International Journal of Technology (2017) 3: 528-538
ISSN 2086-9614

DESIGNING OFFLINE ARABIC HANDWRITTEN ISOLATED CHARACTER RECOGNITION SYSTEM USING ARTIFICIAL NEURAL NETWORK APPROACH

Ahmed Subhi Abdalkafor[1*]

[1]*Career Development Center, University Headquarter, University of Anbar, Anbar, Iraq*

(Received: November 2016 / Revised: February 2017 / Accepted: April 2017)

ABSTRACT

The Arabic language is one of the major languages that has little attention in character recognition field by Arab researchers in particular and foreign researchers in general. Due to the highly cursive nature of handwritten Arabic language, Arabic character recognition is considered one of the most challenging problems in contrast to working with Latin, Japanese or Chinese character recognition. In this paper, we proposed Arabic off-line handwritten isolated recognition system based on novel feature extraction techniques, a back propagation artificial neural network as classification phase. The presented work is implemented and tested via the CENPARMI database. Competitive recognition accuracy has been achieved 96.14%. This result motivates us and other researchers in this field to employ the features extraction techniques that we have used in this research with other Arabic character shapes.

Keywords: Directional features; Regional features; Universe of discourse; Zoning

1. INTRODUCTION

Optical character recognition (OCR) has been an active research field, and still one of the most challenge and dynamic areas of research in computer science. However, the researches in this discipline have a tendency toward maturity and these are associated with a large body of published works. Nevertheless, Arabic character recognition has been given lower proirity related to some researcher's attention, despite of the fact that the Arabic language is considered one of the major languages that is spoken by more than 280 million people (Abuzaraida et al., 2013). Whereas the issue of Arabic character recognition is important not only for native Arabic speakers but also for non-Arabic speakers too, where the Arabic language has been adopted for the alphabetical writing of the non-Semitic languages, such as Kurdish, Malay, Urdu, etc. At these countries, such as Turkey, Malaysia, and India as well as some West African countries, Arabic considered as a second language (Alijla & Kwaik, 2012; AlKhateeb, 2015a; 2015b). Moreover, the researches in the field of optical Arabic character recognition have limited success because the complex natural of Arabic language in general and Arabic scripts in particular. Any OCR system can come in one of two types: printed and handwritten; where in the former, the characters to be identified are printed using the machine, whereas the letter are concerned in recognition (identification) of handwritten characters that written on paper and then scanned by machines (Asebriy et al., 2014). Handwritten recognition systems, on the other hand, come in two major types: On-line and Off-line. In case of on-line handwritten recognition system, the pressure is used on the digital display of an instrument to create a series of points

*Corresponding author's email: ahmed_abdalkafor@yahoo.com, Tel: +9647834120596
Permalink/DOI: https://doi.org/10.14716/ijtech.v8i3.6723

that are traced by pen. However, off-line handwritten recognition system is built on an optical character recognition system and is applied on optically scanned texts. As a thumb of rule, and as a natural result, Off-line handwritten recognition is much harder than On-line handwritten recognition (Rashad & Semary, 2014; Lawgali, 2015). Over the last several years, the working of researchers with Arabic character recognition systems have started to improve efficiency and accuracy recognition. The utilization of an Artificial Neural Network (ANN) in the processes of optical character recognition system applications can considerably improve the performance quality of the recognition with an extremely simple code. An optical character recognition system that is built up based on a principle of 'divide and conquer' algorithm to divide the Arabic characters into four groups was proposed by Hammad and Elhafiz (2015). The proposed system utilized curve tracing and the number of dots of each character as discriminating pieces of information, although Hammad and Elhafiz (2015) used four neural networks to classify each group, but they achieved an accuracy rate of 88.11%, which is comparable with our proposed system that used only one neural network to classify all characters.

In this paper a robust and reliable optical recognition system for Off-line Arabic handwritten isolated characters, based on back propagation artificial neural network is proposed, which takes advantage of two powerful feature extraction techniques which lead to high overall success rate, and stable optical recognition system.

