Asian Journal of Information Technology 5(8): 912-918, 2006

An off Line System for the Recognition of the Isolated Handwritten Arabic Characters

[1]Mohammed Redjimi, [2]Salim Ouchtati and [3]Mouldi Bedda
[1]Department of Computer Science, Skikda University, BP: 26 Route EL_HADAIK Skikda, Algeria
[2]Department of Electrotechnics, Skikda University, BP: 26 Route EL_HADAIK Skikda, Algeria
[3]Department of Electronics, Annaba University, Annaba University 23000 Annaba, Algeria

Abstract: In this study we present an off line system for the recognition of the isolated handwritten Arabic characters. The study is based on the analysis and the evaluation of multi-layers perceptron performances, trained with the gradient back propagation algorithm. It is hoped that the results of the evaluation contribute to the conception of operational systems. The used parameters to form the input vector of the neural network are extracted on the binary images of the characters by the following methods: the centerd moments of the projections sequences, distribution parameters, the Barr features and Coding according the directions of Freeman.

Key words: Optical characters recognition, neural networks, barr features, image processing, pattern recognition, features extraction

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. In this study, the recognition of off-line handwriting is more complex than the on-line case. Complexity due to the presence of noise in the image acquisition process and the loss of temporal information such as the writing sequence and the velocity. This information is very helpful in a recognition process. 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 until today consider constrained recognition problems based on vocabularies from specific domains, e.g. the recognition of handwritten check amounts[4] or postal addresses[5,6]. 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. As a result of intensive research and development efforts, systems are available for English language[7-9], Chinese language[10], Arabic language[11] and handwritten numerals[12]. There is still a significant performance gap between the human and the machine in recognizing unconstrained handwriting. This is a difficult research problem caused by huge variation in writing styles and the overlapping and the intersection of neighboring characters.

Corresponding Author: Mohammed Redjimi, Department of Computer Science, Skikda University, BP: 26 Route EL_HADAIK Skikda, Algeria

A RECOGNITION SYSTEM FOR THE ISOLATED HANDWRITTEN ARABIC CHARACTER

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

- Acquisition and preprocessing.
- Features extraction.
- Recognition.

Acquisition and preprocessing

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 (28 classes) of the isolated handwritten Arabic characters (أ, ب, ت, ث, ج, ح, خ, د, ذ, ر, ز, س, ش, ص, ض, ط, ظ, ع, غ, ف, ق, ك, ل, م, ن, ه, و, ي) with variable sizes and variable thickness and with 100 samples for every class. Let's note that the characters images of our database are formed only by two gray levels: the black for the object and the white for the bottom. The Fig. 2 shows some samples of the used database.

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

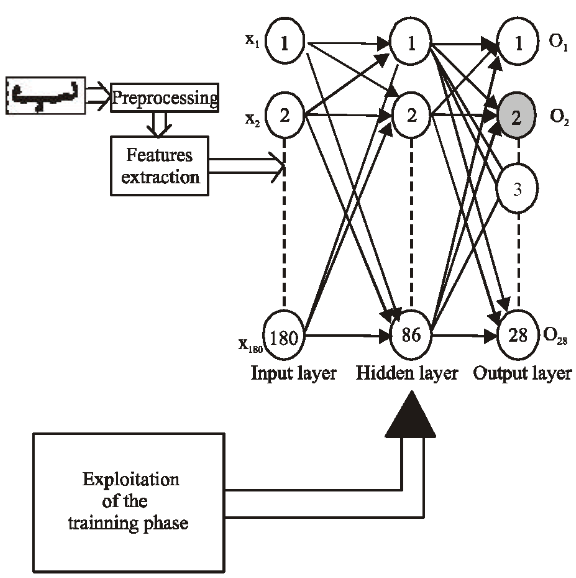

Fig. 1: General schema of our system for the recognition of the isolated handwritten arabic character.

