Indonesian Journal of Electrical Engineering and Computer Science
Vol. 14, No. 3, June 2019, pp. 1518~1524
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v13.i3.pp1518-1524

Recognize printed Arabic letter using new geometrical features

Haidar J. Mohamad[1], Seham A. Hashim[2], Anwar H. Al-Saleh[3]
[1]Mustansiriyah University, College of Science, Department of Physics, Iraq
[2]Middle Technical University, Technical Instructors Training Institute, Technical Electronic Department, Iraq
[3]Mustansiriyah University, College of Science, Department of Computer, Iraq

Article Info

Article history:

Received Dec 10, 2018
Revised Feb 11, 2019
Accepted Feb 27, 2019

Keywords:

Arabic letter
Evaluation of classification
Feature extraction
Image categorization

ABSTRACT

The task of recognizing the shape of Arabic letters using modified algorithms discussed in this paper. The difficulty of recognizing these letters is summarized in the shape of the Arabic letter within a word from a large set of letters has a similar shape. Moreover, the shape of the letter is different depending on its position begin, middle, end within a word. Therefore, it is necessary to introduce new geometric features to categorize each letter. The suggested algorithm with 19 features is used in this paper. These features, like define points for each letter, divide a letter to blocks, edge detection and other features are shown in the suggested algorithm. The introduced geometric features give a high accuracy to recognize printed Arabic letter within a word or text. Minimum distance criteria used to estimate the error of the recognition process between the database and the tested Arabic letter. This method is good to explain the behaviour of the designed algorithm code to distinguish the geometric properties and the accuracy reaches 99.8% for the proposed method. The letter size changes geometry details when the font size is changed. The studied font is Times New Roman with size 30, 36, and 39.

Corresponding Author:

Haidar J. Mohamad,
Mustansiriyah University,
College of Science, Department of Physics, Iraq.
Email: haidar.mohamad@uomustansiriyah.edu.iq

1. INTRODUCTION

Document image analysis (DIA) is one of the most wanted fields nowadays because it deals with converting text to an image that applied to different applications. Therefore, a recognition step, should be considered to analyses a letter component within the document image to get information as a human eye does. Image analysis considers document image analysis as a special case because of spatial properties of the text within the whole image which is different from a satellite image. The demand for using DIA increase every day in document system like wide use of Optical Character Recognition (OCR) in daily routine. For instance, documents in a library converted and archived to an electronic version to be used online and easy to search a word or subject for everyone.

The idea of capturing data from a book is by using an optical scanner which stores data in a form of the picture file. This picture is converted in image processing to a matrix of 1 and 0 value (in case of a binary image) to study the important information. In the case of study letters, the information is to recognize the shape and identify the letter. Therefore, it is important to do some pre-processing stage before performing the recognition process. These steps can be understood as edge detection, thinning, binarization and so on, depending on the suggested algorithm or what the goal of using the image processing technique.

Pattern recognition is a basic knowledge of creating OCR, where OCR interprets characters in an alphabetic shape to another form. These alphabetic characters are rich in information and shape, i.e. font type, size, and Arabic letter position. The principle of OCR based on developing algorithms to recognize

alphabetic shape. Therefore, it is important to reach human recognition accuracy with 100% for the alphabetic shape to be reached for OCR.

Many methodological analysis and scheme have been proposed like using five stages preprocessing like binarization, median filter, Hilditch thinning method, and line segmentation with 24 features. The accuracy of the performance system, unfortunately, reached 48.3% [1]. An algorithm used to recognize car plate letter and numbers with accuracy 95%. But this method works only with car plate number and does not fit with other environments[2]. A chain code algorithm used to compute letter feature depending on neighbourhood function with a normalization. A thinning preprocessing method used with this paper to get high accuracy[3]. Four groups to extract letter features presented in terms of main body, boundary, skeleton, and secondary object feature. This process is complicated to implement all the steps, in the same way, every time [4]. An approach based on structural features and decision tree learning techniques presented which consist of three parts. These parts firstly store letter character in an array after user write it in a special window, secondly generate bounding 5x5 box around the letter then lastly use recognition process to apply tree learning technique. This method depends on a special window to recognize a letter and the accuracy varying between 70 – 93% depending on the tested letter [5]. A neural net recognition method presented in terms of segment letters to upper and same line level, but the accuracy varies between 90-98% depending on the segmentation method [6].

Within this paper, a new strategy presented to recognize and detect Arabic isolated letter depending on new features. These features are selected carefully, and it works with all Arabic letter. Moreover, it is only 18 feature and accurate. The algorithm code designed using MATLAB software. The steps required to have an image to the studied letter then is it automatically recognize the letter. All details are shown in next sections consist of using the algorithm and statistical results.

