International Journal of Engineering and Technologies
ISSN: 2297-623X, Vol. 14, pp 14-30
doi:10.56431/p-nb4392

Submitted: 2018-06-18
Accepted: 2018-08-10
Online: 2018-09-21

New Segmentation Method for Analytical Recognition of Arabic Handwriting Using a Neural-Markovian Method

Khaoula Fergani[a*], Abdelhak Bennia[b]

[a,b]Faculty of technology, Department of electronics, University of Constantine 1, Algeria

*E-mail: khaoula_1190@hotmail.com

Keywords: Recognition of Arabic handwriting, hidden Markov models, fast K-means, Arabic literal amounts, multi-layer perceptron.

Abstract. A new hybrid system of off-line analytical recognition of Arabic handwriting combining a neural network type multi-layer perceptron (MLP) and hidden Markov models (HMM) is presented. We propose a way to cooperate HMM and MLP neural network in a probabilistic architecture taking advantage of both tools dedicated to the recognition of Arabic literal amounts. This description is based on statistical and structural characteristics extraction of the significant character of the handwritten Arabic words, which can be used in the MLP classification module to estimate probabilities used as the observations to perform a recognition by the HMM. The originality of our approach is based on the segmentation into characters taking into account diacritics with the characters that match them. The experiments show the convergence of the global system, even with a random initialization of the neural network.

1. Introduction

The handwriting recognition is in the domain of the pattern recognition which is interested in the forms of characters. The purpose is to develop a system that is closest to human being in his ability to read and to make communication man-machine easier and more flexible. Applications of handwriting recognition have been increasing in registration of bank checks, automatic processing of administrative records, etc. Character recognition has been one of the most fascinating and challenging research areas in the field of image processing and pattern recognition in recent years. Due to the variability in writing style and sizes, recognition of handwritten scripts is even more challenging than printed scripts. In general, character recognition can be defined as the task of transforming text represented in spatial form of graphical marks into its symbolic representation [1]. Automatic recognition of writing consists of the creation of systems capable of recognizing handwritten or printed characters. Despite the great progress of automatic recognition of writing, this area remains very active given the great variability of handwriting. The majority of proposed solutions were tested on Latin writing and then applied as such for the recognition of printed Arabic script. These methods generally assume that characters can be isolated by a segmentation step. This segmentation step is possible in the case of a printed Latin text, but very difficult in the case of cursive or semi-cursive Arabic writing. Therefore, it is clear that off-line recognition of Arabic text is still an open issue. There is still urgent need for high speed recognition rate systems. The improvements in any stage of recognition system will lead to increase the global system efficiency. Therefore, more research is needed in all the recognition system stages especially the segmentation and the classification stages, since they are the most challenging tasks in the off-line Arabic handwritten recognition system.

Although Arabic handwriting is a cursive script, most of the research in this area handles isolated characters[1], some researchers published papers about Arabic character recognition [2]-[10], some about Arabic-Indian numerals [11] and some included both [12]. Different method approaches have been used. Most of these methods are based on neural network and hidden Markov model [13]-[15].

2. Related Work

The cursive nature of the Arabic script, the overlapping between the characters, different forms of each letter depending on its position in the word or the writing style and the presence of the secondary characters like dots, hamza and diacritics, are all factors that increase the difficulty of recognition. Osman [16] developed a segmentation algorithm for Arabic handwriting. The first step in the algorithm was to divide the selected image into lines and sub-words, then trace the sub-word contour. Finally, the algorithm detects the exact points where the contour changes its state from a horizontal to vertical or curved line and consider those point as a segmentation points. The algorithm achieved 89.4% segmentation accuracy on 537 tested words from the IFN/ENIT database. Alma'adeed et al. [17] and Gouda and Rashwan [18] propose a system of segmentation in handwritten Arabic text letters, the potential points of segmentation are at the level of the local minima of weak external contour. These points of segmentation are subject to a set of rules to validate them or reject them. For the validation of their segmentation tool, they use a character recognizer, but this idea remains with the state of project, which it does not seem be carried out thereafter. Boulid et al. [19] present an approach inspired by the perception mechanisms involved in human reading process to automatically extract text lines from Arabic handwritten documents. The proposed approach is based on multi-agent systems to detect and combine the components connected belonging to the same line. Samoud et al. [20] propose three criteria of evaluation for the comparison of two methods of segmentation for Arabic handwritten words. The first method of segmentation is based on a combination of projection and the minima and maxima of the out-line of the picture. The second method is a combination of Hough Transform and mathematical morphology. These methods are developed, evaluated and compared in reference to the IFN/ENIT database. Lawgali et al. [21] exploited the fact that segmentation points, which occur at the end of a character and the beginning of the next, are usually located in the region surrounding the baseline. The segmentation algorithm starts with segmenting the word into sub-words and then the baseline of each sub-word is computed. The vertical projection is used to find the candidate points for the segmentation. The algorithm has been tested using 800 handwritten Arabic words taken from IFN/ENIT database and has achieved 82.98% character accuracy. However, this algorithm could not segment the alphabets (س, ش) into three segments rather it only segmented them into one. Tamen and Drias [22] tried to overcome the over-segmentation problem in the segmentation stage by pasting the segmented parts to rebuild the whole character form after the rejection or the ambiguousness decision in the recognition stage. The training was done using the back propagation algorithm with all the pre-segmented Arabic characters and their different positions written by three different persons.

The field of recognition of handwritten Arabic words is a broad field that contains a large number of methods of classification which are more or less well suited to the handwriting recognition. However, it did not highlight the unquestionable superiority and the choice of a method of classifying compared to others. Different methods have been proposed and high recognition rates are reported for handwriting recognition using hidden Markov models and neural networks. Al-Khateeb et al. [23] have presented an off-line recognition system Arabic handwritten text. The features were extracted from the segmented words using sliding window. The extracted features are fed to the HMM classifier. In order to improve accuracy, the HMM result is further refined by using a re-ranking scheme. Using the IFN/ENIT database, the system has achieved 95.15% recognition rate. Using an explicit segmentation module, El-Zobi et al. [24] have presented an off-line handwriting Arabic words recognition system based on hidden Markov model. Instead of using sliding window based features, they used shape representative features for each letter in each handwritten form. They have used two databases; the IESK-arDB for training and testing, and the IFN/ENIT database samples for validation. The recognition rates have reached 71% on the first database and only 42% for the second. Hussien et al. [25] proposed an optical character recognition Arabic handwritten using the Hopfield neural network. They used a small database for eight Arabic letters with a success rate of 77.25%. El-Adel et al. [26] presented a neural network architecture based on the fast wavelet transformation, learning and classification of the Arabic system of

