A Neural Approach for the Offline Recognition of the Arabic Handwritten Words of the Algerian Departments

Salim Ouchtati, Jean Sequeira, Mouldi Bedda

***Abstract*—**In the context of the handwriting recognition, we propose an off line system for the recognition of the Arabic handwritten words of the Algerian departments. The study is based mainly on the evaluation of neural network performances, trained with the gradient back propagation algorithm. The used parameters to form the input vector of the neural network are extracted on the binary images of the handwritten word by several methods. The Distribution parameters, the centered moments of the different projections of the different segments, the centered moments of the word image coding according to the directions of Freeman, and the Barr features applied binary image of the word and on its different segments. The classification is achieved by a multi layers perceptron. A detailed experiment is carried and satisfactory recognition results are reported.

***Keywords*—**Handwritten word recognition, neural networks, image processing, pattern recognition, features extraction.

I. Introduction

WRITING, which has been the most natural mode of collecting, storing, and transmitting information through the centuries, now serves not only for communication among humans but also serves for communication of humans and machines. The handwritten writing recognition has been the subject of intensive research for the last three decades. However, the early researches were limited by the memory and power of the computer available at that time. With the explosion of information technology, there has been a dramatic increase of research in this field. The interest devoted to this field is explained by the potential mode of direct communication with computers and the huge benefits that a system, designed in the context of a commercial application, could bring. According to the way handwriting data is generated, two different approaches can be distinguished: on-line and off -line. In the former, the data are captured during the writing process by a special pen on an electronic surface. In the latter, the data are acquired by a scanner after the writing process is over. Off-line and on-line recognition systems are also discriminated by the applications they are devoted to. The off-line recognition is dedicated to bank check processing, mail sorting, reading of commercial forms, etc., while the on-line recognition is mainly dedicated to pen computing industry and security domains such as signature verification and author authentication. Optical characters recognition (OCR) is one of the successful applications of handwriting recognition; this field has been a topic of intensive research for many years. First, only the recognition of isolated handwritten characters was investigated [1], [2], but later whole words were addressed [3]. Most of the systems reported in the literature consider constrained recognition problems based on vocabularies from specific domains, e.g. the recognition of handwritten check amounts [4], [5] or postal addresses [6], [7]. Free handwriting recognition, without domain specific constraints and large vocabularies, was addressed only recently in a few papers. The recognition rate of such systems is still low, and there is a need to improve it. Character and handwriting recognition has a great potential in data and word processing, for instance, automated postal address and ZIP code reading, data acquisition in banks, text-voice conversions, etc. Because of intensive research and development efforts, systems are available for English language [8]-[10], Chinese language [11], Arabic language [12]-[14] and handwritten numerals [15]. There is still a significant performance gap between the human and the machine in recognizing unconstrained handwriting. This is a difficult research problem caused mainly by the huge variation in writing styles, the overlapping and the intersection of neighboring characters. Today, the OCR (Optical Characters Recognition) systems are only able to recognition high quality printed or neatly handwritten documents. The current research is now basing on documents that are not well handled and including severely degraded, Omni font machine printed text, and unconstrained handwritten text. A wide variety of techniques is used to perform handwriting recognition. A general model for handwriting recognition is used to highlight the many components of a handwriting recognition system. The model begins with an unknown handwritten character that is presented at the input of the recognition system as an image. Firstly, to convert this image into information understandable by computers, parameterization operation is needed which extracts from the image all of the necessary meaningful information in a compact form, compatible with the computer language. This involves the preprocessing of the image to reduce some undesirable variability that only contributes to

Salim Ouchtati is with the Skikda Electronic Laboratory, Engineering Department, Faculty of Technology, Skikda University, Route El Hadaik, Bp: 26 Skikda 21000, Algeria (Phone: 0021393935198, (e-mail: s.ouchtati@univ-skikda.dz).

Jean Sequeira is with the Laboratory of Information Sciences and Systems LSIS (UMR CNRS 7296) Aix Marseille University Marseille, France (e-mail: Jean.Sequeira@univ-amu.fr).

