A feature extraction method for Arabic Offline Handwritten Recognition System using Naïve Bayes classifier

Ahmed Subhi Abdalkafor
Khattab M. Ali Alheeti
Laith Al-Jobouri

© 2021 IEEE
Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

A Feature Extraction Method for Arabic Offline Handwritten Recognition System Using Naïve Bayes Classifier

Abstract— **Handwriting recognition in the Arabic language is considered one of the most challenging problems and the accuracies in recognizing still need more enhancements due to the Arabic character's nature, cursive writing, style, and size of writing in contrast to working with other languages. In this paper, we propose a system for Arabic Offline Handwritten Character Recognition based on Naïve Bayes classifier (NB). Extraction features preceded by divided the image of character into three horizontal and vertical zones and 3x3 zones in one and two dimensions respectively, then classified by Naïve Bayes. The performance of the system proposes evaluated by using the benchmark CENPARMI database reached up to 97.05% accuracy rate. Experimental results confirm a high enhancement inaccuracy rate in comparison with other Arabic Optical Character Recognition systems.**

Keywords— Arabic Handwritten Character, Starters and Intersections, Minor-Starters, Naïve Bayes Classifier

I. INTRODUCTION

The Arabic language is one of the major languages in the world, but its complex nature has a little interest by researchers, unlike other world languages. [1] [2]. In this context, facing more problems, and challenges for the handwriting recognition systems in the meeting of its difficult work, there is no static representation for the characters, due to the persons themselves in the way how they write these characters and the styles that they adopt in their writing, which they get it by years of experience [3] [4]. Recently, recognition has become a mature discipline due to its use in sensitive and significant applications [5][6], comparing with other languages, the recognition of Arabic characters still the hardest application, due to the nature of both cursive and unconstrained. In addition, there are no rules to control the way, style, and size of writing and in some cases, there is the structure of a very similar character very similar characters structure which in turn makes the recognition a complex issue, due to the huge styles that the humans have in writing their characters, Arabic Off-line handwriting character recognition faces great challenges [7] [8] [9] Starting from the importance of the characters recognizing process which needs an accurate method to ensure that the presented solution would provide a near-optimal recognition process.

Some studies that used different classifiers in the processing phase of Arabic and non-Arabic optical character recognition systems.

[10] Presented a model using a back-propagation ANN to recognize Arabic handwritten character recognition. In this system, we used the Otsu method to convert the gray-level images to binary ones and applied the Universe of discourse and Skeletonization. The features are extracted depending on the zoning method by dividing the image of character into horizontal, vertical zones, and 3x3 zones in one and two dimensions respectively, a result using (CENPARMI) database is the highest among other related works.

A Survey of Arabic handwritten Character Recognition system by Encoded Freeman Chain Code offered by [3], It focuses on survey OCR in handwritten Arabic language and describes approaches for recognizing handwritten isolated Arabic characters using Freeman- chain code. The last stage of this proposed system is to generate the encoded chain code by taking only the first 7 digits and adding additional digits at the beginning of the new chain code to classify each character. The results of this method achieved a high accuracy recognition range from 92% to 97%. The author Schol has been working in the field of Arabic handwritten character recognition represented by three works published in [11] [12] [13], where they built AOCR systems based on the CENPARMI database. The first system is based on the Bio-inspired BAT Algorithm that is implemented to decrease the feature set size and to increase the rate of accuracy. A system has been tested by four classifiers; artificial neural network ANN, KNN, Random Forest (RF), and Bayes Network (BN). The second and also the third systems were built based on two blocks: pre-processing methodology and statistical, structural, and features of topological that extracted from the main and secondary characters, this system used ANN, KNN, SVM classifiers in two and three systems respectively. Recently some studies focused on creating Arabic Offline Handwritten databases for evaluating and matching results [14] [15] [16].