handwritten character tests were conducted using the IESK-arDB data set that includes 6000 segmented characters. The rate of classification for groups of characters is 93.92%. Elleuch et al. [27] introduced an Arabic handwritten characters recognition using deep learning. The approach has been tested on the HACDB database with a classification error rate of 2.1%. Shatnawi et al. [28] model real distortions in Arabic script by using real examples of handwritten characters to recognize characters, they reach a 73.4 % recognition rate. Kef et al. [29] introduced a fuzzy neural network for Arabic handwritten characters recognition. The average result of the recognition rate is 93.8%. Al-Abodi and Li [30] have proposed a new system of recognition based on geometric characteristics of Arabic characters. The IFN/ENIT database has been used in their experimental results. The average result of the recognition rate is 93.3% to 596 words. Lawgali et al. [31] introduced a new framework for the recognition of handwritten Arabic words based on segmentation. An artificial neural network has been used to identify the shape of the character using its characteristics obtained by applying discrete cosine transform. The average result of the recognition rate is 90.73%. Benouareth and Sellami [32] presented a reference system for the recognition of cursive writing off-line based on hidden Markov models, and analytical type without segmentation. The step of recognition is based on extraction of vector of characteristics according to the technique of sliding windows but with different inclinations. The step of combination allows merging the results post-processed to produce the most suitable candidate. The approach proposed by Kundu et al. [33] and Pervez and Al-Ohali [34], is based on a sequential combination of a neural approach with a Markov approach, architecture has properties interesting both in terms of their performance and their relative small size. The method proposed by Pechwitz and Maergner [35] present a recognition system based on HMM of one semi-continuous dimension. Experiments have been made on the four distinct sets (a, b, c, d) of the IFN/ENIT base containing 26.459 handwritten Arabic words, all (a, b, c) is used for learning and overall d is used for the test. The system is obtained a recognition rate about 89%. The method proposed by Dreuw et al. [36] is to introduce a system based on HMM for the off-line recognition of Arabic script that explicitly models the white spaces between the characters and the related pieces of Arabic words (PAW). A visual inspection of the models of learning has shown the need for a precise modeling and adaptation of the lengths of characters. Experiments are conducted on the IFN/ENIT database. The system reached rate recognition of 92.86%. The method proposed by Benouareth et al. [37] describes off-line Arabic word recognition system without constraint-based approach without segmenting (segmentation-free) and semi-continus hidden Markov models with state time explicit, they offer a new version of the Viterbi algorithm taking into account the explicit state duration modeling. The results obtained realize a rate of 90.20% throughout 26.459 Arabic words of the IFN/ENIT database. The approach proposed by Mohamad et al. [38] is based on the combination of three classifiers based on homogeneous HMM. All the classifiers have the same topology as the reference system and differ only in the orientation of the sliding window. The results reported an a single classifier. Al-Khateeb et al. [39] offers off-line Arabic handwritten words recognition system, the vectors of features are used to classify the words using the classifier K-nearest neighbors. The proposed system has been tested successfully on the IFN/ENIT database. The experimental results show a recognition rate of 76.04%. El-Sawy et al. [40] models a deep learning architecture, a convolutional neural network was trained and tested to their database that contains 16800 handwritten Arabic characters. The use of a convolutional neural network led to significant improvements, indeed, they reach a classification error average of 5.1% on the test data.

In this brief, we conclude that HMMs dominate the field of the cursive handwriting recognition, the performance of a classifier rely on the quality of the features and of the classifier itself. A good set of features should represent the characteristics of a class and it is also invariant as possible for changes in this class.

To address the problems faced by researchers, and given the limitations of existing approaches to the recognition of Arabic script and the difficulties of the Arabic word segmentation, many of research are turned to use hybrid methods and in particular to the neural-Markov methods. Such a system needs to take into account a large number of the variability of characters, the main

advantage of the HMM is attributed to their probabilistic framework, which fits well with the nature of the signals noise as the case of handwriting. The main argument for the use of the MLP for the probabilities of output, it is that it is driven in a discriminative way and that no assumption is made on their distributions. Motivated by these advantages, we have proposed a recognition system evaluated using a hybrid model-based on neural network type multi-layer perceptron and hidden Markov models in a limited vocabulary.

3. The Hybrid System

A hidden Markov model is a stationary Markov chain where the observation is a probabilistic function of the state. We also have a notion of observation sequence that appears, meaning, at every given moment, there is a realization of a random variable according to the probability distribution associated with the state currently visited.

The recognition system models the words and characters in the form of Markov models hidden. The system is analytical where patterns of words are constructed by concatenating the character models. The probability densities of observations in each state are modeled by a Gaussian distribution. Learning using the iterative Expectation-Maximization algorithm. During initialization, the observations are affected in states segmenting them linearly then the first parameters are estimated from this assignment [41].

The discrete time Markov models have two disadvantages, the first is that learning is not discriminating in that the parameters of each word model are adjusted only with the word images associated with its class. On the other hand, the weakness of the discrete system is that the observations are often continuous vectors, the use of discrete distribution models thus implies a preliminary phase of vector quantification, with the resulting degradation. It is interesting to include observation densities in Markov models using neural networks, which incorporate discriminant functions as a result of learning. The solution adopted to estimate the probability of occurrence of observations is to use a multilayer neural network.

3.1. *The multilayer perceptron and estimation of the parameters of the neuro-Markovian network:*

Topologically, the multilayer perceptron consists of three layers, the first layer is the input layer consisting of the vector of characteristics obtained from the input image, and the second layer has the hidden units. Finally, the output layer is dimensioned to the number of classes to be discriminated. A neural network is a system that learns from a base of examples containing forms of associated inputs and outputs adjusts its internal settings in our case the synaptic weights. At the end of an optimal training, the output of the multilayer perceptron is a good estimate of probabilities with respect to each class presented at the input of neural network. The outputs of the network asymptotically approximate the Bayesian probabilities belonging to the classes [41], [42].

3.2. *Estimate of the parameters by the MLP network:*

We use a perceptron multilayer to estimate probabilities of issue $p(o_j/e_j, M)$. Each state of the Markov chain will be considered as a network class and observation will be the feature vector extracted from the character. The markovian alignment uses the $p(o/e)$ probability for observation of the character o since the state e, the two terms are related by the formula from Bayes:

$$p(o/e) = \frac{p(e/o) \times p(o)}{p(e)} \tag{1}$$

And for a sequence $o_1, \dots, o_n$ and a path $e_1, \dots, e_n$:

$$p(o_1, \dots, o_n, e_1, \dots, e_n, M) = \prod_{j=1}^{n} p\left(\frac{e_j}{o_j}, M\right) \times p\left(\frac{e_j}{e_{j-1}}, M\right) \times \frac{\prod_{j=2}^{n} p(o_j)}{\prod_{j=2}^{n} p(e_j)} \tag{2}$$

As the product of the probabilities of the segments of image $p(o_j)$ does not depend on the hypothesis of word M, we can write:

$$p(o_1, \dots, o_n, e_1, \dots, e_n, M) = \frac{\prod_{j=1}^{n} p\left(\frac{e_j}{o_j}, M\right) \times p\left(\frac{e_j}{e_{j-1}}, M\right)}{\prod_{j=1}^{n} p(e_j)} \quad (3)$$

The recognition is performed by finding the optimal path that will provide the class (discriminant path) in the case of one model for all classes, for this, the Viterbi algorithm determines the probability of the best path. After decoding, the observation series is associated with an optimal path, that is to say, one that maximizes the observation probability according to the formula (3). This path provides labeling for each observation by indicating the hidden Markov chain state with which it is associated. Learning consists in estimating the probabilities of transitions and observations for all the characters constituting the lexicon. The different probabilities of transitions between states are estimated by counting on the whole learning base. The probabilities of observation of characters are estimated by the outputs of the neural network, this network have as many outputs as possible states.