2. METHODOLOGY

A variety of Arabic handwritten isolated techniques have been proposed, developed and implemented, but do not reach fit to achieve high recognition accuracy. Our proposed system, which consists of two main phases: Database Collection and Character Recognition. Character Recognition Phase can be further subdivided into three major stages: Image Preprocessing, Features Extraction, and Classification. Figure 1 shows the block diagram of our proposed system.

2.1. Database Collection

In this paper, we have used CEMPAMI database since it is a novel database that composed of Off-line Arabic handwritten isolated characters (Alamri et al., 2008). It contains (21,426) characters written by (328) writers.

2.2. Character Recognition

Of four major cascaded stages, where first the character image undergoes a pre-processing step that enables the character image ready for the features extraction stage. Then, the database of feature vectors must be checked to determine if it is suitable to represent the character image and this representation distinguish it from another character, which ensures that the classification via back propagation ANN will yield high performance.

2.2.1. Preprocessing universe of discourse and skeletonization

The universe of discourse of a character image can be defined as the shortest universe that encloses the character itself. In our proposed system, the image is represented as a binary matrix of ones and zeros that represent the character itself and the white space around it respectively. Therefore, the universe of discourse will be the shortest matrix that fits the whole character skeleton. Figure 2 shows the universe of discourse of letters: ***Ha*** **(هـ)** character. Where skeletonization is considered one of the morphological operations where the pixels on the boundaries of object (character) are removed, but at the same time it does not allow the object (character) to break apart. The remaining pixels make up the image skeleton. Figure 3 shows the character ***Ha*** **(ـهـ)** and how it undergoes a skeletonization process.

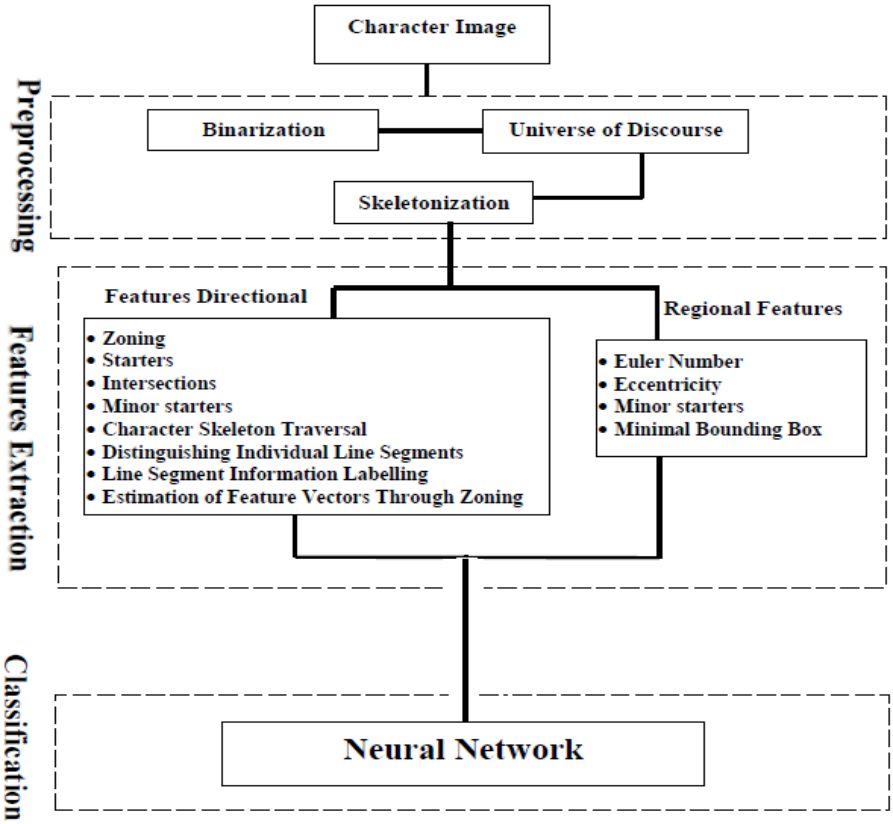

Figure 1 The block diagram of our proposed system

Figure 2 Universe of discourse

Figure 3 The skeleton of ***Ha*** (هـ) character

2.2.2. Features extraction

Two major types of feature extraction techniques were investigated in the implementation of our proposed system: (1) Directional Features; and (2) Regional Features. The following subsections will discuss these types of features extraction.