Mouldi Bedda is with the Electrical Engineering Department, College of Engineering Al-Jouf University, Saoudi Arabia (e-mail: mouldi_bedda@yahoo.fr).

complicate the recognition process. Operations like slant correction, smoothing, normalization, etc. are carried out at this stage. The second step is to extract discrimination features from the image either to build up a feature vector or to generate graphs, string of codes or sequence of symbols. However, the characteristics of the features depend on the preceding step. Features extraction method is probably the most important factor in achieving high recognition performance in character recognition systems, extracted features must be invariant to the distortions, translations, and rotations. The features vector size is also important in order to avoid a phenomenon called the dimensionality problem. Several methods for features extraction are designed for different representations of the characters, such as binary characters, character contour, skeletons (thinned characters), or even gray levels characters [16]. The features extraction methods are valued in terms of invariance properties, and expected distortions and variability of the character. Today, the studies are based not only on how to choose the appropriate features extraction methods, but also on the selection of meaningful and pertinent features from the features vector [17]-[19]. The final step is the character recognition; most recognizers have adopted classical pattern classification methods. Major approaches are statistical based, structural analysis, template matching, and neural network approaches. Significant progress has been made in these classification methods but more work is required to match human performance.

II. A Recognition System for the Arabic Handwritten Words of Algerian Departments

In the setting of the handwritten writing recognition, we proposed an off line system for the recognition of the handwritten Arabic words of the Algerian departments (shown in Fig. 1), this system is divided in three phases:

- Acquisition and preprocessing
- Features extraction
- Recognition

A. Acquisition and Preprocessing

1. Acquisition

Before analyzing the different processing steps, let's recall that we are especially interested at the off line processing. For our case, the acquisition is done with a numeric scanner of resolution 300 dpi with 8 bits/pixels, the used samples are all possible classes of the handwritten words of all the Algerian departments, with variable sizes and variable thickness, and with 10000 samples for every class. Fig. 2 shows some samples of the used database.