2. FEATURE EXTRACTION

In pattern recognition, extract a feature from any letter depends on detecting the essential characteristics of that letter. This considered one of the difficult challenges of pattern recognition. Therefore, the best way to recognize letters is to use a fixed image size of each letter and compare computational approach. This approach is not valid if the letter size expecting to vary. The next step is to find certain features that be useful to characterize any letter and ignore insignificant details. This technique of extracting such features for Arabic letter is divided into three groups like point distribution, structure, and separate/joint letter shape. These features evaluated within the introduced algorithm and updated to the need of recognition Arabic letter.

The results of recognition evaluated using minimum distance error method. This method designed to be within the second algorithm, because the results of comparing between letters and extract features are within the second algorithm.

3. ARABIC LETTER SHAPE

The Arabic letters have an imprint that different comparing with other languages letters. These characters can be described as letter shape which changed according to the letter location within a word. Therefore, the recognition process should solve this issue with all Arabic letters. Figure 1 shows the Arabic letter *Geem* alone, at the beginning, middle, and end of words.

Figure 1. The shapes of the same Arabic letter *Geem* according to its position in a word

The other characters are the similarity in shape between Arabic letters. The only difference is a dot over or under the Arabic letter, and this dot changes the pronunciation of the letter and meaning of the word. The native people who speak this language can distinguish between these letters. However, in the recognition process, this issue should be considered. Figure 2 shows samples of these letters which are similar in shape with a small difference.

ز ر	د ذ	ظ ط	ف ق	س ش
ض ص	ث ب ن ي ت	ع غ	ح ج خ	

Figure 2. Arabic letters with similar shape but different in a dot

4. ARABIC CHARACTER RECOGNITION

The suggested Arabic letters are tested within this work are listed in the Table 1, where the number sequence is SN and this number is shown in the results where it indicates the letter, shape letter is the printed shape in the Arabic language, pronunciation is how this letter can be pronounced in English. These letters have similar properties, these properties are without dots and its position is isolated. There are many cases where the same letter is formed in the word, but we choose this form to be checked and test.

Table 1. The Studied Arabic Letter with Its Sequence and Pronunciation

SN	Shape letter	Pronunciation
1	ع	Ain
2	ا	Alef
3	د	Dal
4	ح	HHA
5	ه	Ha’
6	ل	Lam
7	م	Meem
8	ر	Raa
9	ص	Saad
10	س	Seen
11	ط	Taa
12	و	Waw

The suggested algorithms are designed carefully to be fit with the purpose of recognizing shape letter. First, a database designed for all Arabic letter with its properties to be compared with the results of the inputs letters. This database consists of letters features listed as a table within the algorithm data and indexed with the feature sequence and feature number. The second algorithm test features characters of the input image letter and compares it with the database and feedback the results. The error and the matching between the database and the input letter image checked using minimum distance (MD) criteria within the second algorithm. The introduced diagram steps to recognize the letter shown in Figure 3.

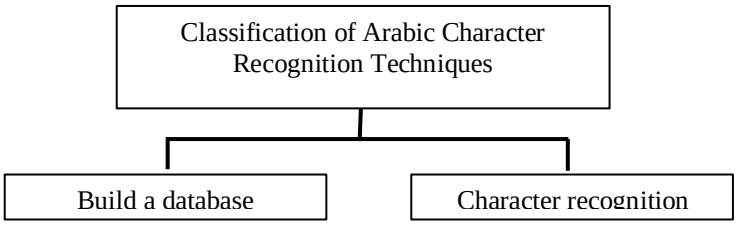

Figure 3. Diagram steps of character recognition

5. EXTRACTING ARABIC LETTER FEATURES

The first step of recognition any letter creates a database consist of 19 features which explained in detail in the Table 2. First, convert the input image to a binary image i.e. zero and one, then remove the black holes inside the letter. Change the input image size to (100x60) pixel for all input images to be uniform. In this case, the letter shape will fill the image to the border to be fit without spaces. Edge detection process using the Soble operator is applied the edges of the input image used as one of the features as shown in

Figure 4 (a). The centre of the tested images is determined and divide it into four square parts and extract number of pixels with 255 and 0 (for the filled letter and edges) representing the area of this section as a new two features as shown in Figure 4 (b). Then calculate the number of pixels whose value is equal to 1 as a feature added within this algorithm. Determine the first pixel in a row and column and give it position x1, y1, then the last pixel in the same row with position x2, y2, then the first pixel in the last row and first column to be x3, y3 position, finally the last pixel in the same row and last column to record x4, y4. Then join these positions with straight lines and calculate the length of each line as new feature namely L1, L2, L3, and L4 as shown in Figure 4 (c). The plot and determine horizontal straight line for each continuous joined white pixels in each row the number of these lines, the longest and shortest straight line with its location T1, T2, Tpo1, and Tpo2 respectively, presents new features as shown in Figure 4 (d). These steps are shown in the algorithm (1).