Directional Features

We depend highly on the novel directional features extraction technique that proposed by (Blumenstein et al., 2003) which were used with English handwritten letters. This novel technique sought to simplify (divide and conquer) the skeleton of a character through identification of individual stroke-on-line segments in the character image. Dileep (2012) implemented this feature extraction technique through zoning the character image into 3×3 (nine) zones and extract the features of each zone individually.

In our proposed system, we used three zoning techniques, two of which were proposed and implemented for English letters by Dileep (2012), especially where divided the character image into three horizontal zones and 3×3 zones, respectively. For further enhancement and to development the distinctive feature of our proposed system, we have divided the image into three vertical zones. Then we extracted the features of each zone individually and concatenated the features of each type of zoning in one feature vector for each character image. From read

the latest researches, no one used this kind of features extraction technique to recognize the Arabic characters. By using this approach, we guarantee that all fine details of Arabic characters are taken into consideration, which help the back propagation ANN in the classification stage dramatically.

Zoning

The universe of discourse is determined, the character image is divided into two types of zoning: (1) zoning along one dimension along the column or row dimension of image; or (2) two-dimension zoning. Figure 4 shows zoning along the column dimension.

Figure 4 Zoning along the column (one dimension)

Figure 5 represents the row zones that are result when the skeletonized character image divided into three horizontal zones. In case of two-dimensional zoning, the character image is zoned into nine equal-sized windows as shown below in Figure 6.

Figure 5 Zoning along the row (one dimension)

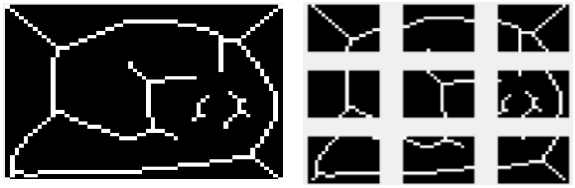

Figure 6 Zoning along 9 zones (two dimension)

Starters, Minor Starters and Intersections

The features extraction technique highly depends on the different typed of segments that can be defined in each zone.

In Figure 7 the zone of 33 positions is extracted and the line segments of this particular zone are highlighted.

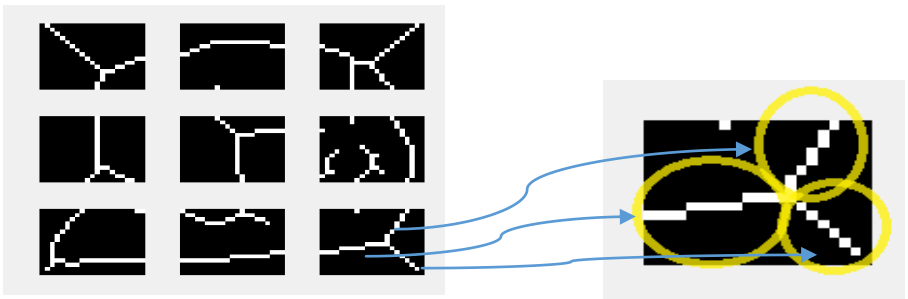

Figure 7 Highlighted line segments of a particular zone

To identify the line segments that will establish our features database, certain pixels in the character skeleton were the first defined as: starters, minor starters, intersections, and the entire skeleton of that zone will be traversed entirely pixel by pixel.