II. METHODOLOGY

In this study, an efficient optical character recognition system is presented for Arabic Off-line handwritten character recognition that depends on four essential mechanisms as basic building blocks: Preprocessing, Feature extraction, and Recognition mechanisms. Firstly, the Preprocessing operations are performed on character images to be ready for the feature extraction phase. Then, extracted features vectors of character image by several techniques. This metric can be considered as a strong indicator of a correctly classification rate via the Naïve Bayes classifier. Figure 1 illustrates the block diagram of our proposed system.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

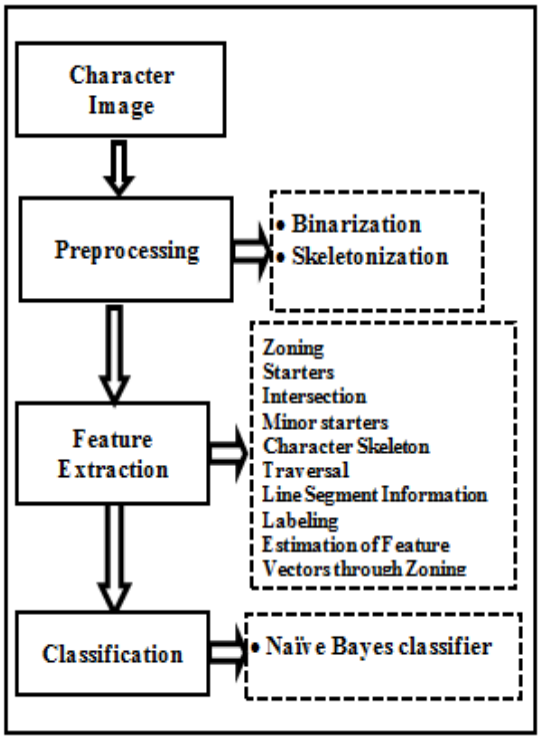


Fig. 1. Block Diagram

A. Image Pre-processing

This stage consists of two processes: (i) Binarization: Converted the grey images to binary, in this work, we applied the Otsu technique [17] but after applying this technique that uses the default value (0.5) we lost the useful information of the images, so we applied the MATLAB software program to selected the intensity scale and calculated by the Otsu technique to replace all pixels of the input image with value one (White) and other pixels with value zero (Black); (ii) Skeletonization: which considered as part of the morphological operations, it operates by removing the pixels on the object boundaries (character) without allowing the object to break apart. The rest of the pixels make up the picture skeleton. Figure 2 shows the skeletonization process of the Kaaf (ك) character.

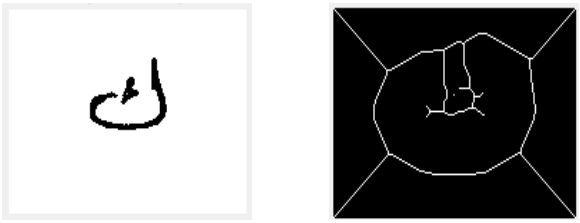


Fig .2 Skeleton of Kaaf (ك) character

B. Features Extraction

The main core of this stage is to extract the important features to distinguish between the scanned of unknown images accurately and efficiently. We have utilized three zoning methods. Two of which are employed by [18] to recognize the English letters, this study divided the image of character into 3 horizontal, vertical, and 3x3 zones then extracted the components of each zone separately and concatenated those features in one vector, so we make sure that all fine details of the skeleton characters have been chosen which helps a Naive Bayes Classifier in the next phase. Figure 3 shows zoning along one and two dimensions of the Kaaf(ك) character image respectively.

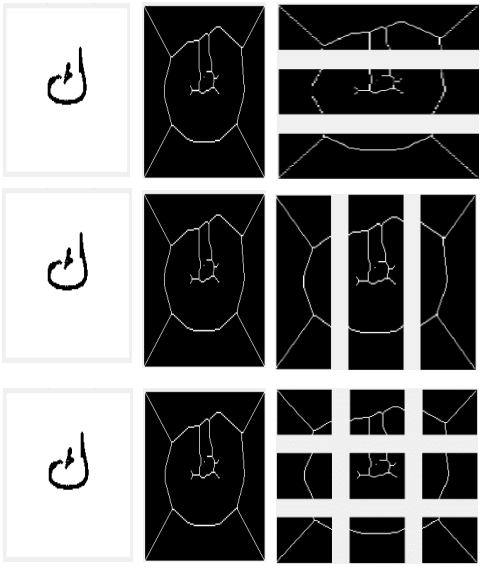
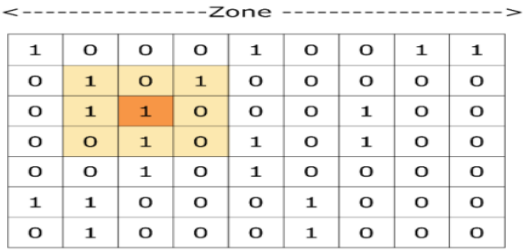


Fig. 3 Zoning Along One and Two Dimensions

To distinguish the line segments that will create the features of a dataset in the proposed system, the specified pixels are determined the starters, intersections, minor starters, and the whole skeleton character then traversed pixel by pixel.

