Indonesian Journal of Electrical Engineering and Computer Science
Vol. 14, No. 3, June 2019, pp. 1443~1451
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v14.i3.pp1443-1451

Off-line handwritten character recognition using an integrated DBSCAN-ANN scheme

Dhurgham Ali Mohammed[1], Alaa Abdul Hussein Mezher[2], Hayder Sabeeh Hadi[3]
[1]Faculty of Education for Girl, Department of Computer Science, University of Kufa, Iraq
[2]Faculty of Computer Science and Mathematics, Department of Computer Science, University of Kufa, Iraq
[3]Faculty of Nursing, University of Kufa, Iraq

Article Info

Article history:

Received Nov 27, 2018
Revised Jan 21, 2019
Accepted Feb 10, 2019

Keywords:

Density based clustering
Features extraction
Handwritten arabic cheaters recognition
Image processing

ABSTRACT

Handwriting character recognition involves a high degree of variability and imprecision. For that, the main factor to judge the recognition accuracy is the technique that is used to extract the features. This paper developed a novel method for handwritten Arabic characters by combining the Density-Based Clustering method with statistical and morphological features. The first stage in recognition of handwritten character image has been done by binarization the image then applies noise removal techniques. The Density-Based Algorithm used to categorize and find any shape of clusters based on pixel information positions. This technique divided the image into characters. Each character will be decomposing into four regions from the centroid followed by feature extraction. These features include vertical and horizontal projections, upper and lower profile, rectangularity and orientation. The results of the present process will transfer to the Neural Network (NN) stage which generates a high level of correctness and accuracy by training. The testing results compared with two of state-of-art researches. The total accuracy of this proposed work observes a better recognition of characters.

Corresponding Author:

Alaa Abdul Hussein Mezher,
Department of Computer Science,
University of Kufa,
An Najaf City, Iraq.
Email: alaa.abdullah@uokufa.edu.iq

1. INTRODUCTION

The task of recognizing the Arabic handwritten alphabets have been an attractive research problem. It is used in Africa and Asia besides Arabic [1]. The challenges in Arabic handwriting are the variety in both size and shape. The shape of a character, overlaps, and interconnections between the neighboring of characters are the main difficulties in addition to the mood of the writer. There are 28 basic Arabic characters. However, the set of alphabet observes 84 different shapes based on the position of the letter related with the beginning, middle or isolated [2], [3]. Also, some Arabic letters have secondary components in (dot) form. The number of dots, dot position and letter position are very important features. The number of dots presents another classification in Arabic alphabetic. It consists from two, three or four elements depending on the number of dots [3]. Table 1 presents some samples.

Another challenge in Arabic characters is the (Hamza) in the letter (Alif (أ)). This character can be drawn with or without it [2]. Many techniques have been presented in this field. Granlund in 1972 used Fourier transformations for feature extraction. The features were genuine shape constants such as size, location, and orientation [4]. Almuallim and Yamaguchi in 1987 applied structural features and skeleton representation for word recognition.

Table 1. Similar Arabic Alphabets Samples

No. of dot	Letter specification	Isolated	Initial	Medial	Final	Letter
One dot	Letter without dot	ع	عـ	ـعـ	ـع	Ain
	Letter with dot	غ	غـ	ـغـ	ـغ	Ghain
mix. dots	Letter with one dot	ف	فـ	ـفـ	ـف	Faa
	Letter with two dots	ق	قـ	ـقـ	ـق	Qaf
two dots	Letter without dot Lower position	ي	يـ	ـيـ	ـي	Yaa
	Letter with dot Upper position	ت	تـ	ـتـ	ـت	Taa
Three dots	Letter without dot	ش	شـ	ـشـ	ـش	Sheen
	Letter with dot	س	سـ	ـسـ	ـس	seen

The main process is to segment the words into "strokes". They achieved 91% of word recognition [5]. Al-Yousefi and Udpa in 1992 proposed a statistical method for character recognition of Arabic. The main idea of this method is to segment the Arabic character into two parts, primary and secondary for instance the dots and small marking. The results accuracy was varied between 81% and 98.79% based on the characteristics [6]. Sano *et al.*, in 1996, proposed a new approach by applying a structural fuzzy relations base on Arabic isolated character recognition. They used multi-patterns based on the number of selected characters. After that the sub-pattern will characterized based on the basic shape elements such as straight line, circle and diacritical points similarity [7]. Dehghani *et al*, in 2001 proposed hidden Markov models (HMMs) for isolated handwritten Persian characters.