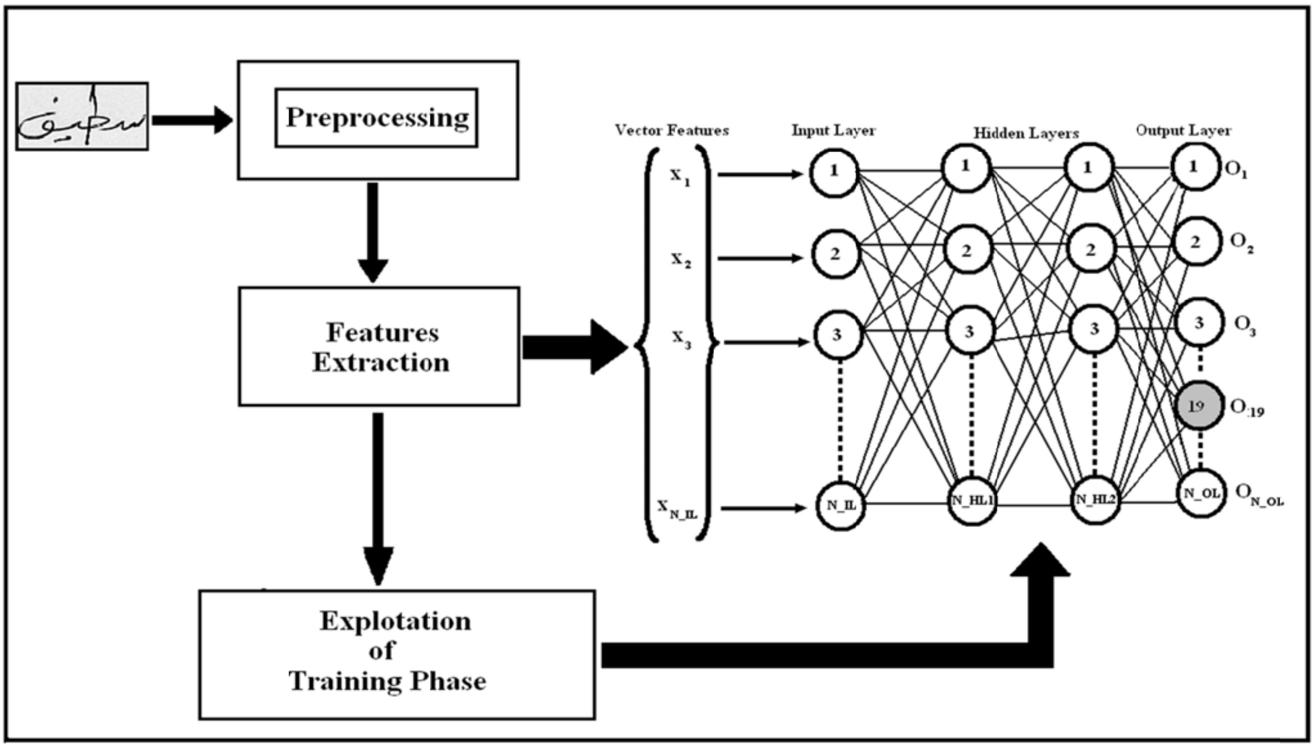

Fig. 1 General schema of our Arabic handwritten words recognition system

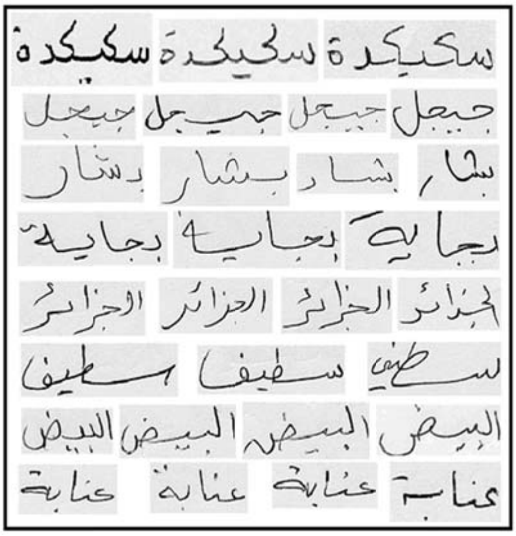

Fig. 2 Some samples of the used database

2. Preprocessing

The preprocessing operations are classical operations in image processing, their objective is to clean and prepare the image for the other steps of the OCR system. The preprocessing attempts to eliminate some variability related to the writing process and that are not very significant under the point of view of the recognition, such as the variability due to the writing environment, writing style, acquisition and digitizing of image. For our case, we used the following preprocessing operations (Fig. 3):

- The binarization: this operation consists at returning the word image to a binary image (black for the bottom and white for the object)
- Normalization of the word image: knowing that the words images have variable sizes, this operation consists at normalizing the image size at 64*192 pixels.
- Dilation: It is the operation that consists in dilating the tracing of the word