1)Starters and Intersections

The pixel that has one neighbor is defined as a starter, where the pixel under consideration contains all pixels that immediately surrounded it which is called the neighborhood. Figure 4 elaborates both of these concepts, where the pixel under consideration is colored in a darker color rather than that of neighbors around it. The neighborhood can come in direct or diagonal directions, which means that each pixel has eight neighbor pixels: four of them in diagonal, two vertical, and the other horizontal. Figure 5 shows the starters and intersection points of Ta (ط) character.



1	0	0	0	1	0	0	1	1
0	1	0	1	0	0	0	0	0
0	1	1	0	0	0	1	0	0
0	0	1	0	1	0	1	0	0
0	0	1	0	1	0	0	0	0
1	1	0	0	0	1	0	0	0
0	1	0	0	0	1	0	0	0

Fig 4 . Direct and Diagonal Neighbourhood of a Pixel

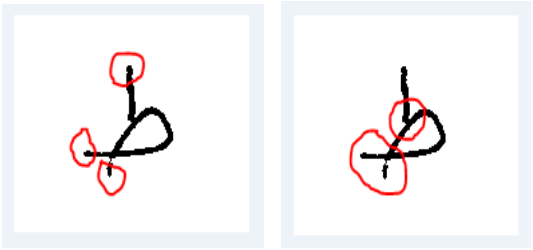


Fig 5. Starters and Intersection of Ta (ط) character

2) Minor-Starters

The pixel that has more than two neighbors has defined Minor starters. If Skeleton's character is traversed then a Minor starter could be found. Figure 6 illustrates this concept.

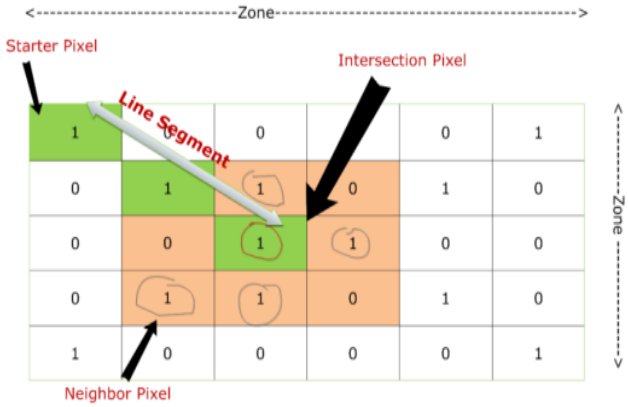


Fig 6. Minor Starters of a Particular Pixel

3) Character Skeleton Traversal

After zoning phase completion, a traversal process is subjected to the character image's skeleton and the line segments' extraction procedure is separately presented. The starters and intersections first will be detected then established in arrays. In addition, skeleton traversal and minor starters are established together at the same time. Starting by detecting the list of the starters, the proposed system's features are extracted by the novel algorithm proposed by [18]. After processing the starters, the line segments and minor starters are obtained simultaneously then established and memorized to be classified and handled. The algorithm will finish after visiting all the character skeleton pixels.

4) Line Segment Information Labeling

The encoding process of white pixels that composed of each segment will take place in the following directional manner. After identifying the starters and intersections of each segment, the segments of the zone should be labeled, which means the directions pixels that compose each segment are identified in one of eight directions: Up, left, right, down, up-left, up-right, down- left and down-right. Where the directions take the following values illustrated in Figure 7. Then the segment takes the label of the most frequent direction. As an example, suppose that we have the following zone shown in Figure 8.