Two types of feature vectors were applied in this method, the performance of this method (V_HMM, H_HMM) and the combination in classifier method reached to 71.82% [8]. Mario Pechwitz and Volker Maergner presented in 2003 semi-continuous one dimension HMM. Pixel value have been used in this method as a rudimentary features detected by rectangular window. The achieved performance was about 89% [9]. Mozaffari *et al*., in 2005, used a skeleton based on statistical features of primitives' partition. The recognition level was 94.44% [10]. El Abed and V. Margner in 2007 applied sliding window based on pixel features extraction. The method used skeleton direction using feature extraction and achieved a89% rate of recognition [11].

Hamdani *et al.*, in 2009, developed a new Arabic Handwriting Recognition method by combining the feature extraction methods with one on-line method. The methods were pixel values, densities and Moment Invariants, and pixel distribution and Concavities. These features correlated with online features in order to segment each part of the word (PAW) based on 21 features. The IFN/ENIT database applied and evaluated in the present system [12].

Jin Chen *et al*., in 2010, used Gabor features vectors method correlated with a set of structure, gradient and concavity features (GSC). The presented work, a Gabor filter is used for features extraction. They applied support vector machine (SVM) for classification. The results observed 79.7%, 82.8% and 84.3%rate of recognition for the combination of a graph with GSC, the combination of proposed Gabor and graph and the combination of proposed Gabor and GSC respectively [13].

Lawgali *et al*., in 2011, Developed a comparison between Discrete Cosine Transformation (DCT) and Discrete Wavelet Transformation (DWT) [3]. The Artificial Neural Network has been used to classify the coefficients of both techniques. The recognition rate of DCT 96.56% and DWT technique was 59.81% in the best cases [3].

Eraqi and Abdelazeem in 2012 used a novel approach for feature extraction and diacritics detection. The method combined the efficient dependent and independent baseline features of the selected image. The process was applied before and after removing the diacritics segments. The rate of recognition was between 96.01% and 96.78% [14]. Sahlol1 and Suen in 2014 used the whole body features and the second component features. The results observe an 88% rate of recognition [1].

Al-Helali and Mahmoud in 2016 developed a framework for recognition of Arabic characters. They have processed Arabic recognition of delayed strokes. The statistical features evaluated for all Arabic characters. Bhuiyan and Alsaade in 2017 proposed a BAMMLP method for Arabic character recognition. This method converts the Arabic characters into a matrix of features (MxN). The organization of the system was by using Bidirectional Associative Memory (BAM) correlated with Multi-Layer Perception (MLP) [15].

Al-Jubouri and Abusaimeh also in 2017 proposed a two-stage recognition system to develop an isolated handwritten Arabic offline recognition. The first stage is Support Vector Machine (SVM) and the second stage is Neural Network (NN). The present purpose of using two stages was to reduce the load of a classifier with a detection rate of 92.2% [16].

This paper develops a reliable offline OCR system for Arabic character recognition using Density-Based Algorithm (DBSCAN). The system is organized with features extraction system and a Neural Network (NN) selected as a training technique for recognizing the characters efficiently.

2. THE PROPOSED METHOD

The proposed Arabic word recognition system is geared towards the state-of-the-art offline text technique methods. The handwritten character images IFN-ENT dataset is used to cover specific shapes of Arabic characters [16]. It consists of more than 2900 various characters with Bitmap image type. The methodology starts with binarization of word image followed by division of the selected word into letter segments. The overall proposed model is represented in Figure 1.