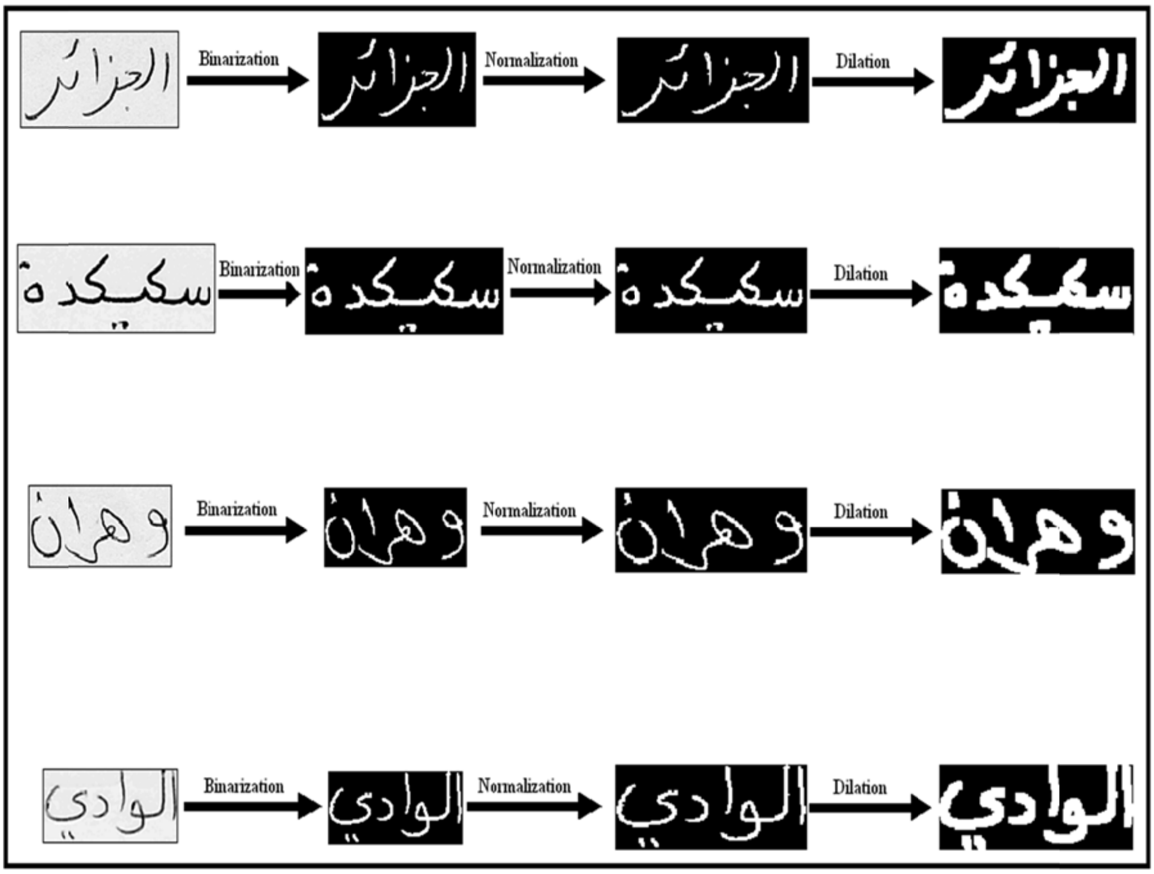

Fig. 3 Example of the different preprocessing operations used for our system

B. Features Extraction

Features extraction is an important step in achieving good performance of OCR systems. However, the other steps also need to be optimized to obtain the best possible performance, and these steps are not independent. The choice of features extraction method limits or dictates the nature and output of the preprocessing step and the decision to use gray-scale versus binary image, filled representation or contour, thinned skeletons versus full-stroke images depends on the nature of the features to be extracted. Features extraction has been a topic of intensive research and we can find a large number of features extraction methods in the literature [16], [20], but the real problem for a given application, is not only to find different features extraction methods but which features extraction method is the best?. This question led us to characterize the available features extraction methods, so that the most promising methods could be sorted out. In this paper, we are especially interested in the binary image of the word, and the methods used to extract the discrimination features of are the following:

1. The Parameters of the Distribution

In this case, we start with dividing the word image in six segments of size 64*32 pixel and every segment will be divided at a determined number of zones, the distribution sequence characterizes a number of the object pixels in relation to the total pixels number in a given zone.

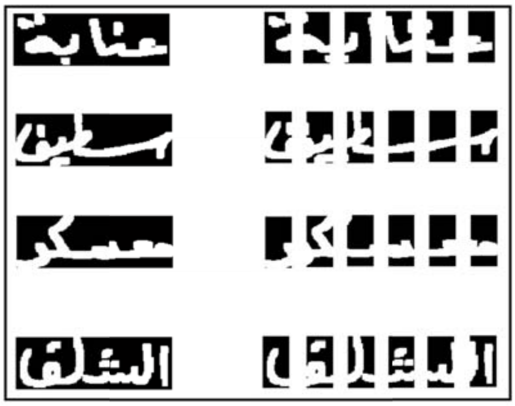

Fig. 4 Example of the segmentation in six segments of some Arabic handwritten words of some Algerian departments