Starters

Starters are defined as those pixels with one neighbour where the neighbourhood of a pixel is defined as all pixels that immediately surround the pixel under consideration. Before the

traversal of the character skeleton begins, all the starters in the particular zone of the character image are identified and then populated in an array. Figure 8 shows the starters of ***Tha*** (ظ) and ***Meem*** (م) characters respectively.

Intersections

Intersections can be defined as those pixels that have more than two neighbours. Although this criterion is not sufficient, it is necessary to define a pixel as an intersection point between multiline segments. Therefore, the process of identification of intersection points is somewhat more complicated than that in case of starters. Figure 9 shows some of intersection points in ***Ta*** **(ط)** character.

Figure 8 Starters of Tha (ظ) and Meem (م) characters

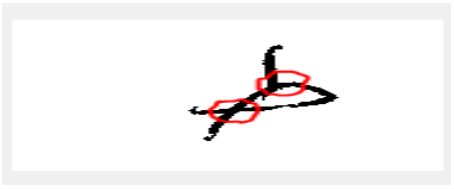

Figure 9 Intersection points in ***Ta*** (ط) character

To overcome this issue, a new property called 'true neighbours' is defined for each Traversed Pixel. Based on the number of true neighbours for a particular pixel, a Traversed Pixel is classified as an Intersection Point (Pixel) or not. For this reason, neighbouring pixels have been classified into two main categories: Direct Pixels and Diagonal pixels, which are used, as illustrated above. Now, for the sake of finding the number of the neighbours for specific pixel, each pixel is under consideration as follows. Firstly, it has to be classified in greater detail, based on the number of neighbours it has in the character skeleton. According to the logics outlined below, there are three cases of neighbourhood that will determine if the current pixel is a true intersection or not (Dileep, 2012).

Case 1: Three Neighbours
If Any one of the direct pixels is adjacent to anyone of the diagonal pixels.
Then The Pixel under consideration cannot be an Intersection.
Else
If None of neighbouring pixels are adjacent to each other.
Then The Pixel under consideration is an Intersection.
Case 2: Four Neighbours
If Each and every direct pixel has an adjacent diagonal pixel or vice versa.
Then The Pixel under consideration cannot be considered as an Intersection
Case 3: Five Neighbours or more
If The Pixel under consideration has five or more neighbours.
Then It is always considered an Intersection

Once all the intersections are identified in each zone of character image, then they have been populated. In Minor Starters, the pixel under consideration has more than two neighbours. It can be found when the skeleton of character is traversed.

Character traversal and distinguishing line segments

Once the zoning phase has been finished, the skeleton of character image undergoes a traversal process whereby each zone is separately subjected to the process of line segments extraction. the novel algorithm was proposed by (Dileep, 2012). We used it in to extract features of our system, first by finding the starters list. Once all the starters are processed in order to obtain the line segments with the minor starters being processed at the same time, then the algorithm starts with minor starters and all the line segments that were obtained are populated and stored to be

labelled and processed later on. Once all the pixels of the character skeleton have been visited, the algorithm stops.

Now, once line segments have been extracted from the character image, they have to be classified into one of the following line segment types: Right Diagonal Line, Left Diagonal Line, Vertical Line, Horizontal Line. Each character pattern comprised of these four types of line segments that are mentioned above. After starters and intersections have been determined, the neighbouring pixels along the thinned character skeleton were followed from the starting points until we reach an intersection point.

Once we arrive at the intersection, the clockwise searching begins in order to determine the beginning and the end of the individual line segments.

The incipience of a new line segment is located **IF**:

1. The previous direction was up-right or down-left **AND** the next direction is down-right or up-left **OR**
2. The previous direction is down-right or up-left **AND** the next direction is up-right or down-left **OR**
3. The direction of a line segment has been changed in more than three types of direction **OR**
4. The length of the previous type is greater than three pixels.

These rules that illustrated above were reviewed for each set of pixels (that compose the segments) in each character sskeleton (pattern).