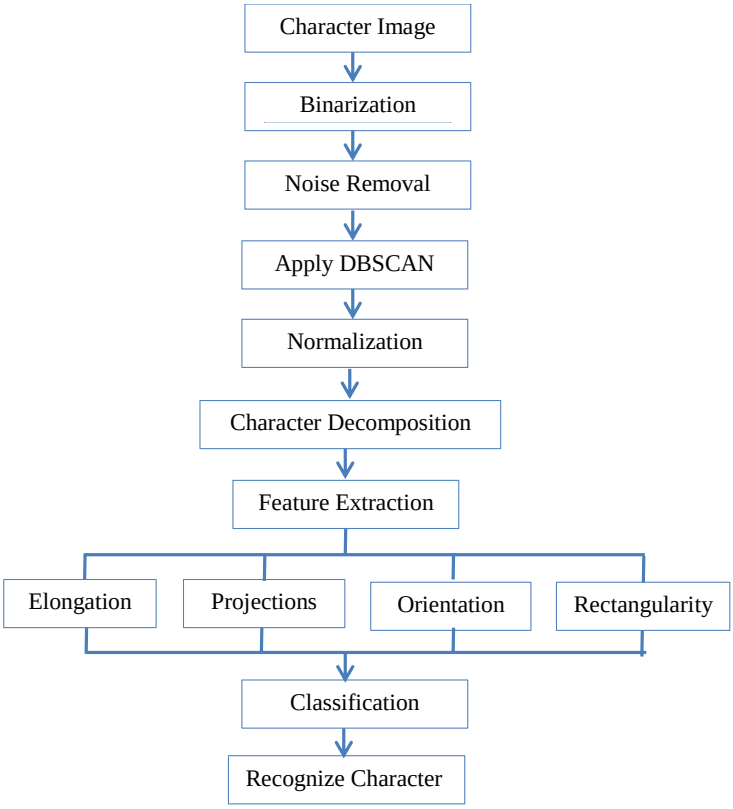

Figure 1. Flow diagram of system methodology

Each of these modules presented in details in subsequent sections. The next sections describe the methodology of the present work including Binarization, noise removal Algorithms, applying DBSCAN technique, normalization, feature extraction techniques, Classification, training and testing phase.

2.1. Binarization

The inserted images are generally segmented from background using binarization, which is actually segmentation into two classes. In this regard, the well-known Otsu's thresholding algorithm (considered to be a benchmark) has been employed to compute threshold from the grayscale image. The Otsu algorithm contains two classes of pixels (background and foreground) using the histogram based image t hresholding [17].

2.2. Noise removal

The noise removal technique can be described as the effect of slightly distorting of the real image, median filtering has been used in this work for reducing random noise. The present filter applied the sort of median filter all over the image with scattered pixels of noise and effectively got rid of the noise [18].

2.3. Density-Based Algorithm

Density based algorithm (DBSCAN) is defined as a data clustering method. This method developed by Ester in 1996 to discretize the area into several small typical density points [19]. The main idea from this method is to specify a position p in the continuous ID domain (x- axis) based on the formula [20], [21].

The basic concept of this technique is to analyze the domain data in order to propose a logical division. In our case, the set of points in a domain are closely packed together in order to investigate the relationship between the pixels. The specified position p contains m-by-n neighborhood. The domain will define all the information in the pixel domain based on the topological and statistical features as in the formula below:

DBSCAN can categorize and find any shape of clusters based on pixel information positions that lie close to each other in Arabic character by computing process of four definitions.

Definition 1: (Eps-neighborhood): The Eps-neighborhood of a point Ƥs is defined by the cluster region Nr that represents the space character area. The Eps-neighborhood has the existing character. It also has a center point that represents the center of character area.

Definition 2: (directly density-reachable): Directly density-reachable is the character center point Ƥs which can be reached.

Definition 3: (Density-reachable): Density-reachable is the point that can be reached through the specified character area.

Definition 4: (cluster): In the present use of DBSCAN algorithm, consider each cluster C is density-reachable with maximum rank of Ƥ from point Ƥs; Hence: "∀ Ƥ ϵ C" is density-reachable from Ƥs with respect to Eps-neighborhood.

The aim of using this method is to detect the pixel information to use them in grouping the data, to find each group specification including topological features such as endpoints, pixel ratio and height to width ratio and to specify the statistical features such as connected components in the domain. The method working in both x-direction and y-direction is supported by the mathematical model as shownin Table 2.