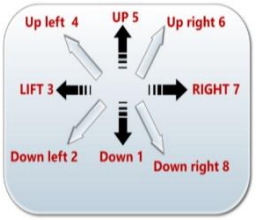


Fig 7. Pixel direction

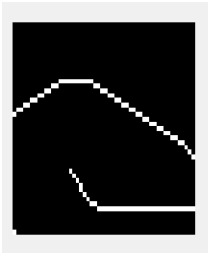


Fig. 8 Identified Segments of the given zone

We can identify two segments in this zone, as illustrated in Figure 9:

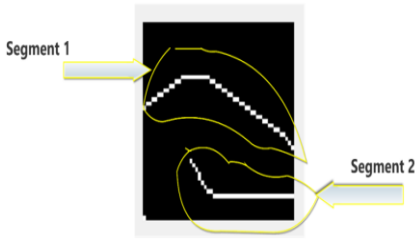
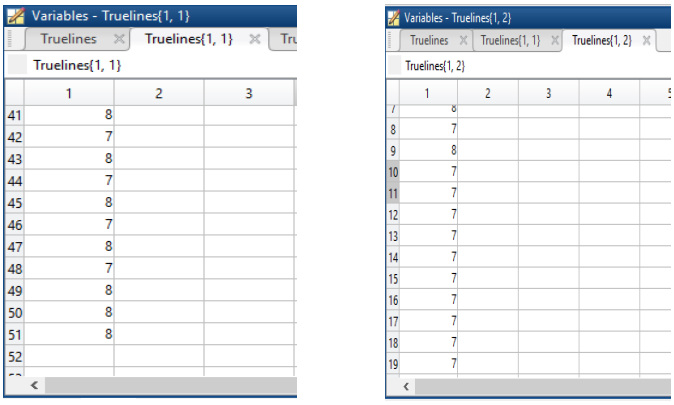


Fig. 9 Identified Segments of the given zone

However, these segments contain spurious pixels that are composed in another direction, thus, we identify the direction of each pixel, and the most frequent direction is given as a label for that segment as illustrated in Figure 10. As we note, the direction (8) which represents downright is the most frequent in segment 1, whereas the pixel direction of (7) (which represents the right direction) is the most frequent in segment 2 as illustrated in Figure. 10 (B). Therefore, we can conclude that segment (1) has the label of (8) and segment (2) take the label of (7), which are numerical values that will be processed to be suitable as inputs for the Naïve Bayes classifier.



A B

Fig. 10 Pixels' directions of segment 1 and segment 2

5) *Estimation of Feature Vectors through Zoning*

A methodology for compatible feature vector creating is improved by [19] who proposed the mentioned feature extraction techniques to be used as compatible and uniform size inputs Naïve Bayes classifier. The zoning process of the character pattern is the first step that is marked with direction information into windows of equal size. First, we check the matrix of the image, it will be padded floating-point values

between (-1 and 1) with added background pixels along the length of columns and rows if not divisible in an equal manner then extracted the direction information that consists of line segment direction, line segment direction, intersection points, starter points, length and it expressed in for each window. The extraction and storing algorithm of the line segment information is completed by finding the starting and intersection points then extracted the length and number of the line segments which create 9 floating-point values as input vectors: The number and total length of right and left diagonal lines, The number and total length of horizontal and vertical diagonal lines and the number of intersection points.

III. CLASSIFICATION

This classification phase is built on the Bayes theorem proposed by Thomas Bayes. This theorem is used to calculate probabilities of a particular problem hypothesis explicitly and at the same time, it does not affect by the noise associated with input data. The Bayesian classifier has a high capability to minimize misclassification probability[20].

A. Naïve Bayes

Naïve Bayes classifier can be described as an independent feature model due to the underlying probability theory that forms the basis for this type of classifier. Simply, a naïve Bayes classifier depends heavily on an essential assumption that the presence of a specific feature of a particular class is not concerning the presence or absence of any other features [21].

A) Naïve Bayes Probabilistic Model

In this type of classifier, the probability model is the conditional model that can be mathematically expressed in probability Eq. (1) [22].

$$Pro(Cla|Fea_1, Fea_2, \dots\dots\dots\dots\dots, Fea_n) \quad (1)$$

where over a dependent class variable Cla with a small set of outcomes (classes) that are conditional on many feature variables represented by vector Eq. (2):

$$(Fea_1, Fea_2, \dots\dots\dots\dots\dots, Fea_n) . \quad (2)$$

Now, using Bayes Theorem will yield in Eq. (3):

$$Pro(Cla|Fea_1, Fea_2, \dots\dots\dots\dots\dots, Fea_n) = \frac{Pro(Cla)Pro(Fea_1, Fea_2, \dots\dots\dots\dots, Fea_n|Cla)}{Pro(Fea_1, Fea_2, \dots\dots\dots\dots, Fea_n)} \quad (3)$$

