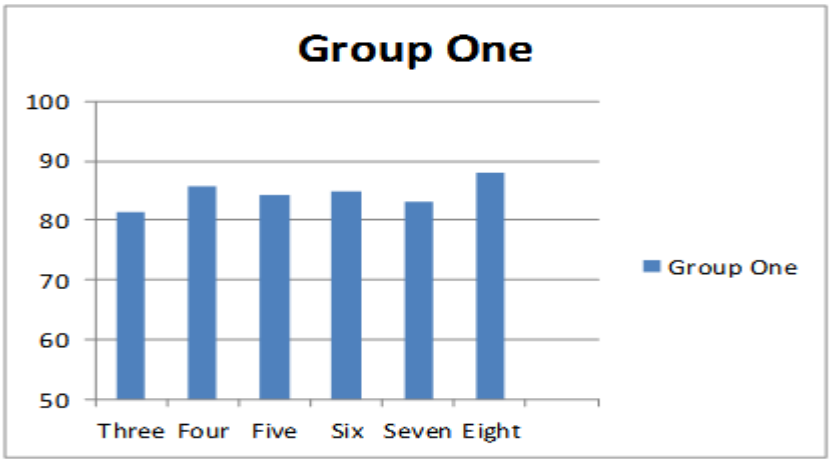


Fig. 10. Cross Validation for Group OneScheme

VII. CONCLUSION AND FURTHER WORK

Arabic character recognition methods are affected by several factors like characters shape, complexity of drawing the character and the amount of rotation of the character on the horizontal line . To overcame these problems we propose divide and conquer system the propose method divide the Arabic characters into 4 groups .A neural network has been designed to classify each groups.

The result of this experiment can give however these result show that more directions should be tested.

Finally we look in this study to apply the proposed method to recognize other languages handwritten characters and the development method to be more performed and efficient in online and offline handwritten characters.

REFERENCES

[1] Abdurazzag Ali ABURAS and Salem M. A. REHIEL ,(2007),"Off-line Omni-style Handwriting Arabic Character Recognition System Based on Wavelet Compression", International Islamic University Malaysia, Electrical and Computer Engineering, PP10, 50728, 53100.

[2] Haraty, R. and Ghaddar, C. Neuro-Classification for Handwritten Arabic Text. Proceedings ACS/IEEE International Conference on Computer Systems and Applications,2003.

[3] Amin, A. Recognition of Hand-Printed Characters Based on Structural Description and Inductive Logic Programming. Pattern Recognition Letters, vol. 24, pp. 3187-3196, 2003.

[4] A. Alaei, P. Nagabhushan, U. Pal, "A new two-stage scheme for the recognition of Persian handwritten characters," in Proc. of 12th ICFHR, pp.130-135, 2010.

[5] A. A. Desai, "Gujarati handwritten numeral optical character reorganization through neural network," Pattern Recognition, vol. 43, no. 7, pp. 2582-2589, July 2010.

[6] D. V. Sharma, P. Jhajj, "Recognition of isolated handwritten characters in Gurmukhi script," International Journal of Computer Applications, vol. 4, no. 8, pp. 9-17, 2010.

[7] M. Kumar, M. K. Jindal, R. K. Sharma, "k-NN based offline handwritten Gurmukhi character recognition," in Proc. of ICIIP, pp. 1-4, 2011.

[8] JiaYonghong, "Digital image processing(The Second Edition)".WuHan China: Wu Han university press, pp 114-116,2010.

[9] Lam, Seong-Whan Lee, and Ching Y. Suen, "Thinning Methodologies-A Comprehensive Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 14, No. 9, 1992.

[10] Kohavi, R.," A Study of Cross Validation and Bootstrap for Accuracy Estimation and model Selection", Proceedings Of the 15th International Conference on Artificial Intelligence (IJCAI). pp. 1137-1143, 1995.