The next step is recognizing the input letter using the algorithm (2). The input image of the Arabic letter checked with the first algorithm to extract all the features and compare it with database. The database contains all Arabic letters features as a coded table. Then the recognition step implemented in the second algorithm, where the minimum distance equation is used. The letter feature compared with database and the minimum distance calculate the error percentage with the database which has all Arabic letter features. The lower value of this criteria means the high matching with the database letter. The results of the MD data shown in Figure 5 for the input images letters. Figure 5 shows the three sizes used to compare algorithm outputs as a function of letter size. There is a small difference when the letter size is changed as shown in Figure 5 (d). However, the letter recognition still effective and gives a high matching. The MSD that used in the algorithm (2) is effective and simple to act and giving results. The tilt dashed line in Figure 5 shows the lowest value of MD for Arabic letter which consists of the SN sequence. The accuracy is high to make the error percentage reaches 99.8%. While other researchers reach accuracy value 88.38%[7], 94.44%[8], 95.64%[9], 96.84%[10], 97.23%[11], and 97.3%[12].

The accuracy can be calculate using (ACC=[(MD-average (MD))]/MD* 100%), and the error percentage EP=(100-ACC)*100%[13]. Therefore, for the letter HHA with font size 30, the MD is 0.02657 and average (MD) equal 0.02199, as a result, the EP is 99.8%.

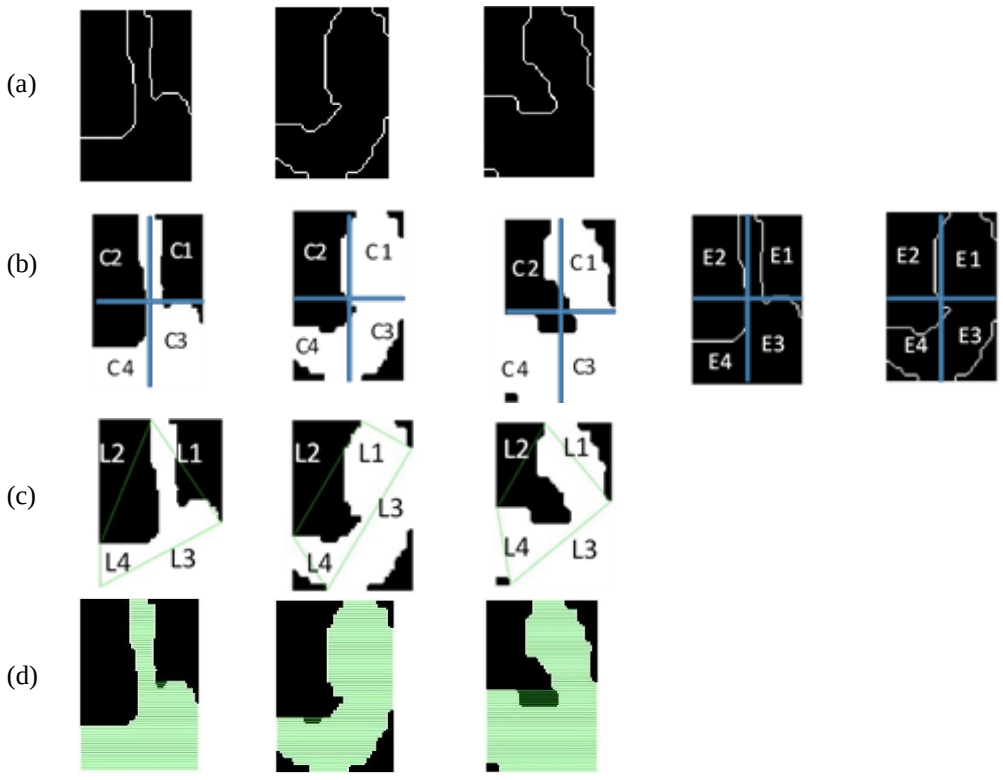

Figure 4. The output images of the algorithm (1) for the new features (a) edge detection (b) dividing filled letter and edge image into four parts (c) length of the joined pixel (d) horizontal straight line