4. The Experimental Phase

4.1. *Architecture of the system proposed:*

We present a new method of segmentation for recognition of handwritten words based on a hybrid system type neuro-Markovian incorporating neural networks and hidden Markov models in a complementary architecture. The most common strategy in hybrid systems neuro-Markovian is to use neural networks to estimate Markovian model observation probabilities in their usual formalism. The proposed analytical method takes into account points and diacritics inclinations and the false positions in writing. Combination step merges the outputs produced by the HMM to choose the most suitable candidate word. Development start with the pre-processing of the images, then the extraction of the representative characteristics of those words to serve as input to the proposed classifier. In what follows, we will detail the work, giving first the general scheme of the system used before moving on to describe each completed step.

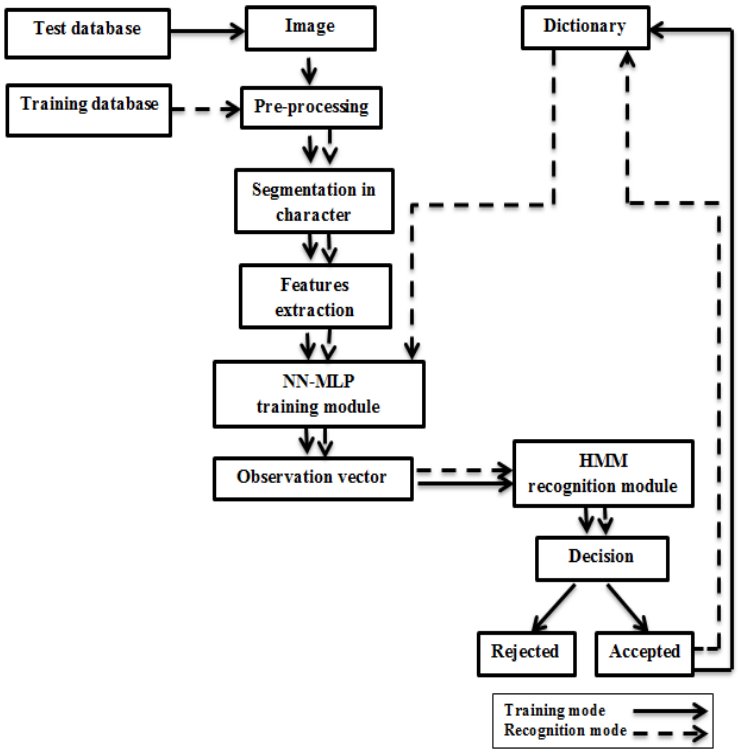


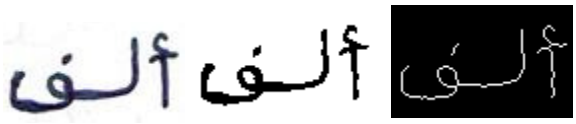
Figure 1. Methodology of the Arabic handwriting recognition.

4.2. *Data collection:*

The main goal here is to collect images of Arabic handwritten characters and words written by many writers and to make the databases as much representative as possible. So, a form is designed to do so, the form consists of 34 alphabets and 37 words written by 120 scripters. Also we have asked writers to use their everyday writing in order to get the most natural and unconstrained way of writing. No restrictions were imposed on the writing instrument. Hence, word produced with a number of different writing instruments is included in the database (ballpoint pens, ink pens, and pencils) all with various stroke widths. An example of a filled form is shown in Figs. 6-8. All form pages were scanned using a high quality scanner. The output of the scanner can be either a (.jpeg) or (.bmp) format. After collecting all forms, characters and words were extracted automatically and pre-processed in order to remove noises.

4.3. *Pre-processing:*

The task of pre-processing is an important step in any system of recognition, the purpose of this step in the handwritten word recognition is to improve the readability of the image and remove details that do not have power discriminative in the recognition process. The proposed system takes as input an image scanned from a handwritten Arabic word. To ease the difficult task of identification, we use the binarization, smoothness, standardization, the skeletonization and the estimate of the baseline [43], [44].



(a) (b) (c)

Figure 2. Image preprocessing "word ألف":
(a) Original image, (b) binary image and (c) skeleton image.

In our work, we examined the standardization operation that brings back the images of words and even the images of the characters to standard sizes and reduce all types of variations. Indeed, the size of a character can vary one entry to another, which can cause instability of parameters. The operation of smoothing is applied to eliminate the noise in the image and to describe it by a sequence of vectors of at least stable features. For this purpose, we use the algorithm of smoothing proposed in [14]. Extraction of characteristics (i.e. diacritical points) needs to estimate the baseline of word. The method described in [15], gives a good estimate of the baseline. It is based on the analysis of the histogram of horizontal projection after transformation to a binary image word [45]. To get the skeleton word we use the algorithm of Hilditch [46], he proceeds by successive refinements. The skeletonization is used to reduce the variability of writing style and make simple extraction of certain features. Generally, this will take a lot of time, and sometimes its application to Arabic script can remove diacritical points that are relevant primitives for the word discrimination. The algorithm of Hilditch has lower complexity and its application preserves the diacritical points. Fig. (2c) shows the result of the application of this algorithm in Fig. (2a).

4.4. *Segmentation method:*

Most of the currently proposed segmentation algorithms do not resolve the problem of duplication of characters in Arabic script. Segmentation stage is the most difficult stage, and the main source of errors in recognition. The purpose of this step is extraction of the characteristics obtained in vertical segmentation. The segmentation module that we implemented is achieved in two steps, segmented word in characters and the characters into related parts. It is a non-uniform segmentation based on the analysis of vertical projection histogram [45]. This algorithm is based on the skeleton of the word and the detection of the essential points (branch or cross). The general idea of this process is to make segmentation between each two branch points or crossing and the segmentation column must contain a single pixel after having extracted the skeleton of the word,

this technique involves the word segmentation in individual characters, segmentation points are identified at the end of a character and the beginning of the next. The skeleton word representation we permit to obtain some characteristics that are difficult to extract from the bitmap representation. In order to isolate the characters, our approach is to find the characteristic points of the most common characters. These points are usually close to the baseline. The general idea of characters segmentation is to determine the baseline for the given word based on the horizontal projection. The vertical projection on the word done by summing the pixels of the characters vertically, we find the characteristic points of character, where segmentation should take place if the width of the segmented part does not have to be very small and there is a rapid change in the vertical projection in the vicinity of the point (the case of the character "س").