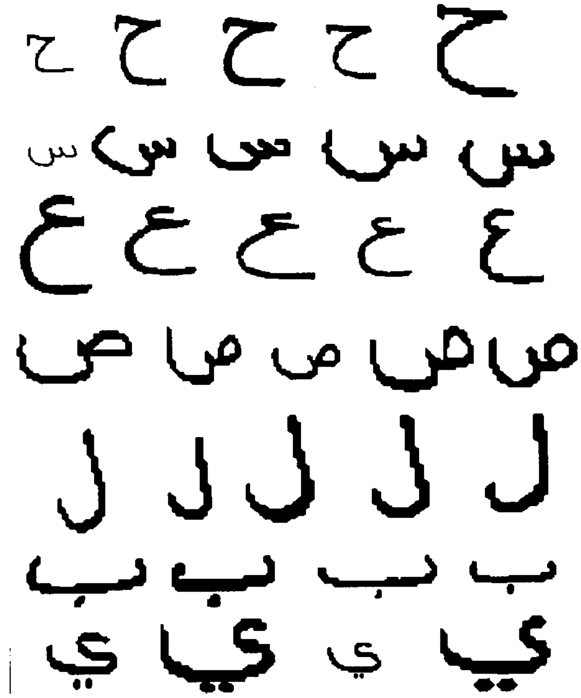

Fig. 2: Some samples of the used database

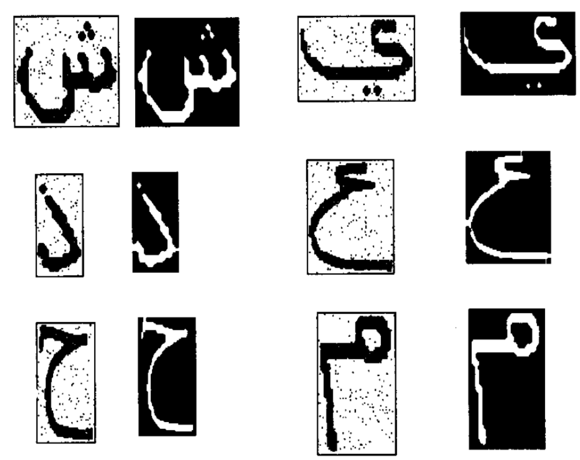

Fig. 3: Filtering and inversion of the gray levels of some handwritten digits

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:

Filtering and inversion of the gray levels: This operation consists in eliminating the noises in the binary image due to different reasons (bad Acquisition conditions , bad writing conditions, the writer's mood.. etc.), in our case, some digits are marked by the noise of type studies and

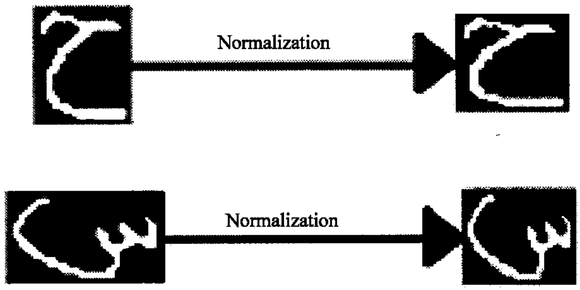

Fig. 4: Normalization of some handwritten digits

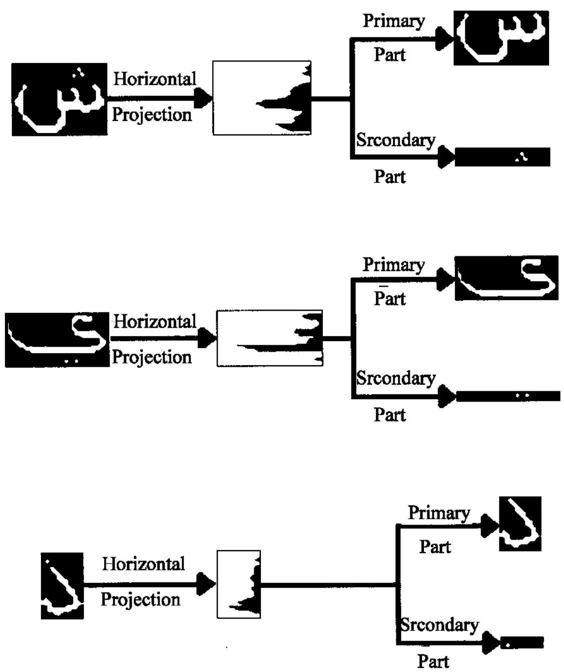

Fig. 5: Detection and isolation the secondary part

salt, the application of the filter median on the digit image permitted us to eliminate easily this type of noise. Let's note that for reasons of calculation we reversed the gray levels of the character image (black for the bottom and white for the object). The Fig. 3 shows us the filtering and inversion operation of the gray levels of some handwritten digits.

Normalization of the digit image: Knowing that the digits images have variable sizes, this operation consists at normalizing the image size at 64 * 64 pixels (Fig. 4).