The mathematical formula in Eq. (3) can be plainly understood as in Eq. (4):

$$Posterior = \frac{Prior\ x\ Likelehood}{Evidence} \quad (4)$$

Practically speaking, the numerator is equal to the joint probability that is given by function probability in Eq. (5):

$$Pro(Cla, Fea_1, Fea_2, \dots\dots\dots\dots\dots, Fea_n) \quad (5)$$

By repeated application of the basic definition of conditional probability, Probability Eq. (5) can be re-written as in Eq. (6):

$$= Pro(Cla)Pro(Fea_1, Fea_2, \dots\dots\dots\dots\dots, Fea_n|Cla)$$
$$Pro(Fea_{4,\dots,}Fea_n|Cla, Fea_1, Fea_2, Fea_3) = Pro(Cla)Pro(Fla_1|Cla)Pro(Fea_{2,\dots,}Fea_n|Cla, Fea_1) Pro(Fea_{3,\dots,}Fea_n|Cla, Fea_1, Fea_2) Prp(Cla)$$
$$Pro(Fea_1|Cla)Pro(Fea_2|Cla, Fea_1)$$
$$Pro(Fea_3|Cla, Fea_1, Fea_2)$$
$$Pro(Fea_{4,\dots,}Fea_n|Cla, Fea_1, Fea_2, Fea_3)$$
$$= Pro(Cla)Pro(Fea_1|Cla)Pro(Fea_2|Cla, Fea_1)$$
$$Pro(Fea_3|Cla, Fea_1, Fea_2) \dots$$
$$Pro(Fea_n|Cla, Fea_1, Fea_2, Fea_3, \dots, Fea_{n-1}) \quad (6)$$

Now, the assumption of "Naïve "conditional independence plays its role as follows:

Assuming that each feature vector($\boldsymbol{Fea_i}$) is conditionally independent of every other feature vector ($\boldsymbol{Fea_j}$) in the dataset such that $\boldsymbol{j \neq i}$, then you get Eq. (7):

$$Pro(Fea_i|Cla, Fea_j) = Pro(Fea_i|Cla) \quad (7)$$

For each $\boldsymbol{j \neq i}$, therefore the joint probability model can be expressed as follows Eq. (8):

$$Pro(Cla, Fea_1, \dots\dots, Fea_n) = Pro(Cla)Pro(Fea_1|Cla)Pro(Fea_2|Cla)Pro(Fea_3|Cla) \quad (8)$$

Under the assumptions in Eq. (8) above, the conditional distribution over the class variable $\boldsymbol{Cla}$ can be expressed mathematically as in Eq. (9):

$$Pro(Cla|Fea_1, \dots\dots, Fea_n) = \frac{1}{Z} Pro(Cla) \prod_{i=1}^{n} Pro(Fea_i|Cla) \quad (9)$$

where Z represents the evidence that we talked about in the explanation of the Bayes theorem in terms of simple English language, which is a scaling factor that depends only on $\boldsymbol{Fea_1, \dots\dots, Fea_n}$, thus, it becomes a constant if the values of these features variable are known.

The Equations above represent a manageable model of a certain problem, where we can factor it into a so-called class prior $\boldsymbol{Pro(Cla)}$ and independent probability distribution $\boldsymbol{Pro(Fea_i|Cla)}$. Then the Classifiers function as in Eq. (10):

$$classify\ (fea_i, \dots, fea_n) = argmax\ Pro(Cla) = (\prod_{i=1}^{n} Pro(Fea_i = fea_i|(Cla = cla) \quad (10)$$

IV. RESULTS AND DISCUSSION

This section presents the discussion of results obtained from training and testing the proposed Offline Arabic handwritten recognition system using the CENPARMI database [23] which contains 21426 Arabic characters handwritten and divided into 32 sets of characters as Table 1. Naïve Bayes classifier has been built for the sake of performance enhancement. Table 1 illustrates the accuracy rate achieved for each character.

Table 1. Accuracy Rates For Arabic Characters.