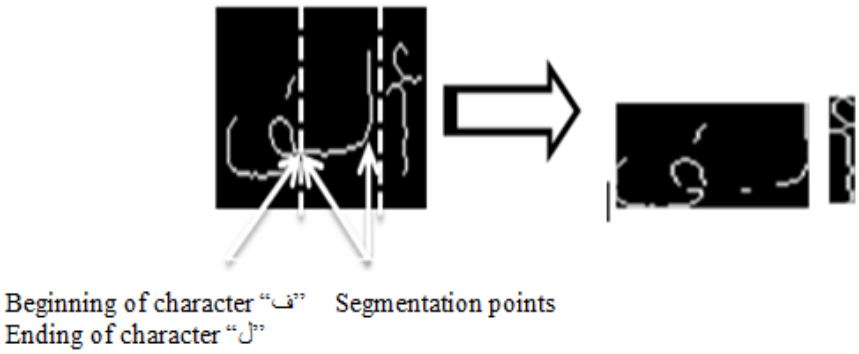


Figure 3. Segmentation into characters.

4.5. *Extraction of characteristics and vectorization:*

Extraction of primitives is to transform an image (character, grapheme, word...) in a vector of fixed size primitives. This transformation is to change the space of data representation of the image to an N-dimensional space ($\Re^{N)}$. The choice of characteristics is very decisive for the stage of recognition. We used in our system a mixture between statistical and structural characteristics from the literature that give better results. After several tests, we chose features that seem relevant to the discrimination of Arabic words and all of these features have been selected for the description of each character [47].

a. *Statistical characteristics:*

For these kinds of characteristics, we use the moments of projections: the mean μ, variance σ^2 which are calculated for different angles projection (horizontal, vertical, diagonal 45° and 135°), therefore, 08 features are extracted from the histograms of projection [48], [49]. For the description of a character, we choose Fourier descriptors which are introduced in the vector of primitives whose goal is to encode the contour in a more compact way. Fourier descriptors are invariant by translation, rotation and scaling [49], [50]. Also, we include the moments of Hu, these moments are descriptors overall since they take into account the internal organization of the shape of the character and a small number of these moments are used to describe a character. The moments of Hu are invariant to translations, rotations and scale changes [49], [51]. For these reasons, we are interested in this kind of attributes. Other moments are used are moments of Zernike, they are based on the principle of orthogonal polynomials. As a result, the reconstruction of the form from these moments is possible. A relatively small set of Zernike moments can characterize the overall shape of an object. If we use Zernike moments of higher order, the more precise is the reconstruction of the image of the object. The interest of calculating Zernike moments lies in their invariance vis-a-vis a translation, a change of scale or a rotation of a given form. We chose to calculate the Zernike moments from the skeleton of the word [49]. To extract density by zoning, the word is divided into three horizontal zones and each zone is divided into three vertical areas, we get finally nine areas. Densities in each area should be normalized by dividing by the surface of the area, since the words are not all the same size [52]. We get profiles of the handwritten word take into account diacritics, it is determined on a number of lines, in general, spread evenly over the

height of the character, the distance between the left edge (respectively, right, upper, bottom and oriented) character and the first black pixel met on this line. The set of these distances defines a left profile (respectively, right, upper, bottom and oriented) character. Profiles must be normalized (dividing by the width of the image of the character), therefore we get a feature vector that contains eight profiles [61]. The direction of the plot is sufficient to define the shape of the skeleton, the most famous code is the Freeman code of 8-connectivity. For this method, we have obtained a vector that contains the distribution on the eight orientations, over the whole character from which we draw a vector of characteristics of dimension eight [43], [53].

b. *Structural characteristics:*

Features which are calculated from the skeleton image correspond to firstly, characteristic points, they represent the black pixels in the skeleton of a word with a number of different neighbor of 0 and 2. There are two types, the extreme points and junction points. An extreme point corresponds to a segment beginning/end of a line segment. Junction points connect three or more in the skeleton word, and are divided into points of intersection and branch points [31], [43]. Secondly, we extract the ascenders and descenders, generally, the first step is the determination of the median area of the word that distinguishes the letters to ascenders or descenders. The method used is based on the analysis of the horizontal histogram. We are looking at the top line data index (m1) as well as in the lower part, the indices of the minima of the histogram respectively M1 and M2. In the ideal case, these two minima mark the middle zone [43]. Also, we extract and classify the secondary tracks, diacritical points corresponds to black pixels whose skeleton with a number of neighbors is equal to 0, these points are distinguished by their position (above or below the base line).

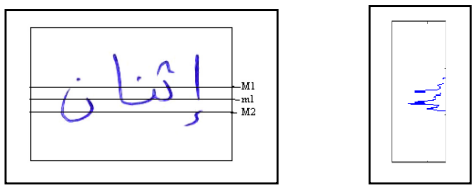


Figure 4. Detection the median area by horizontal projection.

In Arabic words, diacritics parties (hamza, point...) which can be considered as a secondary trait are an integral part of the characters, their number and their position above or below the character, change the meaning of the latter. This property has led us to detect the diacritics parts in the characters. For this purpose, we have adopted the compact setting by δ such as δ is the number of pixel. Therefore, discrimination between the main and secondary tracks on the one hand and between the secondary tracks themselves, can be done by application of the following chart:

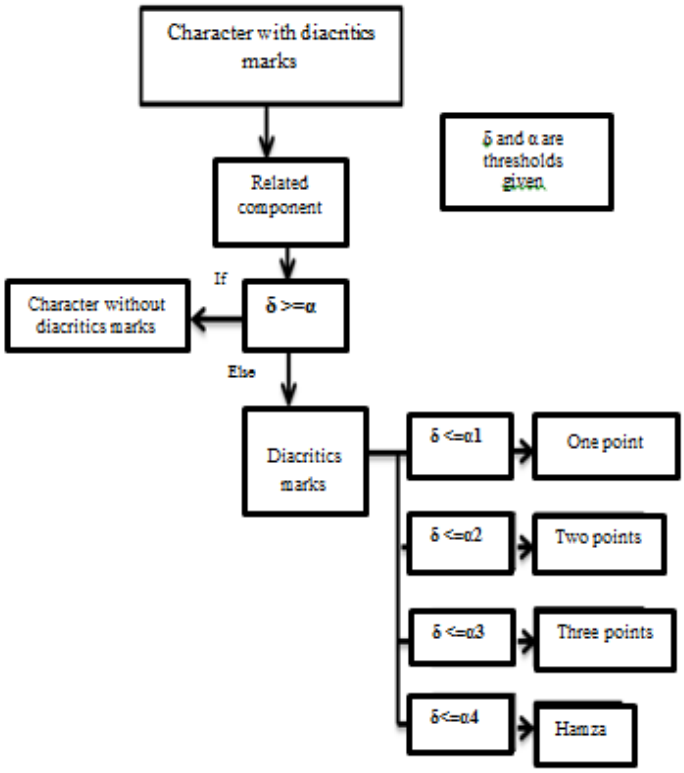


Figure 5. Traces classification flowchart.