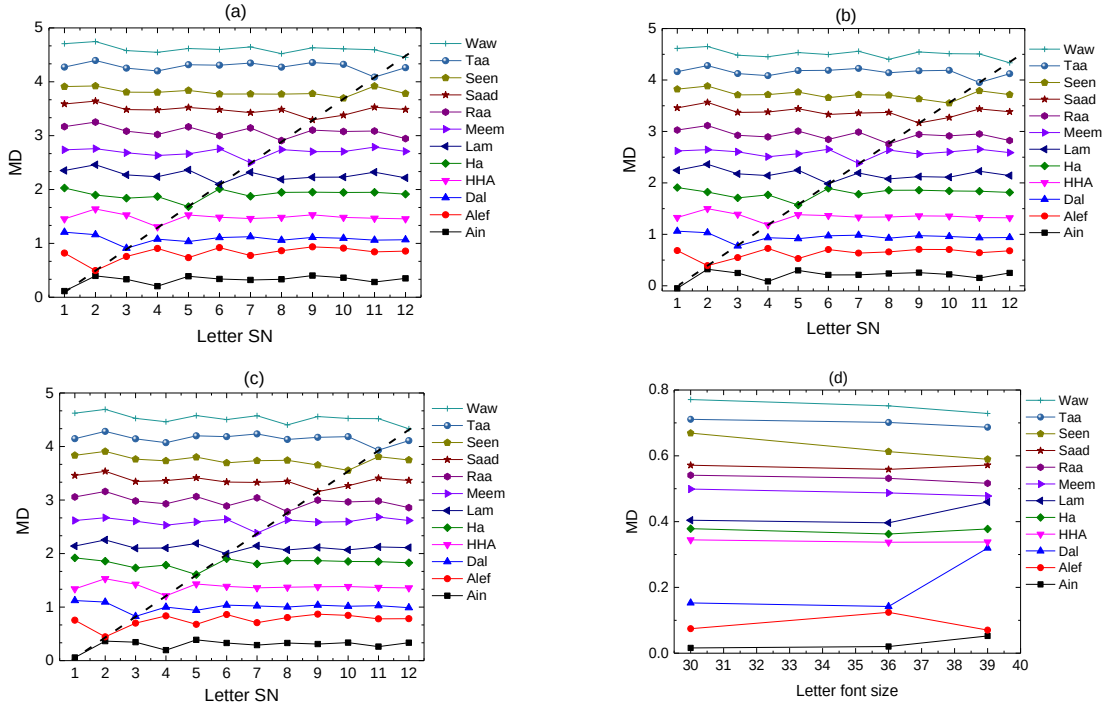

Figure 5. The MD for each letter where the letter SN from the Table (1) for different font size (a) 30 (b) 36 (c) 39 and (d) the letter font size as a function of MD

Algorithm (1): create a database
Input: Insert images of the Arabic letters.
Output: Database (DB) contains 19 features of 12 letters (DB (19x12)) at size 16 of Times New Roman.
Algorithm Steps:

1) Convert the input image to a binary image with 0 and 1.
2) Remove black holes with values 0 that surround with 1.
3) Resize of the input image to 100 x 60 pixels.
4) Find the centre of the resized image (cx, cy).
5) Calculate the number of pixels with value 1.
6) Four parts (C1:4) divided image and the area is calculate for all parts.
7) Apply Soble edge detection method.
8) Four parts (E1:4) divided image edges and the area is calculate for all part.
9) Find the first pixel of 1 in the first row, last column, last row and first column, (x1, y1), (x2, y2), (x3, y3), and (x4, y4), respectively. Connect these points to get four lines L1, L2, L3, and L4 then calculate the length of each line.
10) For all rows find the first and last pixel that equal to 1. Plot straight line by connecting these two points.
11) Account the number of lines that do not have zeroes in the middle, and those that have zeroes in its middle.
12) Find the longest and shortest straight line without zeroes in middle and its location in every row namely T1, T2, Tpo1, and Tpo2 respectively.

Algorithm (2): letter recognition
Input: load letter image, then load database DB (19x12) from the algorithm (1).
Output: Recognize the Arabic letter.

1) Input the Arabic letter image.
2) Extract 19 geometric features as in algorithm (1).
3) Use minimum distance technique to find the matching between the extracted 19 features between the input letter image and the DB as:

$$MD = \frac{|A-DB|}{|A+DB|}$$

4) End.