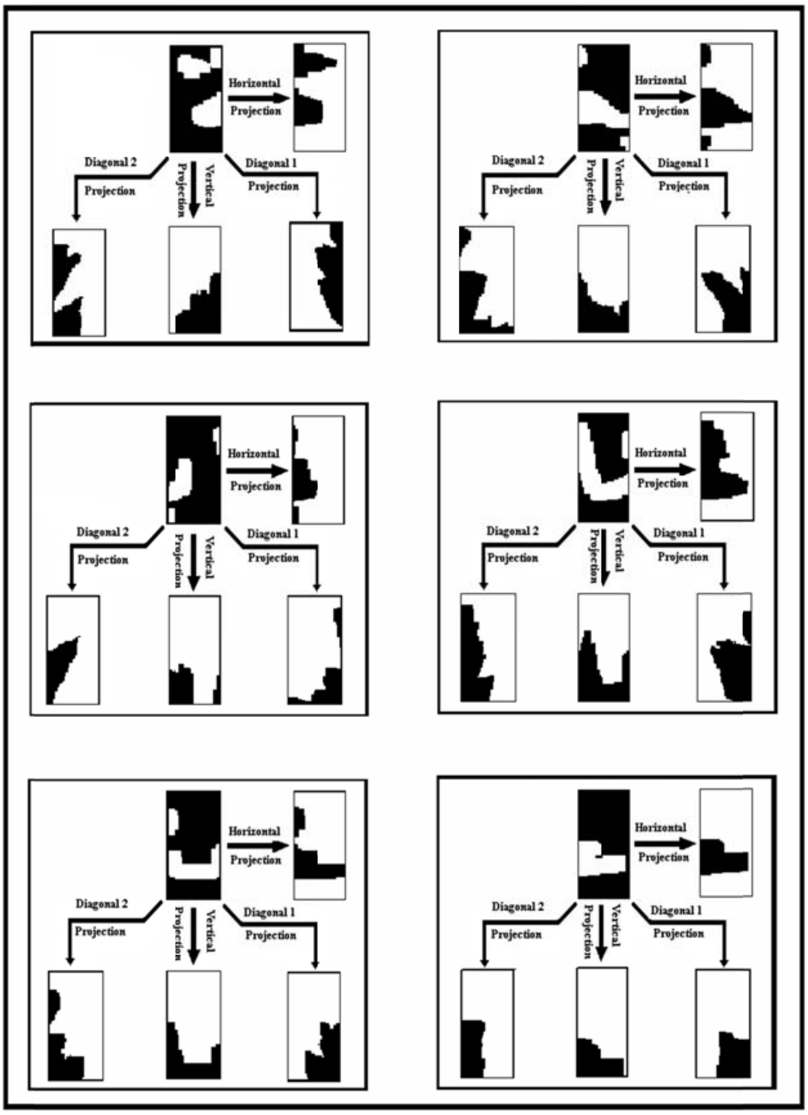

Fig. 5 The different projections of the different segments of the Arabic handwritten word "عنابة"

For our application, every segment is divided in 32 zones and every zone is of size 8*8 pixels (Fig. 4), and the values of the distribution sequence are defined by:

$$x_i = \frac{N_i}{N} \quad (1)$$

with, x_i : is the ith value of the distribution sequence, N_i :is a number of the object pixels in the ith zone., N: is a total pixels number in the ith zone.

2. The Centered Moments of the Different Projections of the Different Segments

The projection of an image in a given direction is the number of objects pixels in the direction in question, from that point of view; we can define the following projections:

- The horizontal projection: For a given line, the value of the projection is equal to the number of the object pixels in this line.
- The vertical projection: For a given column, the value of the projection is equal to the number of the object pixels in this column
- The projections according to the two diagonals: For a given diagonal, the value of the projection is equal to the number of the object pixels according to the direction in question.

Let us note that for our application, the different projections are also calculated on segments gotten after the division of the word image in six segments, (every segment is of size 64*32). Fig. 5 shows the different projections of the different segments of the Arabic handwritten word **"عنابة".**

The discrimination parameters are the centered moments of the different projections of the segments gotten by:

$$u_k = \sum_{i=1}^{M} (x_i - \bar{x})^k . p(x_i) \quad (2)$$

$$\bar{x} = \sum_{i=1}^{M} x_i \, p(x_i) \quad (3)$$

with: $\bar{x}$ is the mean value of the distribution sequence. $p(x_i)$ is the probability of l 'element x_i in the distribution sequence M is the size of the projection sequence. k is the order of the moment.

For our application, we choose the first six moments for every projection.

3. The Centered Moments of Image Coding According to the Directions of Freeman

This method consist to divide the word image into four zones of equal size (size of each zone is 32 * 96 pixel) and to encode each zone according to the eight directions of Freeman (Fig. 6). In other words, for a given zone, each parameter is the accumulated object pixel in one of the directions of Freeman. So each zone will be encoded by eight parameters, and the word image will be encoded by a sequence of thirty-two parameters. The centered moments of the sequence of Freeman are obtained by (2), and (3).

For our application, we chose the first six moments.