Table 2. Results of DBSCAN Technique

No.	DBSCAN process	Feature
1	Maximum point in x-direction	Upper Profile
2	Minimum point in x-direction	Lower Profile
3	Maximum point in y-direction	Baseline profile
4	Zero pixel in x-direction	Extract the separated characters
5	Lowest pixels density	Extract the connected characters
6	Cluster character	character elements, area, pixel density
7	Determine the centroid	Centroid

$$Cov_i = \sum_{i=0}^{m} x_i \quad (1)$$

$$Cov_j = \sum_{j=0}^{n} y_j \quad (2)$$

$$Sum_i = \sum_{i=min}^{i=max} Cov_j \quad (3)$$

$$Sum_j = \sum_{j=min}^{j=max} Cov_i \quad (4)$$

$$A = \iint (Cov_i \, , Cov_j \,) \, dxdy \quad (5)$$

$$C_{cen} = \iint \frac{(Cov_i)}{A} \, dA \quad (6)$$

Where Cov is covered pixels, A is Area, Ccen is centroid, Sum is pixels summation, n is raw numbers and m is column numbers. The present sequence will provide a set of results. Table 2 shows the result and specify a set of image features. The overall proposed method is illustrated in Figure 2 through a block diagram.

Each of these modules is discussed in detail in the subsequent sections. The proposed sequence of operations is performed based on the scanned image. The main process is enhancing the input image to be suitable for segmentation. The DBSCAN process applied on the words that have been analyzed, specifying the upper line, lower line, and baseline, passing by the extraction of characters and specifying the centroid as shown in Figures 3 and 4.