Estimation of feature vectors through zoning

Blumenstein et al. (2003) who suggested the feature extraction techniques mention above, have also developed a methodology for creating appropriate feature vectors in such a way that it is suitable and uniform in its size to be used as inputs to the back propagation artificial neural network (ANN). The first step of this methodology is to zone the character pattern that is marked with direction information into windows of equal size. Now, if the image matrix was not divisible in equal manner, then it was padded with additional background pixels along the length of both its rows, and its columns. The next step is to extract the direction information out of each individual window where the direction information includes: line segment direction, the intersection points, length, starter points and it is expressed in floating point values between (-1 and 1). The extraction and storing algorithm of a line segment information proceeds in the following steps.

In the first step is to locate the starting point and the intersection points in the window under consideration, in the second step is to extract the number and the length of line segments. These steps yield an input vector composed of nine floating-point values. Now, each value in this vector is defined as follows the number of right diagonal lines, the total length of right diagonal lines, the number of horizontal lines, the total length of horizontal lines, the number of vertical diagonal lines, the total of vertical diagonal lines, the number of left diagonal lines, the total length of left diagonal lines and number of intersection points.

Regional Features and Euler number of an image

In addition to the directional features that extracted above, we have extracted the regional features of an image, which include Euler number. The Euler number is one of the most important regional features of an image that is used to describe the topological structure of a binary image. The Euler number is expressed mathematically as explained in Rosenfeld and Kak (1982) as shown in Equation 1:

$$E = N - H \quad (1)$$

where N is the number of regions of the image or in other words, it represents the number of connected components of an object. H represents the number of holes in the image or the number of isolated regions of the image's background. In order to illustrate the Euler number implemented in our proposed system, we apply Equation 1 to two Arabic characters**: *Saad* (ص), *Faa* (ف)**. Figures 10 and 11 illustrate the application of Euler number.

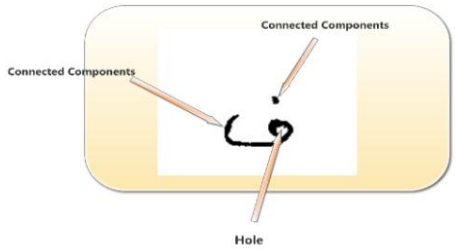

Figure 10 Number of regions and holes in the ***Faa*** (ف) Character

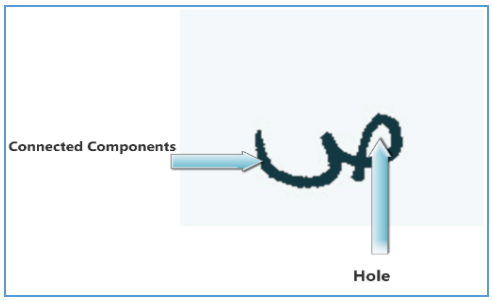

Figure 11 Number of regions and holes in the ***Saad*** (ص) Character

3. RESULTS AND DISCUSSION

Multi-layer back propagation ANN is considered the classification engine of our proposed system, which consists of two major dependent phases: Training and Testing. In the Training phase, the optimal weights, network parameters and network structure that will be fed as inputs to Testing phase are defined. In the training stage of our back propagation multi-layer neural network handwritten recognition system, we have achieved a mean square error (MSE) of (0.0209) after 1000 epochs with an overall recognition rate of (93.72%).The number of hidden neurons of the back propagation neural network has been chosen to be high enough to model our problem at hand, but at the same time not too high to avoid over-fitting. Therefore, the hidden layers and the neurons (nodes) of each one has been chosen to have optimal performance in the testing phase. Figure 12 illustrates the methodology in which each neuron in multi-layer perceptron processes the coming information.