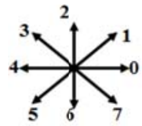

Fig. 6 The eight Freeman directions

4. Barr Features

The Barr features have been used with success in several works [21], [22], they are calculated on the binary images. Four images parameters are generated, and every image parameter corresponds to one of the following directions: east (e), North (n), Northeast (ne), Northwest (nw). Every image parameter has a whole value representing the Barr length in the direction in question. The features are calculated from the images parameters using zones that overlap to assure a certain degree of smoothing. Fifteen rectangular zones are arranged in five lines with three zones for every line; every zone is of size [(h/3)*(w/2)] where h and w are respectively the height and the width of the image. The high corners on the left of the zones are at the positions {(r0, c0): r0=0, h/6, 2h/6, 3h/6, 4h/6 and c0=0, w/4, 2w/4}. The values in every zone of the parameters images are added and the sums are normalized, and the dimension of the features vector is 15* 4=60. If we suppose f_1, f_2, f_3, f_4 are the images parameters associated at a shape in entry and Zi (i=1,2… .15) is an rectangular zone of size [(h/3)*(w/2)] with the top corner on the left is (r_0,c_0), the value P_{ik} of the parameter associated to the Zi zone for the image parameter f_k (k=1,2,3,4) is given like follows:

$$P_{ik} = \frac{1}{N} \sum_{r=r_0}^{r_0+\frac{w}{2}} \sum_{c=c0}^{c0+\frac{h}{3}} f_k(r,c) \quad (4)$$

In this work, this method is applied in two different manners:

a) Barr Features Applied on the Segmented Image of the Word

In this case, we start by dividing the word image in six segments of size 64*32 pixel (Fig. 7), and for each segment, we apply the cited previously method, therefore, the application of the Barr Features method on the segmented image will allow us to obtain 60 * 6 = 360 parameters.

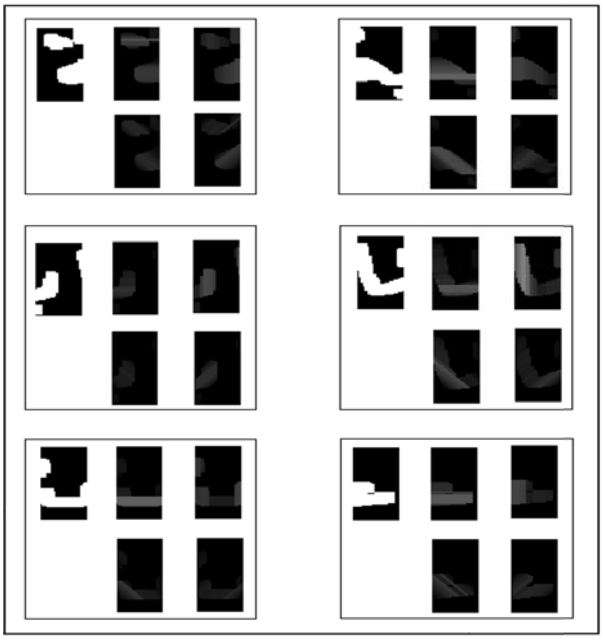

Fig. 7 The different images parameters of the different segments of the Arabic handwritten word “عنابة”

b) Barr Features Applied on the Whole Image of the Word

In this case, the Barr Features method is applied overall image of the word (Fig. 8), and this allows us to obtain sixty parameters.

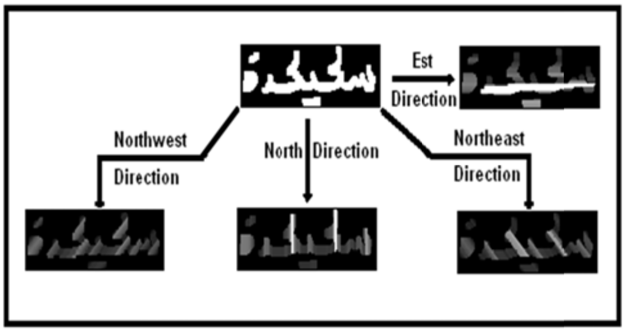

Fig. 8 The Arabic handwritten word "سكيكدة"and its four images parameters

C. Word Recognition

The handwritten word recognition is a problem for which a recognition model must necessarily take in account an important number of variability, dice at the time, the recognition techniques based on the neural networks can bring certain suppleness for the construction of such models. For our system, we opted for an MLP (Multi-Layers Perceptron) which is the most widely studied and used neural network classier. Moreover, MLPs are efficient tools for learning large databases. The used MLP in our work is trained with the back propagation with momentum training algorithm. The transfer function employed is the familiar sigmoid function.