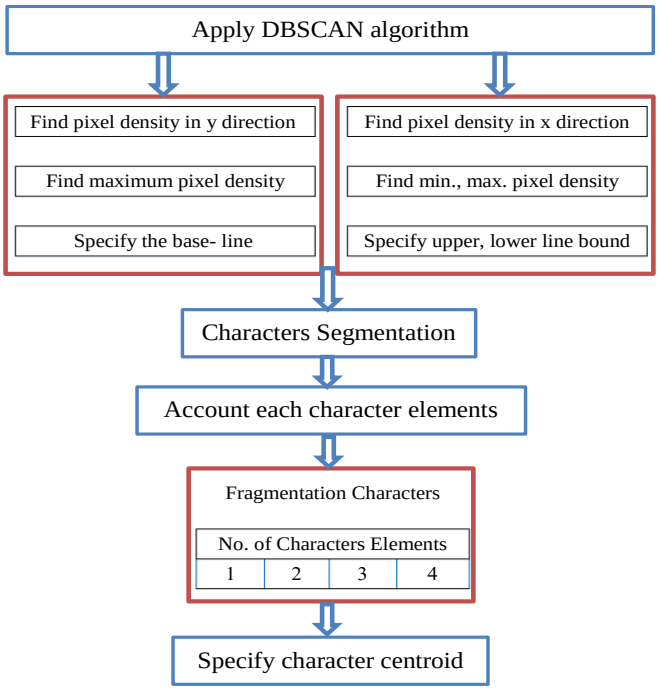

Figure 2. DBSCAN technique adopted in the present method

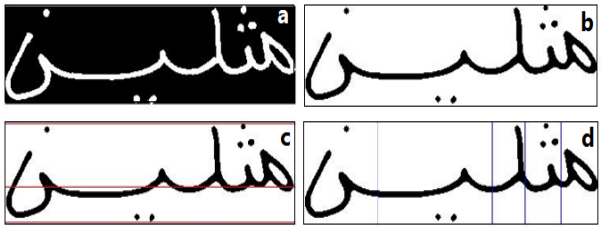

Figure 3. Characterizing the image

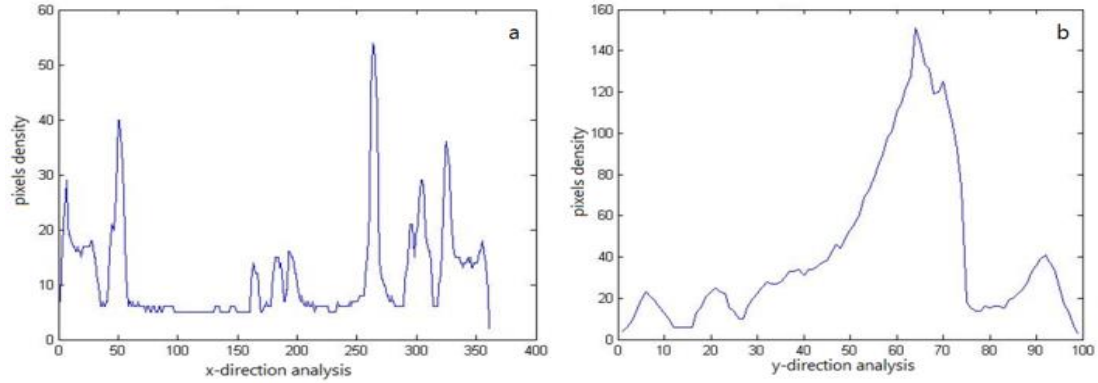

Figure 4. DBSCAN analyses the data

3. CHARACTER DECOMPOSITION

After specifying the centroid, the process of image decomposition on the four regions based on the image centroid. This will be applied to divide the image into four regions as shown in Figure 5.

The reason for this step is to investigate the image statistics in topological features based on the character components. The character component will use the centroid as a unique position for each character.

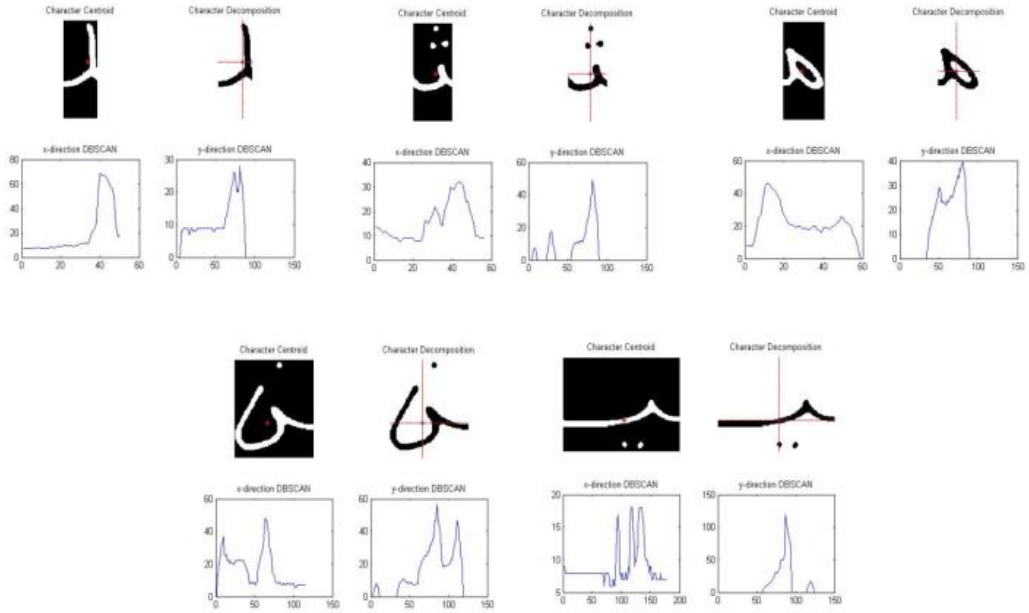

Figure 5. Decompose the image

4. FEATURE EXTRACTION

4.1. Horizontal and Vertical Projections

Projections give a count of the number of black pixels in the row and in the column of the fragmented images [22], number of horizontal projection pixels generated by counting of black column pixels of fragmented images. Similarly, vertical projection counts the black number of fragmented image pixels in each raw. The horizontal and vertical projection can be taken from the DBSCAN data results.

4.2. Orientation

Orientation features is applied to compute the direction or slope of a stroke in the fragmented image [23]. The orientation of the fragment is measured based on the angle between the major axis and the x-axis of an ellipse approximating the fragment. The orientation of the Arabic letters can be vertical such as (غ,ل,ر,ح,ا) or horizontal such as (ف,ظ,ن,ك,ت). For the Arabic script, it is clear that some letters are vertically oriented as (غ,ل,ر,ح,ا) and others are oriented horizontally as (ف,ظ,ن,ك,ت).