Character	Accuracy Rate
أ(Alif)	100%
ب (Baa)	100 %
ت (Taa)	90.434%
ث(Thaa)	94.063%
ج(Jeem)	88.793%
ح(Haa)	100%
خ (Kha)	100%
د (Daal)	100%
ذ(Thaal)	100%
ر(Raa)	100%
ز(Zaay)	100%
س (Seen)	100%
ش(Sheen)	92.086%
ص(Saad)	100%
ض(Daad)	100%
ط(Ta)	100%
ظ (Tha)	100%
ع(Ayn)	100%
غ(Ghayn)	100%
ف(Faa)	100%
ق(Gaaf)	100%
ك (Kaaf)	100%
ل (Laam)	93.264%
م(Meem)	100%
ن(Noon)	100%
هـ(Ha)	88.695%
ه(Ha2)	94.495%
و(Waaw)	89.655%
ؤ(Waaw2)	100%
ء(Hamza)	100%
ي(Yaa)	100%
ئ(Yaa2)	88.695%

As shown in the above table, the results are very promising which investigated the total accuracy rate that reached up to 97.05% for all characters although we get success (Accuracy) rate reached up to (100%) for almost all characters. Despite the computational complexity of this system, it is suitable for real-time applications because the run time is acceptable where (0.376214) seconds, which means (0.02144) second for each character. From recent studies in the field of pattern recognition, a few researchers used the Naïve Bayes classifier to classify Arabic Offline handwritten character and comparison with classifiers such as ANN, SVM [13] [25] [26], etc. So, we will compare the obtained results with various classification algorithms and techniques in the same database and others because of the highly effective feature extraction methods applied in this work. Table 2 is dedicated to comparing our results with the results of other systems. These results are arranged according to the descend accuracy rate. Table 2 shows a brief of the most recent works in Arabic handwritten character recognition. As seen, they are organized according to years, database, the approach of the classification, and accuracy rate.

. **Table 2.** Comparison with Other Systems That Built On Same And Different Database.

Character	Year	Datasets	Approach	Rate
Althobaiti et al.[3]	2017	own dataset	Encoded Freeman Chain Code	Ranges from 92% to 97%
Abdalkafor (**Our previous work**) [8]	2017	CENPARMI	Feedforward neural network	96.14%
Torki et al.[27]	2014	AIA9K	SVM with RBF kernel	94.28%
Gumah et al. [28]	2010	28 images collected from 48 different writers	Fast Wavelets Transform (FWT)	94.18%
Schol et al. [9]	2016	CENPARMI	Bio-inspired BAT Optimization Algorithm	91.59%
Jamal et al [26].	2015	CENPARMI	SVM	90.88%
Schol, et al.[12]	2014	CENPARMI	SVM	89%
Sahlol, et al.[13]	2014	CENPARMI	Feedforward neural network	88%
Alkhateeb [29]	2015	Total 28000 images written by 100 different writers 10 for each character.	Feedforward neural network	87.75%
Our Method	2019	CENPARMI	Naïve Bayes classifier	**97.05**

From Table 2, various techniques have been proposed for Arabic characters [10] [12] [13] and [26] have been used in the same database, namely, CENPARMI, with various classifiers and achieved a high accuracy rate ranged from 96.14% to 88%. Our results have superiority and considered a strong indicator of our feature extraction techniques associated with Naïve Bayes as a classification technique proved recognition 97.05 % of characters correctly. The reason for this highest result is to implement the method proposed by Dileep for Arabic characters and add an extra way that divides the image into three vertical zones to guarantee examination of all details in the image entering into the proposed system. This result motivates us to employ the features extraction techniques with other shapes of languages character or to integrate our proposed into handwritten Arabic text recognition systems to improve further the quality of recognition systems.

V. Conclusions and Future Work

In this paper, we have introduced a method for off-line Arabic handwriting character recognition. In our work, all the tiny details of the character image curve are taking into account by dividing it into three horizontal, vertical, and 3x3 zones that prove the high capabilities of the feature extraction techniques optimally and enabled us to improve the low accuracy rate problem. The accuracy rate obtained considered promising results in an off-line Arabic handwritten character

recognition field. By further investigation, since the handwritten Indian and Arabic digits are much easier than the isolated handwritten Arabic characters, we recommend using our proposed system to recognize it and we expect higher performance (fully recognized). In addition, we can use our proposed system for isolated printed Arabic character recognition and we expect also higher recognition accuracy due to its smoothest forms of printed characters. Also, we intend to use the proposed system to recognize other types of language such as Japanese, Chinees, Persian, and Urdu.