1. Used Features Vector

It is the features vector used to characterize the word image, and with which, we will nourish the recognition module. For our case, the vector used is formed by: one hundred ninety-two (192) parameters of distribution method, hundred forty four (144) parameters of the centered moments of different projections (the first six values of the centered moments of the four projections of the six segments), the six (6) parameters of the centered moments of image coding according to the Freeman directions, the three hundred sixty (360) parameters of the Barr features method applied on the segmented word image and the sixty (60) parameters of the Barr features method applied on the whole image word.

2. The Input Data

The database consists of 480000 binary images. These images represent all classes possible of the Arabic handwritten word of the Algerian departments with variable sizes and variable thickness, and with 10000 samples for every class. This database is divided to two sets, 70% for training the neural network and 30 % for testing it.

3. Neural Network Parameters

The input layer nodes number is equal to the size of the used features vector (N_IL=762) the output layer nodes number is equal to the classes number to recognize (N_OL=48), for the hidden layers, we used a double hidden layer with 360 nodes for the first hidden layer and 157 for the second. The number of the hidden nodes is fixed by groping (N_HL1=360, N_HL2=157). The initial connection weights are in the range [-1, 1].

4. The Training Process

For training the neural network, back propagation with momentum training method is followed. This method was selected because of its simplicity and because it has been previously used on a number of pattern recognition problems. The method works on the principle of gradient descent. The algorithm uses two parameters, which are experimentally set, the learning rate and momentum. These parameters allow the algorithm to converge more easily if they are properly set by the experimenter. For our case, we have opted for the following values: η=0.35 and μ=0.9. The training process for the network is stopped only when the sum of squared error falls below 0.001.

III. The Experimental Results

The neural network performances are measured on the entire database (training or learning set and testing set). During this phase, we present the word image to recognize to the system entry, and we collect at the exit its affectation to one of the possible classes.

The results can be:

- Recognized word: the system arrives to associate one and only one prototype to the digit to recognize.
- Ambiguous word: the system proposes several prototypes to the digit to recognize.
- Rejected word: the system does not take any decision of classification.

- Non-recognized word: the system arrives to take a decision for the presented digit, but it is not the good decision.

The results and the different rates are regrouped in Table I:

TABLE I
RESULTS AND DIFFERENT RATES

Symbol	Quantity
R_R(%)	97.325
A_R(%)	0.375
J_R(%)	1.215
NR_R(%)	1.085

- R-R: Recognizer rate
- A-R: ambiguity rate
- J-R: Reject rate
- NR-R: No Recognizer rate

IV. Conclusion and Perspectives

The recognition of the Arabic handwritten words is a problem for which a model of recognition must necessarily take in account an important number of variables and constraints due at the variation of the word shape of the same class (variation of the writing styles, use of different writing instruments, variation of writing of a writer to another.. etc). In our work, we presented an off line system for the recognition of the Arabic handwritten words of the names of Algerian department. This work is based principally on the evaluation of neural network performances, trained with the gradient back propagation algorithm. The used parameters to form the input vector of the neural network are extracted on the binary images of the handwritten word by several methods: the Distribution parameters, the centered moments of the different projections of the different segments, the centered moments of the word image coding according to the directions of Freeman, and the Barr features applied on the whole binary image of the word and on its different segments. The classification is achieved by a multi layers perceptron.