Another characteristic is used pieces of Arabic word (PAW), the extraction of PAW procedure also called labeling of the pixels, is widely used in recognition of forms to segment the binary images. The technique is to group the neighboring pixels into a related component called set. Each set is disjoint from the others and can be easily isolated. We have use an algorithm works in a single pass, following the criterion of 8-connectedness [36].

All these features have been selected for the description of each character. As a result we get a vector of characteristics with statistical and structural characteristics for each character. The final result of vectorization is the composition of a matrix of characteristic for each manuscript word of size $l * z$ such that l is the number of character and z and the total number of primitives. The choice of this type of features is amply motivated by the simplicity and robustness of their calculation, as well as by their power discriminative.

$$\text{Features matrix} = \begin{pmatrix} \text{Features vector of } 1^{st} \text{ character in } pos_1 \\ \text{Features vector of } 2^{nd} \text{ character in } pos_2 \\ . \\ . \\ . \\ \text{Features vector of } Z^{th} \text{ character in } pos_Z \end{pmatrix}$$

4.6. *Learning:*

Neural network requires a large data of images of handwritten characters training to get a good result. Available training data are divided into two different sets, set of learning and validation set. There should not be any overlap between these two sets of data in order to improve the ability of generalization of a neural network. The test is designed to access the network generalization ability.

Pos8	Pos7	Pos6	Pos5	Pos4	Pos3	Pos2	Pos1	العدد	الرقم	
				ـد	ـحـ	ا	و	واحد	1	واحد
					ـدـ	ـحـ	ا	احد	2	احد
			ن	ـا	ـذـ	ـثـ	ا	اثنان	3	اثنان
				ـا	ـذـ	ـثـ	ا	اثنا	4	اثنا
			ة	ـث	ا	ـلـ	ث	ثلاثة	5	ثلاثة
			ة	ـع	ـر	ر	ا	اربعة	6	اربعة
				ـة	ـمـ	ـمـ	خ	خمسة	7	خمسة
					ـة	ـتـ	س	ستة	8	ستة
				ـة	ـع	ـبـ	س	سبعة	9	سبعة
		ة	ـيـ	ث	ـمـ	ـمـ	ث	ثمانية	10	ثمانية
				ـة	ـع	ـسـ	ت	تسعة	11	تسعة
				ة	ـر	ـشـ	ع	عشرة	12	عشرة
					ـر	ـشـ	ع	عشر	13	عشر
			ن	و	ـر	ـشـ	ع	عشرون	14	عشرون
		ن	ـو	ث	ا	ـلـ	ث	ثلاثون	15	ثلاثون

Figure 6. Overview of database 1.

Figure 7. Overview on the test database.

Learning is to establish two types of classes, a class of words and a class of letters. First a list ranked by order of each character position (Fig. 6). The goal is to get the correct character in the first position in this list. Second, a list classified according to the letter (Fig. 8).

Isolé	Fin	Milieu	Début	الحرف	الرقم
ا	ـا	ـا	أ	أ	1
ب	ـب	ـبـ	بـ	ب	2
ت	ـت	ـتـ	تـ	ت	3
ث	ـث	ـثـ	ثـ	ث	4
ج	ـج	ـجـ	جـ	ج	5
ح	ـح	ـحـ	حـ	ح	6
خ	ـخ	ـخـ	خـ	خ	7
د	ـد	ـدـ	دـ	د	8
ذ	ـذ	ـذـ	ذـ	ذ	9
ر	ـر	ـرـ	رـ	ر	10
ز	ـز	ـزـ	زـ	ز	11
س	ـس	ـسـ	سـ	س	12
ش	ـش	ـشـ	شـ	ش	13
ط	ـط	ـطـ	طـ	ط	14

Figure 8. Preview on the database 2.

For the dictionary design, it is based on sequences of representative training vectors. This procedure is to present each class by a characteristic vector, then, it optimizes iteratively the partition of the dictionary by using the fast K-means algorithm [43]. The result of this fast K-means is used to initiate a multi-layer perceptron, of which the number of classes of output will be the same as the number K of partitions of the fast K-means. Learning of the HMM and MLP is done separately in four steps:

1) Decode words with the hybrid system databases to create a base of characteristic vectors annotated to the neural network.

2) Train neural network by retro spread of the gradient.

3) Use the new neuron network to calculate the probabilities of observation.

4) Optimize the transition probabilities of the states of the HMM by the Baum-Welch algorithm [42].

This iterative process is repeated throughout the base training up to performance saturation using the Viterbi alignment [42]. It only exploits the sequence of states that make up the best path and forward-backward probabilities, which to redistribute probabilities subsequently on several classes and no longer targets on a single.

4.7. *Classification and recognition:*

Recognition model combines MLP and HMM works to recognize a word as shown in Fig. 2. The Markov model used is organized into columns of states, and each state may only issue a single class of observations. Each column includes N states where N is the number of all possible classes of characters making up the words in the lexicon. In our application, the MLP is upstream with the HMM. Emission probability of the observations are calculated by MLP, transition probabilities are estimated by counting and priori probabilities are obtained by calculating the number of occurrences of each state (class) in the learning base. We adopted a model (Discriminant Path) that has one model for all classes of words, by searching for the optimal path which is based on Bayesian probability. To recognize a word by the system, the likelihood of words is calculated as the sum of probabilities over all possible paths through the HMM model. In this case, the recognition is to determine the path corresponding to the sequence of observation, that is to be found in the model, the suite of states, called sequence of states of Viterbi, that maximizes the quantity $p(e/o,A)$, the Viterbi algorithm is used to find the optimal path representing the recognized word. For a word image to be recognized, we will retain one and only a succession of states (classes) provided by the HMM which maximizes the probability of observation. The performance of a classifier can be measured by calculating the three following rates:

$$\text{rejection rate} = \frac{\text{number of rejected forms}}{\text{total number of forms}} \quad (4)$$

$$\text{recognition rate} = \frac{\text{number of recognized forms}}{\text{total number of forms}} \quad (5)$$

$$\text{substitution rate} = \frac{\text{number of males recognized forms}}{\text{total number of forms}} \quad (6)$$

The procedure of learning and recognition for a character is the same for a handwritten word.

5. Tests, Results and Discussion

Our goal was to implement a new method of segmentation into characters and test the effectiveness of this algorithm in a recognition system hybrid neural-Markovian of handwritten Arabic words. To validate the proposed approach, we conducted experiments on two databases of handwritten words, one on a lexicon (Database 1) of size 37 for the literal amounts (Fig. 7), the other (Database 2) of size 34 but formed by the handwritten Arabic letters (Fig. 8). Each base contains about 4440 samples of words manuscripts and 4440 samples for handwritten letters that are written by 120 scripters (secondary students and personal administrative). These databases have been developed by our research team at the laboratory of automatic of Constantine. We have divided the basis of words in two sub-bases, the first contains 80% of the words for the operation of learning and the second contains 20% of the words for testing (Fig. 9), for the character database we also added about 5500 samples for the different letter formats from the DBAHCL database [54].