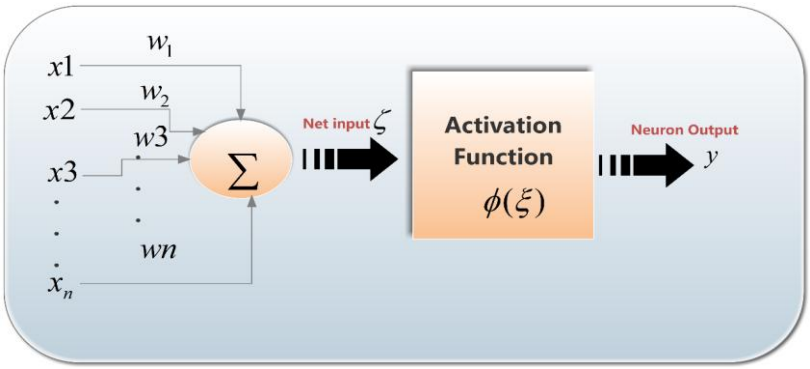

Figure 12 Information processing by the i^{th} neuron of the j^{th} layer

After training phase has been performed and the optimal weights of the Back Propagation Artificial Neural Network have been obtained, then these optimal weights are fed into testing phase of our proposed system. Our proposed has achieved low MSE of 0.00917 associated with overall recognition accuracy, that reached up to 96.14% and 100 % for many characters. The recognition rates achieved by our proposed system for each character are presented in Table 1.

Table 1 Achieved recognition rates for Arabic characters in testing phase

Character	Recognition Percentage	Character	Recognition Percentage
أ(Alif)	100%	ظ(Tha)	93.55%
ب(Baa)	96.72 %	ع(Ayn)	93.44%
ت(Taa)	96.77%	غ(Ghayn)	90.16%
ث(Thaa)	93.55%	ف(Faa)	93.44%
ج(Jeem)	93.85%	ق(Gaaf)	100
ح(Haa)	96.72%	ك(Kaaf)	96.92%
خ(Kha)	90.48%	ل(Laam)	100%
د(Daal)	100 %	م(Meem)	98.46%
ذ(Thaal)	92.19%	ن(Noon)	100%
ر(Raa)	87.30%	هـ(Ha)	100%
ز(Zaay)	100%	ه(Ha2)	95.08%
س(Seen)	93.75%	و(Waaw)	92.19%
ش(Sheen)	88.33%	ؤ(Waaw2)	100%
ص(Saad)	100%	ء(Hamza)	96.83%
ض(Daad)	98.36%	ي(Yaa)	100%
ط(Ta)	100%	ئ(Yaa2)	98.39%

As shown in Table 1 the achieved results are very promising where the achieved overall accuracy reached up to (96.14 %) for all letters, although we were even more successful when the (recognition) rate reached up to (100%) for some letters, as highlighted in the same table. Table 3 and Table 4 are dedicated to compare our experimental results with other existing optical character recognition techniques and systems, we first compare the performance of the proposed system with other systems that built on the same data sets, namely, CENPARMI. Then we compare our experimental results with that achieved by other systems that used different databases for Off-line Arabic handwritten isolated characters and built on different classification algorithms and techniques.

Table 2 Comparison between previous studies that built on the same database

Previous study	Approach	Accuracy
Jamal et al. (2014)	Support vector machine	90.88%
Sahlol et al. (2014a)	Support vector machine	89.2%
Sahlol et al. (2014b)	Feed forward neural network	88%
Proposed method	Feed forward neural network	96.14%

A deep insight in the experimental results in Table 2 will reveal that our proposed system can be efficiently utilized as a major part of the system that was proposed by Jamal et al. (2014) to recognize end-word classes (Isolated). This will enhance the overall handwritten word recognition accuracy of their system. This technique utilized the Arabic writing characteristics in the process of Arabic text segmentation. This shape analysis-based technique has two main stages: (1) metric-based segmentation; and (2) recognition-based segmentation or End-Shape Letter ESL-based segmentation. In the metric-based segmentation phase, the distance between adjacent components was evaluated utilizing a gap metric that will calculate the mean gap between text words based on an estimated threshold. This stage segmented the text line into its words. The ESL-based segmentation will specify the word segment. ESL can come into two forms: Isolated or Part-of-Word (PAW). This stage begins by end-shape recognition where the isolated letter or the last letter of a PAW will be identified and then the last part of the word will be extracted based on width, heights and position of the baseline. Sahlol et al. (2014a, 2014b) used the same classifier and database (CENPARMI) that we used during our proposed system.