References

[1] J. Mantas, "An Overview of Character Recognition Methodologies", Pattern Recognition, Vol.19, No.6, 1986, pp.425–430.
[2] S. Mori, C. Y. Suen and K. Yamamoto, "Historical Review of OCR Research and Development", Proceedings of the IEEE, Vol. 80, No. 7, 1992, pp. 1029-1058.
[3] A. L. Koerich, R. Sabourin, C. Y. Suen & A. El-Yacoubi, "A Syntax Directed Level Building Algorithm for Large Vocabulary Handwritten Word Recognition", In 4th International Workshop on Document Analysis Systems (DAS 2000) , Rio de Janeiro, Brazil, December 2000.
[4] L. S. Oliveira, R. Sabourin, F. Bortolozzi and C. Y. Suen, "A Modular System to Recognize Numerical Amounts on Brazilian Bank checks", 6th International Conference on Document Analysis and Recognition (ICDAR 2001), Seattle-USA, IEEE Computer Society Press, September 10-13, 2001. pp 389-394.
[5] S. Ouchtati, M. Bedda, F. Bouchareb and A. Lachouri "An Off Line System for the Handwritten Numeric Chains Recognition", International Journal of Soft Computing (IJSC), ISSN: 1682-9503, Vol. 1, No. 4, 2006, pp. 279-287.
[6] A. Filatov, N. Nikitin, A. Volgunin, and P. Zelinsky, "The Address Script TM Recognition System for Handwritten Envelopes", In International Association for Pattern Recognition Workshop on Document Analysis Systems (DAS'98), Nagano, Japan, November 4-6 1998, pp. 157–171.
[7] A. El-Yacoubi, "Modélisation Markovienne de L'´écriture Manuscrite Application `a la Reconnaissance des Adresses Postales", PhD thesis, Université de Rennes 1, Rennes, France, 1996.
[8] J. Hu, M. K. Brown and W Turin, "HMM based on-line handwriting recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, October . 1996, pp. 1039-1045.
[9] G. Kim and V. Govindaraju, "A lexicon driven approach to handwritten word recognition for real-time applications", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, April 1997, pp. 366-379.
[10] R. Buse, Z-Q Liu, and T. Caelli, "A structural and Relational Approach to Handwritten Word Recognition", IEEE Trans. Systems, Man and Cybernetics, Part-B, Vol. 27, October. 1997, pp. 847-861.
[11] K. Liu, Y. S. Huang and C. Y. Suen, "Identification of fork Points on the Skeletons of Handwritten Chinese Characters," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, October. 1999, pp. 1095-1100.
[12] A. Amin, "Off-Line Arabic Character Recognition – The State of the Art (Review)", Pattern Recognition, vol. 31, No. 5, 1998, pp. 517-530.
[13] M. Redjimi, S. Ouchtati and M. Bedda, "A New off Line System for the Recognition of the Isolated Handwritten Arabic Characters", Asian Journal of Information Technology (AJIT), ISSN: 1682-3915, Vol. 5, No. 8, 2006, pp. 912-918.
[14] S. Ouchtati, L. Bennacer and M. Bedda, "An Off Line System for the Arabic Handwritten Words Recognition ", International Review on Computers and Software (I.RE.CO.S.), ISSN: 1828-6003, Vol. 3, No. 6, 2008, pp. 579-585.
[15] J. Cai and Z-Q Liu, "Integration of Structural and Statistical Information for Unconstrained Handwritten Numeral Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, March 1999, pp. 263-270.
[16] Oivid Due Trier, Anil K. Jain, & Torfinn Taxt, "Feature Extraction Methods for Character Recognition – A Survey", Pattern recognition, Vol. 29, No. 4, 1996, pp.641-662.
[17] M. Bedda, M. Ramdani et S. Ouchtati., "Sur le Choix d'une Représentation des Caractères Manuscrits Arabes", Proceeding du 2ème Conférence Internationale Signaux, Systèmes, et Automatique SSA2'99, Université de Blida, Algérie, 10-12 Mai 1999, pp73-84.
[18] F. Grandidier, "Un nouvel Algorithme de Sélection de Caractéristiques Application à la Lecture Automatique de L'écriture manuscrite", thèse de doctorat en génie PH.D, Ecole de Technologie Supérieure, Université du Quèbec Canada, Janvier 2003.
[19] N. Benahmed, "Optimisation des Réseaux de Neurones Pour la Reconnaissance des Chiffres Manuscrits Isolés, Sélection et Pondération des Primitives par Algorithmes Génétiques", Thèse pour l'obtention de la Maîtrise en Génie de la Production Automatisée, Montréal le 01 Mars 2002.
[20] Salim Ouchtati, Mohammed Redjimi and Mouldi Bedda, "A Set of Features Extraction Methods for the Recognition of the Isolated Handwritten Digits", International Journal of Computer and Communication Engineering (IJCCE), ISSN: 2010-3743, Vol. 3, No. 5, September 2014, pp. 349-355.
[21] S. Ouchtati, M. Redjimi and M. Bedda, "An Offline System for the Recognition of the Fragmented Handwritten Numeric Chains ", International Journal of Future Computer and Communication (IJFCC), ISSN: 2010-3751, Vol. 4, No. 1, February 2015, pp. 33-39.
[22] S. Ouchtati, M. Bedda, and A. Lachouri "Segmentation and Recognition of Handwritten Numeric Chains", Journal of Computer Science (JCS), ISSN: 1549-3636, Vol. 3, No. 4, 2007, pp. 242-248.