4.3. Rectangularity

Rectangularity is defined as the ratio of element area to its total bounding box area. The term bounding box can be defined as a smallest rectangular that enclose the shape of writing in a fragment [24]. Also, all the data can be taken from the DBSCAN results.

4.4. Elongation

Elongation represents the aspect ratio of the fragmented character. It helps to discriminate between non-elongated and elongated shapes. Elongation is defined as the height to width ratio in bounding box [24]. Figure 6 shows a bounding box extracted from a fragment which encloses a stroke. Elongation can be expressed as below:

$$\text{Elongation} = \frac{\text{lb}}{\text{sb}} \tag{7}$$

Where lb is the long side in the bounding box and sb is shorter side of bounding box as shown in Figure 6.

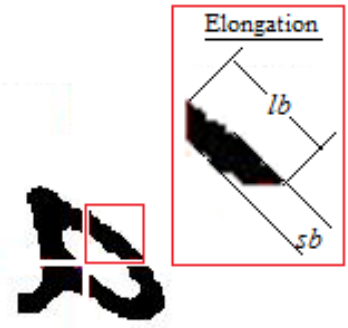

Figure 6. Elongation with bounding box

5. INTEGRATED DBSCAN-ANN

An integrated DBSCAN-ANN scheme has been developed based on character features extraction and character recognition. The neurons of input representation can be determined by feature vector length. Also, the input characters considered 168 elements based on 28 neurons as an output layer. The processes identified the characters based on two layer log-sigmoid transfer function which considered as perfect for learning. The function generates output range between 0 and 1. Also, the network date randomly divided into two categories. The first is for tanning which is considered 80% of the data and the second is 20% which is used for testing the system. Back propagation training method is used based on principle of gradient descent. Gradient descent is an optimization algorithm applied to minimize a cost function (cost) and is used to find the values of parameters (coefficients). The training process stopped when the square error summation falls below 0.001. The neurons number of hidden layers specified by trial and error, also the starting number was 20 neurons.

6. RESULTS AND DISCUSSION

There are 308 characters used as a test set, eleven different samples of each of the 28th characters. The experimental results shown in Table 3 represent the rate of recognition of each character.

Table 3. Results of Arabic Letters Recognition

Character	Rate of Recognition		Character	Rate of Recognition	
	Sahlol and Suen 2014	Present study		Sahlol and Suen 2014	Present study
أ	96%	98%	ض	72%	81%
ب	87%	95%	ط	91%	100%
ت ث	69%	96%	ظ	83%	86%
ج	76%	91%	ع غ	66%	79%
ح خ	100%	100%	ف	99%	100%
د ذ	94%	100%	ق	100%	100%
ر	100%	100%	ك	61%	82%
ز	83%	87%	ل	81%	89%
س ش	88%	88%	م	96%	100%
ص	89%	89%	ن	92%	93%
	80%	89%	ه	100%	100%
	78%	88%	و	97%	100%
	88%	100%	ي	100%	100%
	100%	100%		---	88%

The results showed that the rate of characters recognition was 93.54% for all letters which represent better than the previous studies such as Sahlol (obtained 88%) and Al-Jubouri (obtained 92.2%). The statistical and structural features obtained from the small character fragments present superior results. The division represents one of the most crucial steps. It provides four subgroups based on the character number of elements. Also, each subgroup character divided into four fragments based on the centroid of the character then the features of each fragment detected. The present process improves the recognition rate because of character singularity. The similarity of [(ت) Taa and Thaa (ث), (س) Seen and (ش) Sheen] solved by the differences in element account. The character (ض) has the lowest recognition due to the big variety in the character as shown in Figure 7.