This approach was built based on two principle building blocks: Novel prepressing operations include a variety of noise removal and dilation techniques, and structural, statistical and topological features that are extracted out of the main and secondary components of each Arabic character. Structural features include the upper and lower profiles that used to capture the outline shape of the connected portion of the character, which is composed of the horizontal and vertical projection profiles too. The statistical features are those features that result from discrimination between the individual foreground pixels and the set of all foreground pixels, which yields the connected components of the character. Finally, the topological features that are composed of the end- points, pixel ratio and height to width ratio features of the character. After examining the recognition accuracy for each character, we found that the recognition rate is between 100% for the some characters, such as Daal(د), Zaay(ز), Saad(ص), and 90.16% and 93.75% for Ghayn(غ), and Seen(س), respectively. Contrary to what was achieved by Sahlol et al. (2014a, 2014b), where they received accuracy rates of 61% and 66% for Gaaf(ق) and Ayn(ع) respectively, which is considered below the results that we achieved for the same characters. however, we have achieved better results either in terms of overall recognition accuracy or in terms of accuracy rate for each character.

Table 3 Comparison between previous studies that were built on a different database

Previous study	Approach	Accuracy
Bahashwan & Abu-Bakar (2015)	Feed forward neural network	90.03%
Alkhateeb (2015a)	Feed forward neural network	87.75%
Alkhateeb (2015b)	Hidden Markov Model	88.25%
Hammad & Elhafiz (2015)	Feed forward neural network	88.11%
Amrouch et al. (2008)	Hidden Markov Model	85.71%
Proposed method	Feed forward neural network	96.14%

From the previous Table 3, our proposed system shows its superiority in comparison with other OCR systems such as Hammad and Elhafiz (2015). The proposed system was realized based on an Off-line isolated character database that was collected and processed by the recognition group of Sudan University for Science and Technology (SUST ARG) and a neural network was used to classify the characters in each set. The results were due to highly effective feature extraction techniques that have been employed in our proposed system.

4. CONCLUSION

This paper presents an Off-line Arabic Handwritten isolated recognition system based on novel feature extraction techniques and powerful machine learning algorithm: Back Propagation artificial neural network. It proves that the proposed system provides better recognition accuracy rather than that achieved by other approaches. The proposed system achieved (96.14%) overall recognition accuracy and a recognition rate of (100%) for some letters, which considered to be promising results in the field of Arabic handwritten recognition. Zoning techniques into horizontal, vertical and 3x3 square zones enabled us to exploit the high capabilities of the feature extraction techniques in optimal way where the zoning techniques give the tiny details of the character curve and overcome the low recognition accuracy problem of other currently available Arabic handwritten optical isolated character recognition systems. Depending on the high performance that has been achieved in this paper, we recommend using other classifiers in association with the novel feature extraction techniques that have been used in this paper, such as K-nearest neighbors algorithm (KNN), Fuzzy Logic (FL), Dynamic Bayesian Networks (DBN) and Self-Organized Map (SOM) network or other types of artificial neural networks. The recognition accuracy of this research is highly dependable on many factors related to the used database, namely, CENPARMI in our case. These factors include the

way in which the letters are written by the writers who were involved in the process of creating this database, such as if the letters are poorly written or drawn in unusual style or using the different writing style such as AL-Rukaa style, and since CENBARMI contains most of those writing defects, we expected the low accuracy of overall performance.