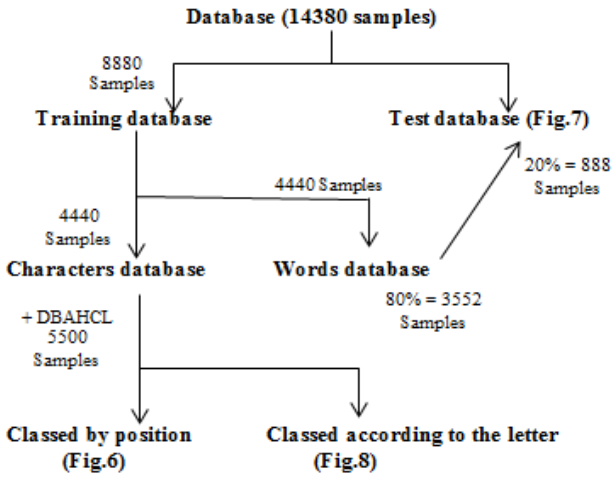


Figure 9. Creating database flowchart.

Several tests were conducted to assess the recognition rate of the system according to the segmentation procedure of the words image in print.

Table 1. Success and error rates obtained for different position of the arabic word.

Characters groups	Isolated	Beginning	End	middle
Recognition rate (%)	100	98.7	97.2	86.4
Error rate (%)	0	1	1.7	4.6
Substitution rate (%)	0	0.3	1.1	9

We have a recognition rate of character at the beginning of word of 98.7%, this shows that the majority of the characters are recognized in the first position, where a perfect recognition of the overall word (Table 1). According the results obtained (Tables 1-3), we notice that it is no big error at recognition except for characters س and ش because of the similarity of these characters. We also reach substitution rates of 0% for the characters (ا, ب ت, ل, ن, ه and ي), it shows the great potential and the ability of recognition of our system to recognized handwritten Arabic characters. We can also see that some characters are very confused with similar characters (ex: ح, عand خ), the small distinction of the structure and position of the diacritical point causes challenges for some similar characters. There are also certain traits of the characters are missing with the structure of the character. Other characters may have an extra stroke.

Maybe these arguments of the difficulty of recognition of Arabic characters are the main reason why some words are very difficult to recognize as (ستة and تسعة).

Table 2. Success and error rates obtained from the database of the character.

Arabic letters	Rates (%)			Arabic letters	Rates (%)		
	recognition	reject	substitution		recognition	reject	substitution
ا	100	0	0	غ	98	0	2
ب	100	0	0	ف	97	1	2
ت	100	0	0	ق	90	7	3
ث	97	2	1	ك	98	2	0
ج	94	2	4	ل	100	0	0
ح	99	0	1	م	98	1	1
خ	98	1	1	ن	100	0	0
د	98	1	1	ه	100	0	0
ذ	89	4	7	ئ	93	3	4
ر	97	1	2	ء	100	0	0
ز	98	0	2	ى	98	1	1
س	89	7	4	ي	100	0	0
ش	87	5	8	و	96	1	3
ط	95	2	3	ؤ	90	4	6
ظ	93	1	2	ص	98	2	0
ع	99	0	1	ض	94	2	4

Table 3. Success and error rate obtained from the database of the word.

Arabic Words	Rates (%)			Arabic words	Rates (%)		
	recognition	reject	substitution		recognition	reject	substitution
واحد	98.25	0.3	0.2	تسعون	91.56	3.4	5.04
اثنان	98.76	1	0.04	مائة	97	1	2
ثلاثة	86.5	11.3	2.2	مئة	94	2	4
اربعة	98.06	1	0.94	مائتان	86.7	5	8.3
خمسة	88.67	11	0.33	ثلاثمائة	98.5	1.1	0.4
ستة	91.04	5.45	3.51	أربعمائة	83	5	12
سبعة	96.38	2.42	12	خمسمائة	91.56	3.4	5.04
ثمانية	97	2	1	ستمائة	93	3	4
تسعة	93	3	4	سبعمائة	97	2	1
عشرة	98.7	0.9	0.4	ثمانمائة	86.7	5	8.3
عشر	97	1	2	تسعمائة	100	0	0
عشرون	98.7	0.9	0.4	ألف	99	1	0
ثلاثون	94	2	2	آلاف	98.7	0.8	0.5
اربعون	98.7	0.9	0.4	ألفا	100	0	0
خمسون	91.56	3.4	5.04	ألفان	100	0	0
ستون	91.56	4.3	5.04	دينار	98	1	1
سبعون	96.38	2	12.42	جزائري	90.24	6.76	3
ثمانون	97	2	1	احد	99	1	0

Table 4. Recognition and error rate achieved on our databases.

Rates (%)	Words database	Character database
Recognition rate	94.7	96.31
Rejection rate	02.79	01.56
Substitution rate	03.09	02.12

At character level, the recognition error is 3.37% and the success rate is 97.17%. This shows that the majority of the characters are recognized hybrid network that gave us a rate of recognition words of 94.7%. Taking into account the context by hybrid system brings an increase in the rate of recognition (Table 4).

We conducted a precise research at the level of each step in the process of recognition and we noticed that the images of these words have lost information at the level of the acquisition and sometimes the skeletisation. So this is the poor quality of the images that has disrupted the recognition process and the bad writing of some writers. Fig. 10 shows typical examples of recognition errors

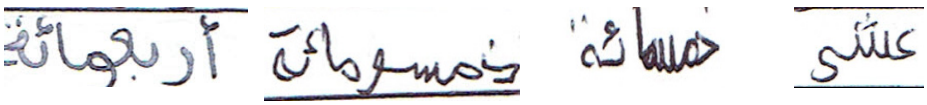


Figure 10. Samples examples for bad writing.

6. Conclusion and Prospects

Our goal is off-line recognition of handwritten Arabic words with a limited vocabulary using an analytical method with application on the Arabic literal amounts. In order to achieve an off-line handwritten Arabic words recognition system we have made first necessary transformations to process forms. They show different aberrations of the acquisition system which must be to guard before starting any procedure of recognition. In order to better characterize the image of the word,

we segment it into series of characters and from each of these characters, we calculate a vector of parameters. This allows us to describe the image of the word by a matrix of features or each line represents a vector of characteristics of the character of the word. For Arabic handwriting recognition rate depends on the segmentation so we based on the analytical approach for the development of segmentation technique because a good segmentation maintains the full character making it easy for operations that follow (extraction of parameters and recognition). The work on the recognition of handwriting showed recognition system performance can be improved by using a combination of statistical and structural characteristics. Indeed, a bad choice of primitives influences negatively on the results even if a very powerful classifier is used. Recognize an object is decided that the vector characterizing the object analyzed to recognize close to the vector memorized at the time of the training. The choice of classifier is based on its speed and its ability to deal with heterogeneous data. HMMs are considered among the most frequently used and successful methods in the recognition of off-line Arabic manuscript words. As the Arabic script is cursive nature, one of the major obstacles encountered in an effort to improve the Arabic handwriting recognition is the lack of effective and efficient solutions to the problem of segmentation. The originality of our approach is that it is possible to propose a recognition system taking into account diacritics with their characters in the segmentation, the main interest of this technique that it is easier to find the set of potential segmentation points. In this system, the model applied to a vocabulary of the literals amounts. A new hybrid model type neural-Markovian allows incorporating more contexts by setting the diacritics to their characters. The reduction of the lexicon using a sophisticated segmentation algorithm while introducing a new modeling of words.