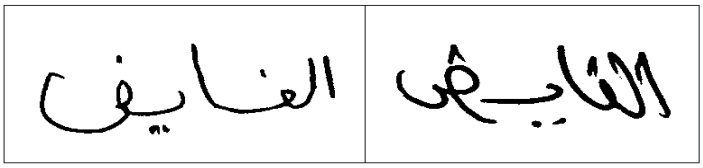

Figure 7. The character (ض) in different words

Analyzing the character observe a big characteristic differences as shown in Figure 8. It is seen the centroid position, character decomposition and DBSCAN analysis are different. Also, the dot shape causes differences in the number of pixel density in y-direction; they were separated in Figure 8(a) and connected in Figure 8(b).

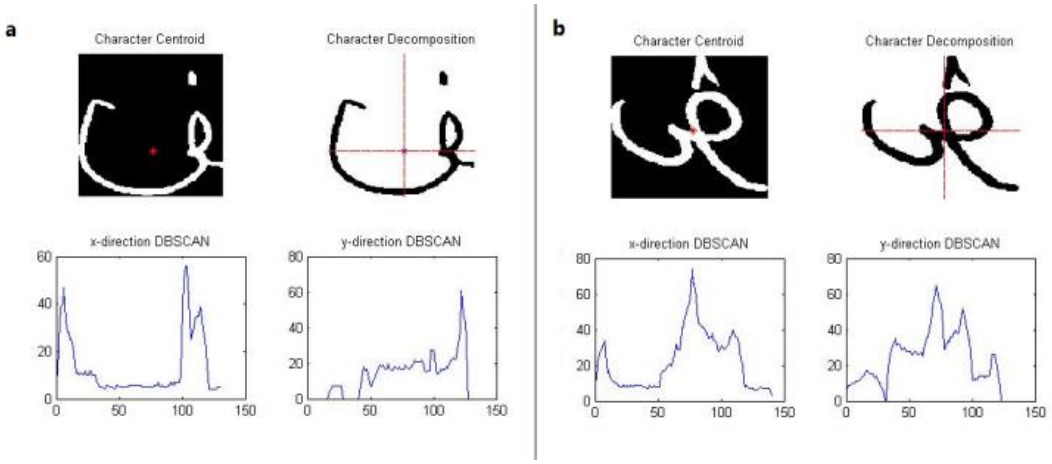

Figure 8. The character (ض) analysis

7. CONCLUSION

The present scheme is used for extracting the Arabic character features to achieve high recognition accuracy. The used techniques during the characters processing started with binarization and noise removing. These techniques presented to enhance the image letter for Density-based process. The algorithm clustered the letters and extracts the statistical features. The structural features are also investigated to obtain six features for each character. It is concluded from this work that the character elements are one of the major factors in results accuracy. The character elements features reflect the character specification. For future work, further investigation will be extended to other causable patterns.

REFERENCES

[1] Sahlol, A, and C Suen. 2014. "A Novel Method for the Recognition of Isolated Handwritten Arabic Characters." arXiv Preprint arXiv:1402.6650. http://arxiv.org/abs/1402.6650.

[2] Kacem, Afef, Nadia Aouiti, and Abdel Belaïd. 2012. "Structural Features Extraction for Handwritten Arabic Personal Names Recognition." *Proceedings - International Workshop on Frontiers in Handwriting Recognition, IWFHR*, 268–73. https://doi.org/10.1109/ICFHR.2012.276.

[3] Lawgali, A, and A Bouridane. 2011. "Handwritten Arabic Character Recognition: Which Feature Extraction Method." *International Journal of Advanced Science and Technology* 34 (September):1–8. http://nrl.northumbria.ac.uk/1908/.

[4] Granlund, Gösta H. 1972. "Fourier Preprocessing for Hand Print Character Recognition." IEEE Transactions on Computers 100 (2). IEEE:195–201.

[5] Almuallim, Hussein, and Shoichiro Yamaguchi. 1987. "A Method of Recognition of Arabic Cursive Handwriting." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 5. IEEE:715–22.

[6] Al-Yousefi, H, and S S Udpa. 1992. "Recognition of Arabic Characters." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14 (8). IEEE:853–57.