5. REFERENCES

Abuzaraida, M.A., Zeki, A.M., Zeki, A.M, 2013. Recognition Techniques for Online Arabic Handwriting Recognition Systems. IEEE, *In*: 2012 International Conference on Advanced Computer Science Applications and Technologies *(ACSAT)*, 26-28 November, 2012, pp. 518–523

Alamri, H., Sadri, J., Suen, C.Y., Nobile, N., 2008. A Novel Comprehensive Database for Arabic off-line Handwriting Recognition. *In:* Proceedings of 11th International Conference on Frontiers in Handwriting Recognition, ICFHR, Volume 8, pp. 664–669

Alijla, B., Kwaik, K., 2012. OIAHCR: Online Isolated Arabic Handwritten Character Recognition using Neural Network. *The International Arab Journal of Information Technology*, Volume 9(4), pp. 343–351

Alkhateeb, J.H., 2015a. Off-Line Arabic Handwritten Isolated Character Recognition. *International Journal of Engineering Science and Technology* (*IJEST*), Volume 7(7), pp. 251–257

AlKhateeb, J.H., 2015b. A Database for Arabic Handwritten Character Recognition. *Procedia Computer Science*, Volume 65, pp. 556–56

Amrouch, M., Elyassa, M., Rachidi, A., Mammass, D., 2008. Off-line Arabic Handwritten Characters Recognition based on a Hidden Markov models. *In*: Image and Signal, Springer Berlin Heidelberg, pp. 447–454

Asebriy, Z., Bencharef, O., Raghay, S., Chihab, Y., 2014. Comparative Systems of Handwriting Arabic Character Recognition. IEEE, *In:* Complex Systems (WCCS), Second World Conference, pp. 90–93

Bahashwan, M.A., Abu-Bakar, S.A.R., 2015. Off-line Handwritten Arabic Character Recognition using Features Extracted from Curvelet and Spatial Domains. *Research Journal of Applied Sciences, Engineering and Technology*, Volume 11(2), pp. 158–164

Blumenstein, M., Verma, B., Basli, H., 2003. A Novel Feature Extraction Technique for the Recognition of Segmented Handwritten Characters. IEEE, *In:* Document Analysis and Recognition, Proceedings, Seventh International Conference, pp. 137–141

Dileep, D., 2012. *A Feature Extraction Technique based on Character Geometry for Character Recognition*. Cornell University Library, Volumes 1–4

Hammad, N.H., Elhafiz, M.A., 2015. Divide and Conquer Method for Arabic Character Recognition. *IOSR Journal of Engineering (IOSRJEN)*, Volume 5(4), pp. 36–41

Jamal, A.T., Nobile, N., Suen, C.Y., 2014. End-shape Recognition for Arabic Handwritten Text Segmentation. *In:* IAPR Workshop on Artificial Neural Networks in Pattern Recognition, Springer International Publishing, pp. 228–239

Lawgali, A., 2015. A Survey on Arabic Character Recognition. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, Volume 8(2), pp. 401–426

Rashad, M., Semary, N.A, 2014. Isolated Printed Arabic Character Recognition using KNN and Random Forest Tree Classifiers. *In:* International Conference on Advanced Machine Learning Technologies and Applications, Springer International Publishing, pp. 11–17

Rosenfeld, A., Kak, A.C., 1982. *Digital Picture Processing*. Academic Press Inc., New York

Sahlol, A.T., Suen, C.Y., Elbasyouni, M.R., Sallam, A.A, 2014b. A Proposed OCR Algorithm for the Recognition of Handwritten Arabic Characters. *Journal of Pattern Recognition and Intelligent Systems*, Volume 2(1), pp. 90–104

Sahlol, A.T., Suen, C.Y., Elbasyouni, M.R., Sallam, A.A, 2014a. Investigating of Preprocessing Techniques and Novel Features in Recognition of Handwritten Arabic Characters. *In*: Artificial Neural Networks in Pattern Recognition, Springer International Publishing, pp. 264–276