We showed that the results are promising with a general recognition rate of 98.59% for words and 94.85% for characters. Poor recognition of the word or characters is because of several factors, including the number of characters, the complexity of each character, the similarity between the characters, in addition to the quality of the database and the means of exercising it.

Conflict of Interest

The author(s) declare(s) that there is no conflict of interest.

Acknowledgments

The authors would like to thank the anonymous reviewers for any comments and suggestions that enhance the technical and scientific quality of this paper. The authors would also like to thank the members of laboratory, for their valuable suggestions and comments, which helped us to improve this paper.

References

[1] I. Yousef, A. Shaout, Off-line Handwriting Arabic Text Recognition: A Survey, Inter. J, Advanced Research in Computer Science and Software Engineering. 7(4) (2014).

[2] A.M. Ali, A classifier for Arabic handwritten characters based on supervised self-organizing map neural network, in: Proc. Inter. Conf. Mathematical models for engineering science, 2010.

[3] M. Ali et al., Fuzzy Logic approach to Recognition of Isolated Arabic Characters, Int. Jour. Computer Theory and Engineering. 1(2) (2010) 119-124.

[4] Y. El-glaly, F. Quek, Isolated Handwritten Arabic Character Recognition using Multilayer Perceptrons and K Nearest Neighbor Classifiers, unpublished, 2011.

[5] D. Laslo, A. Al-Hamadi, M. El-Zobi, An Active Shape Model based approach for Arabic handwritten character recognition, in: Proc. IEEE 11th Inter. Conf. Signal Process (ICSP), Vol. 2, 2012.

[6] F.H. Zawaideh, Arabic Hand Written Character Recognition Using Modified Multi-Neural Network, Int. J. Emerging Trends in Computing and Information Sciences. 7(3) (2012) 1021-1026.

[7] A. Sahlol, S. Cheng, A Novel Method for the Recognition of Isolated Handwritten Arabic Characters," Inter. J, Computer Vision and Patt. Recogn., preprint, 26 Feb 2014, arXiv:1402.6650.

[8] S.A. Azeem, M. El-Meseery, Arabic Handwriting Recognition Using Concavity Features and Classifier Fusion," in: Proc. IEEE 10th Inter. Conf. Machine Learning and Applications and Workshops (ICMLA), Vol. 1, 2011, pp. 200–203.

[9] S. A. Mahmoud, S. O. Olatunji, Handwritten Arabic numerals recognition using multi-span features & Support Vector Machines, IEEE 10th Inter. Conf. Information Sciences Signal Processing and their Applications (ISSPA), 2010, pp. 618-621.

[10] M.T. Parvez, S. Mahmoud, Arabic Handwritten Alphanumeric Character recognition using Fuzzy Attributed Turning Functions," Inter. J, Patt Recogn. 46(1) (2013) 141-154.

[11] I. Lawal et al., Recognition of handwritten Arabic (Indian) numerals using freeman's chain codes and abdicative network classifiers, in: IEEE 20th Inter. Conf. Patt. Recogn., 2010, pp. 1884–1887.

[12] G.F. Soleimanian, E.A. Zadeh, Artificial Neural Network Application in Letters Recognition for Farsi/Arabic Manuscripts, Inter. J. Scientific & Technology Research. 8(1) (2012) 90-94.

[13] A. Boukharouba, A. Bennia, Recognition of Handwritten Arabic words using a neuro-fuzzy network, Proc, 1st Mediterranean. Confer. Intell. Systems and Automation, 2008, pp. 254-259.

[14] E. Augustin, Reconnaissance de mots manuscrits par systèmes hybrides Réseaux de Neurones et Modèles de Markov Cachés, PhD thesis, Rene Descartes Univ., Paris V, 2001.

[15] A. Boukharouba, A. Bennia, Recognition of Handwritten Arabic Literal Amounts Using a Hybrid Approach, Cognitive Computation. 2(3) (2011) 382–393.

[16] Y. Osman, Segmentation algorithm for Arabic handwritten text based on contour analysis, IEEE Inter. Conf, Computing, Electrical and Electronics Engineering (ICCEEE), 2013.

[17] S. Alma'adeed, C. higgens, D. Elliman, Recognition of off-line handwritten arabic words using hidden markov model approach, in: Proc. 16th Inter. Conf, Patt. Recogn. 3 (2002) 481-484.

[18] A.M. Gouda, M.A. Rashwan, Segmentation of connected Arabic characters using hidden markov models, IEEE Inter. Conf. Comput. Intell, Measurement Systems and Applications CIMSA, 2004, pp. 115-119.

[19] Y. Boulid, A. Souhar, M.Y. Elkettani, Segmentation approach of Arabic manuscripts text lines based on multi agent systems, Inter. J. Comput. Information Systems and Industrial Management Applications. 8 (2016) 173-183.

[20] F.B. Samoud, S.S. Maddouri, H. Amiri, Three Evaluation Criteria's Towards a Comparison of Two Characters Segmentation Methods for Handwritten Arabic Script, Inter. Conf. Handwriting Recogn., 2012.

[21] Z. Tamen, H. Drias, How to overcome some segmentation problems in a constrained handwritten arabic character recognition system, IEEE 10th Inter. Conf, Information sciences signal processing and their applications (isspa), 2010.

[22] A. Lawgali et al., Automatic segmentation for Arabic characters in handwriting documents, in: 18th IEEE Inter. Conf. Image Processing (ICIP), 2011, pp. 3529-3532.

[23] J.H. Al-Khateeb et al., "Offline handwritten Arabic cursive text recognition using Hidden Markov Models and re-ranking," Patt. Recogn. Letters. 8(32) (2011) 1081-1088.

[24] M. El-zobi et al., A Hidden Markov Model-Based Approach with an Adaptive Threshold Model for Off-Line Arabic Handwriting Recognition, in: 12th IEEE Inter. Conf. In Document Analysis and Recognition, 2013, pp. 945-949.

[25] R.S. Hussien, A.A. Elkhidir, M.G. Elnourani, Optical Character Recognition of Arabic handwritten characters using Neural Network, in: Proc. Inter. Conf. Comput. Control. Networking. Electronics and Embedded Systems Engineering, 2015, pp. 456-461.