REFERENCES

[1] K., Jayech, Mahjoub, M.A., Essoukri, N. & Amara, B,. Jayech, K., Mahjoub, M. A., & Amara, N. E. B, "Synchronous Multi-Stream Hidden Markov Model for offline Arabic handwriting recognition without explicit segmentation" . Neurocomputing,. 214, pp. 958-971, 2016.

[2] R., Sarkhel, Das, N., Das, A., Kundu, M., &Nasipuri, M, "A multi-scale deep quad tree based featureextraction method for the recognition of isolated handwritten characters of popular Indic script"'. Pattern Recognition,. 71, PP. 78-93, 2017.

[3] H., Althobaiti, & Lu, C, "A survey on Arabic optical character recognition and an isolated handwritten Arabic character recognition algorithm usingencoded freeman chain code ". In Information Sciences and Systems (CISS),51st Annual Conference on, pp. 1-6.2017.

[4] A.S Abdalkafor, "Survey for Databases On Arabic Off-line Handwritten Characters Recognition System" . In 2018 1st International Conference on Computer Applications & Information Security (ICCAIS).pp.1-6. 2018.

[5] M. Z., Khedher, & Abandah, G. "Arabic character recognition using approximate stroke sequence". In Proc. Workshop Arabic Language Resources and Evaluation: Status and Prospects and 3rd Int'l Conf.,2002.

[6] A.S,"Abdalkafor, DFRS-database for fingerprint recognition system using Ink-On-Paper technique, Journal of Engineering and Applied Sciences, 13(17), pp. 7401–7407, 2018.

[7] J., Al Abodi, & Li, X, "An effective approach to offline Arabic handwriting recognition". Computers & Electrical Engineering, 40(6), pp.1883-1901, 2014.

[8] A, JihadA. Abdalkafor, A. S,"A Framework for Sentiment Analysis in Arabic Text", Indonesian Journal of Electrical Engineering and Computer Science, Vol 16, No 3.2019.

[9] A. S., Abdalkafor, &Sadeq, A,"Arabic Offline Handwritten Isolated Character Recognition System Using Neural Network". International Journal of Business and ICT, 2(3), pp. 41-50, 2016.

[10] A. S, Abdalkafor, "Designing Offline Arabic Handwritten Isolated Character Recognition System using Artificial Neural Network Approach". International Journal of Technology, 8(3), pp. 528-538, 2017.

[11] A. T Sahlol,., Suen, C. Y., Zawbaa, H. M., Hassanien, A. E., &Elfattah, M. A, 2016. Bio-inspired bat optimization algorithm for handwritten arabic characters recognition. In Evolutionary Computation (CEC), 2016 IEEE Congress on(pp. 1749-1756). IEEE.

[12] A. T Sahlol,., Suen, C. Y., Elbasyoni, M. R., &Sallam, A. A, "Investigating of preprocessingtechniques and novel features in recognition of handwritten Arabic characters. In IAPR Workshop on Artificial Neural Networks in Pattern Recognition, (pp. 264-276), 2014.

[13] A. T Sahlol,., Suen, C. Y., Elbasyouni, M. R., &Sallam, A. A, "A proposed OCR algorithm for the recognition of handwritten Arabic characters". J. Pattern Recognit. Intell. Syst, 2, pp. 8-22, 2014.

[14] A. S., Abdalkafor, T. A. Emad, W. A. Khalid and N. M. Aiman. 2019. A Novel Database for Arabic Handwritten Recognition (NDAHR) System, In 2019 2st International Conference on Computer Applications & Information Security (ICCAIS), 2019.

[15] N. Lamghari and S. Raphay. "DBAHCL: database for Arabic handwritten characters and ligatures",International Journal of Multimedia Information Retrieval, vol. 6, pp. 263-269, 2017.

[16] A. S., Abdalkafor, Awad W. K, Alheeti, K. M.A.. Novel Comprehensive Database For Arabic And English Off-Line Handwritten Digits Recognition, Journal Of Southwest Jiaotong University, in press, 2019.

[17] N, Otsu,. "A threshold selection method from gray-level histograms",IEEE Transactions on Systems, Man, and Cybernetics: 9(1), pp. 62–66, 1975.