Table 2. Geometric Features of the Arabic Letter

Property	Meaning	Details
W		Width of Image (100)
H		Height of Image (60)
Image size		W*H (100*60)
Black pixels	Convert image to binary	Convert image to zero one
	Clear the black space	Take the character shape
Holes	Fill image regions and holes	Take the letter without spaces
	Determine the centre of the character image	The centre of character shape (cx, cy).
	Calculate the number of pixels whose value is equal to 1	Calculate the number of white dots associated with the character (one feature)
Points	Dividing image into four equal parts (C1:4)	Calculate the area of the letter in each part (four features).
	Detection the edges of image.	Calculate the number of points in the edges of a character using the Sobel coefficient.
	Dividing image into four equal parts (E1:4)	Calculate the length of the rib drawn between the four points (four features).
straight line	Find the (x1, y1), (x2, y2), (x3, y3), (x4, y4) coordinates of the first pixel that equal to 1 in the first row, last column, last row and first column, respectively.	Plot square shape and calculate the length of each side, L1, L2, L3, and L4 (four features).
	determine the coordinates of the first and last pixel that equal to 1	Calculate the longest straight line connected and its location (one feature).
	Account the number of lines that don't contain zeroes pixel in the middle, and those that contain zeroes in its middle.	Calculate the shortest straight line connected and its location (one feature).
	Find the longest and shortest straight line that doesn't contain zeroes in middle and its location in any row	Calculate the longer and shorter straight and intermittent line and position T1, T2, Tpo1, and Tpo2 (four features).
	Applying the minimum distance technique to recognize input image character	Find the error of the matching between the database and the recognized letter.

6. CONCLUSION

The geometrical character of some Arabic letter changed with increase of the font size in the word software. This can be noticed from Figure 5 (d), the MD is changing with the font size for some Arabic letter while some of them are constant. The suggested new features give high accuracy to recognize Arabic letter depending on newly designed two algorithms. The 19 features geometry selected carefully to recognize similar Arabic letter shapes. The criteria to distinguish between the results depends on minimum distance. This is shown from the dashed line in Figure 5 which locate the minimum value of the MD.

REFERENCES

[1] I. Supriana, A. Nasution, Arabic Character Recognition System Development, Procedia Technology, 11 (2013) 334-341.
[2] M. Sarfraz, M.J. Ahmed, S.A. Ghazi, Saudi Arabian license plate recognition system, in: International Conference on Geometric Modeling and Graphics, 2003, pp. 36-41.
[3] Hesam Izakian, Seyed Amirhasan Monadjemi, Behrouz Tork Ladani, K. Zamanifar., Multi-Font Farsi / Arabic Isolated Character Recognition Using Chain Codes, in Conference Proceedings, World Academy of Science, Engineering and Technology 2009, pp. 58-61.
[4] G.A. Abandah, M.Z. Khedher, Analysis of Handwritten Arabic Letters Using Selected Feature Extraction Techniques, International Journal of Computer Processing of Languages, 22 (2009) 49-73.
[5] Ahmad T. Al-Taani, Saeed Al-Haj, Recognition of On-line Arabic Handwritten Characters Using Structural Features, Journal Of Pattern Recognition Research, 1 (2010) 23-37.
[6] Z. Abdelmalek, On Multiple Typeface Arabic Script Recognition, Research Journal of Applied Sciences Engineering and Technology, 2 (2010) 428-435.
[7] Neila Mezghani, A. Mitiche, On-line recognition of handwritten Arabic characters using A Kohonen neural network, in Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02), IEEE, Canada, 2002, pp. 490–495.
[8] S. Mozaffari, K. Faez, M. Ziaratban, Structural Decomposition and Statistical Description of Farsi/Arabic Handwritten Numeric Characters, in Proceedings of the Eighth International Conference on Document Analysis and Recognition, IEEE Computer Society, 2005, pp. 237-241.
[9] N.B. Amor, M. Zarai, N.E.B. Amara, Neuro-Fuzzy approach in the recognition of Arabic Characters, in 2006 2nd International Conference on Information & Communication Technologies, 2006, pp. 1640-1644.
[10] H. Majid, M. Dzulkifli, R. Abdolreza, Deductive method for recognition of on-line handwritten Persian/Arabic characters, in 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE), 2010, pp. 791-795.

[11] B.M.F. Bush of, M. Spann, Segmentation and recognition of Arabic characters by structural classification, Image and Vision Computing, 15 (1997) 167-179.
[12] S.A. Mahmoud, A.S. Mahmoud, The use of Hartley transform in OCR with application to printed Arabic character recognition, Pattern Analysis and Applications, 12 (2008) 353.
[13] W. Guang, M. Baraldo, M. Furlanut, Calculating percentage prediction error: A user's note, Pharmacological Research, 32 (1995) 241-248.