[26] A. El-Adel et al., Dyadic Multi-resolution Analysis-Based Deep Learning for Arabic Handwritten Character Classification, in: Proc. 27th IEEE Inter. Conf. Tools with Artificial Intelligence (ICTAI), 2015, pp. 807-812.

[27] M. Elleuch, N. Tagougui, M. Kherallah, Arabic handwritten characters recognition using Deep Belief Neural Networks, in: Proc. 12th Inter. Multi-Conf. Systems, Signals & Devices (SSD), 2015, 1-5.

[28] M. Shatnawi and S. Abdallah, Improving Handwritten Arabic Character Recognition by Modeling Human Handwriting Distortions, ACM Trans. Asian Low-Resour, Lang Inf. Process. 15 (2015) 1-12.

[29] M. Kef, L. Chergui, S. Chikhi, A novel fuzzy approach for handwritten Arabic character recognition, Pattern Analysis and Applications. (2015) 1-16.

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

[31] A. Lawgali, M. Angelova, A. Bouridane, A Framework for Arabic Handwritten Recognition Based on Segmentation, Inter. J. Hybrid Information Technology. 7 (2014) 413-428.

[32] A. Benouareth, M. Sellami, Proposition d'une méthode structurelle pour la reconnaissance des mots arabes manuscrits par approche globale, Jour. Communication INI, Alger, 1998, pp. 121-131.

[33] A. Kundu et al., Arabic handwriting recognition using variable duration HMM, in: 9th IEEE Inter. Conf. In Document Analysis and Recognition. 2 (2007) 644-648.

[34] A. Pervez, Y. Al-Ohali, Arabic Character Recognition: Progress and Challenges, Inter. Conf. Advanced Comput. Science Applications and Technologies. (2012).

[35] M. Pechwitz, V. Maegner, HMM Based approach for handwritten Arabic Word Recognition Using the IFN/ENIT– DataBase, ICDAR'03, 2003, pp. 890-894.

[36] P. Dreuw, S. Jonas, H. Ney, White-space models for offline Arabic handwriting recognition, in: 19th Inter. Conf, Patt Recogn, 2008, pp. 1- 4.

[37] A. Benouareth, A. Ennaji, M. Sellami, Arabic Handwritten Word Recognition Using HMMs with Explicit State Duration, EURASIP. J. Advances in Signal Processing. ID 247354 (2008).

[38] R.A. Mohamad, L. Likforman-Sulem, C. Mokbel, Combining slanted-frame classifiers for improved HMM-based Arabic handwriting recognition, IEEE Trans. Patt. Analysis and Machine Intell. 7(31) (2009) 1165-1177.

[39] H. Al-Khateeb et al., Word-based Handwritten Arabic Scripts Recognition using DCT Features and Neural network Classifier, in: 5th Inter.Multi-Conf, Systems, Signals and Devices, 2008, pp. 1–5.

[40] A. El-Sawy, M. Loey, H. EL-Bakry, Arabic Handwritten Characters Recognition using Convolutional Neural Network, WSEAS Trans. Comput. Research. 5 (2017) 2415-1513.

[41] A.F. Gernot, Markov models for pattern recognition from theory to applications, Advances in Computer Vision and Pattern Recognition, 2nd edition, Springer, October 2013.

[42] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, IEEE Proc. 2(77) (1989)2 57–286.

[43] R.C. Gonzalez, R.E. Woods, Digital Image Processing, 2nd edition, Addison Wesley, 2001.

[44] A. Boukharouba, A. Bennia, Novel feature extraction technique for the recognition of handwritten digits, Applied Computing and Informatics. (2016).

[45] N. Otsu, A threshold selection method from gray-scale histogram, IEEE Tran. System, Man, and Cybernetics. 9 (1979) 62–66.

[46] J. Hilditch, Linear skeletons from square cupboards, Machine Intelligence. 4 (1969) 404–420.

[47] F. Lauer, C.Y. Suen, G. Bloch, A trainable feature extractor for handwritten digit recognition, Pattern Recogn. 40 (2007) 1816– 1824.

[48] C.H. The, R.T. Chin, On image analysis by the methods of moments, IEEE Trans. Pattern Anal. Mach. Intell. 10 (1988) 496–513.

[49] M. Cheriet et al., Character Recognition Systems: A Guide for students and Practioners, John Wiley & Sons Inc., Hoboken, New Jersey, 2007.

[50] H. Kauppinen, T. Seppanen, M. Pietikamen, An experimental comparison of autoregressive and Fourier-based descriptors in 2D shape classification, IEEE Trans. Pattern Anal. Mach. Intell. 17 (1995) 207–210.

[51] A.L. Koerich, Unconstrained handwritten character recognition using different classification strategies, in: Proc. Inter. Workshop, Artificial Neural Networks, Patt. Recogn., 2003, pp. 52–56.

[52] D. Impedovo, Zoning methods for handwritten character recognition: a survey, Patt. Recogn. 3(47) (2014) 969-981.

[53] C.L. Liu et al., Handwritten digit recognition: investigation of normalization and feature extraction techniques, Patt. Recogn. 2(27) (2004) 265–279.

[54] A. Lawgali, M. Angelova, A. Bouridane, HACDB: Handwritten Arabic characters database for automatic character recognition, Workshop, Visual Information Processing (EUVIP), 2013, 255-259.

[55] S. Benchaou, Features extraction for offline handwritten character Recognition, Europe and MENA Cooperation Advances in Information and Communication Technologies, 2007, 209-217.

[56] H. Nemmour, Y. Chibani, Artificial Immune Algorithm for Handwritten Arabic Word Recognition, Inter. J. Information Technology. 2(14) (2017).

[57] Z.Q. Liu, J. Cai, R. Buse, Handwriting recognition soft computing and probabilistic approaches, Springer. 133 (2010) 31-57.

[58] A. Belaıd, C. Choisy, Human reading based strategies for off-line arabic word recognition, in: Proc, Inter. Conf, Arabic and Chinese Handwriting Recognition. 4768 (2006) 36-56.

[59] T. Sari, L. Souici, M. Sellami, Off-line handwritten Arabic character segmentation algorithm, in: Proc. 9th IEEE Inter. Workshop, Handwriting Recognition, Computer Society, 2002, pp. 452.

[60] H.M. Eraqi, S. Abdelazeem, HMM-based Offline Arabic Handwriting Recognition: Using New Feature Extraction and Lexicon Ranking Techniques, in: Inter. Conf. Handwriting Recogn., 2012, pp. 554–559.

[61] R. El-Hajj, L.S. Laurence, C. Mokbel, Arabic handwriting recognition using baseline dependant features and hidden Markov modeling, in: Proc. 8th IEEE Inter. Conf. Document Analysis and Recognition, 2005, pp. 893-897.

[62] H. El-Abed, V. Margner, Comparison of Different Preprocessing and Feature Extraction Methods for Offline Recognition of Handwritten Arabic Words, in: 9th Inter. Conf. Document Analysis and Recognition. 2 (2007) 974-978.