[7] Sano, M, T Kosaki, and F Bouslama. 1996. "Fuzzy Structural Approach for Recognition of Handwritten Arabic Characters." In Proc. Int. Conf. on Robotics, Vision and Parallel Processing for Industrial Automation, Ipon, Malaysia, 252–57.
[8] Dehghani, A, F Shabini, and P Nava. 2001. "Off-Line Recognition of Isolated Persian Handwritten Characters Using Multiple Hidden Markov Models." In Information Technology: Coding and Computing, 2001. Proceedings. International Conference on, 506–10.
[9] Pechwitz, Mario, and Volker Maergner. 2003. "HMM Based Approach for Handwritten Arabic Word Recognition Using the IFN/ENIT-Database." In Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on, 890–94.
[10] Mozaffari, Saeed, Karim Faez, and Majid Ziaratban. 2005. "Structural Decomposition and Statistical Description of Farsi/Arabic Handwritten Numeric Characters." In Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on, 237–41.
[11] Abed, Haikal El, and Volker Margner. 2007. "Comparison of Different Preprocessing and Feature Extraction Methods for Offline Recognition of Handwritten Arabicwords." In Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, 2:974–78.
[12] Hamdani, Mahdi, Haikal El Abed, Monji Kherallah, and Adel M Alimi. 2009. "Combining Multiple HMMs Using on-Line and off-Line Features for off-Line Arabic Handwriting Recognition." In Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on, 201–5.
[13] Chen, Jin, Huaigu Cao, Rohit Prasad, Anurag Bhardwaj, and Prem Natarajan. 2010. "Gabor Features for Offline Arabic Handwriting Recognition." In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, 53–58.
[14] Eraqi, Hesham M, and Sherif Abdelazeem. 2012. "HMM-Based Offline Arabic Handwriting Recognition: Using New Feature Extraction and Lexicon Ranking Techniques." In Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on, 554–59.
[15] Bhuiyan, Md Al-Amin, and Fawaz Waselallah Alsaade. "On Arabic Character Recognition Employing Hybrid Neural Network." *International Journal of Advanced Computer Science and Applications* 8.6 (2017): 96-101.
[16] Al-Jubouri, Mohamed Anas Hussein. "Offline Arabic Handwritten Isolated Character Recognition System Using Support vector Machine and Neural Network." *Journal of Theoretical & Applied Information Technology* 95.10 (2017).
[17] Abdul, Hameed M, and Taghreed A Naji. 2015. "Images Segmentation Based on Fast Otsu Method Implementing on Various Edge Detection Operators I Ntroduction Otsu Thresholding Method" 28 (3):18–28.
[18] Aminuddin, Nur Shazwani, Masrullizam Mat Ibrahim, Mohd Ali, Syafeeza Ahmad Radzi, Wira Hidayat, and Mohd Saad. 2017. "A New Approach To Highway Lane Detection By Using Hough Transform Technique." *Journal of Information and Communication Technology* 2 (2):244–60.
[19] Ester, Martin, Hans P Kriegel, Jorg Sander, and Xiaowei Xu. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 226–31. https://doi.org/10.1.1.71.1980.
[20] Moreira, Adriano, Maribel Y Santos, and Sofia Carneiro. 2005. "Density-Based Clustering Algorithms – DBSCAN and SNN." University of Minho - Portugal, 1–18.
[21] Trikha, Priyanka, and Singh Vijendra. 2013. "Fast Density Based Clustering Algorithm." *International Journal of Machine Learning and Computing* 3 (1):10–12. https://doi.org/10.7763/IJMLC.2013.V3.262.
[22] Pradeep, J., E. Srinivasan, and S. Himavathi. "Diagonal based feature extraction for handwritten character recognition system using neural network." Electronics Computer Technology (ICECT), 2011 3rd International Conference on. Vol. 4. IEEE, 2011. https://doi.org/10.5121/ijcsit.2011.3103.
[23] Ramadevi, Y, T Sridevi, B Poornima, and B Kalyani. 2010. "Segmentation and Object Recognition Using Edge Detection Techniques." *International Journal of Computer Science & Information Technology (IJCSIT)* 2 (6):153–61. https://doi.org/10.5121/ijcsit.2010.2614.
[24] Descartes, Paris. 2009. "PhD of Université Paris Descartes Classification of Handwritten Documents : Writer Recognition."