Use of the horizontal projection for the detection and the isolation of the secondary part that will be recognized in an ulterior phase (Fig. 5).

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

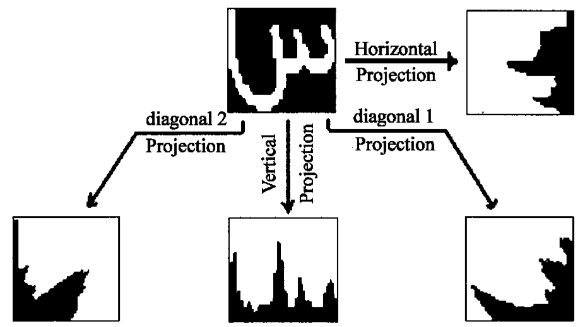

Fig. 6: The four projections of the character SIN

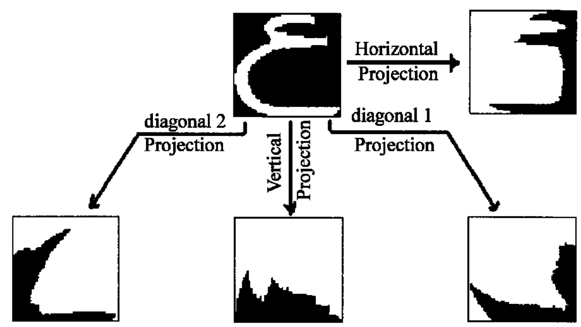

Fig. 7: The four Projections of the character AIN

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, 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 study, we are especially interested in the binary image of the digits and the methods used to extract the discrimination features are applied on the image of the primary part of the character (the secondary part is recognized in an ulterior phase). These methods are the following:

The centred moments of the projections sequences: The projection of an image in a given direction is the number of objects pixels in the direction in question, in this case the parameters of discrimination are the centred moments of the following projections: vertical, horizontal and according the two diagonals (Fig. 6 and 7).

The centred moments of k order for every sequence of projection are given by:

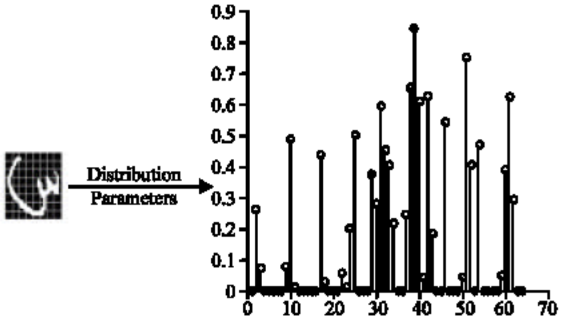

Fig. 8: The digits eight and five and their distribution sequences

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

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

- $\bar{x}$ is the mean value of the sequence of projection.
- $p(x_i)$: is the probability of 1 'element in the sequence of projection.
- M: is the size of the projection sequence.

The parameters of the distribution: While dividing the character image 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. For our application, the digit image is divided in 64 zones (Fig. 8) and the values of the distribution sequence are defined by:

$$R_i = \frac{N_i}{N} \quad (3)$$

With:

- R_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.

Barr-features: The Barr features have been used with success in several works[13,14], they are calculated on the binary images of the characters. Firstly, 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 of the parameter associated to the Zi zone for the image parameter (k=1,2,3,4) is given like follows:

Fig. 9: The images parameters of character SIN

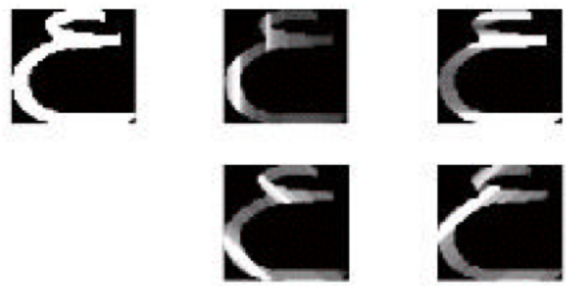

Fig. 10: The images parameters of character AIN

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

The Fig. 9 and 10 shows the images parameters of the characters SIN and AIN.

Coding according the directions of freeman: This method consists to dividing the image of the character in four zones (Z1, Z2, Z3 and Z4) and for every zone we calculate the number of object pixels in the directions of Freeman (Fig. 11). Therefore each zone will be coded by eight parameters and the image of the character will be coded by thirty two parameters.

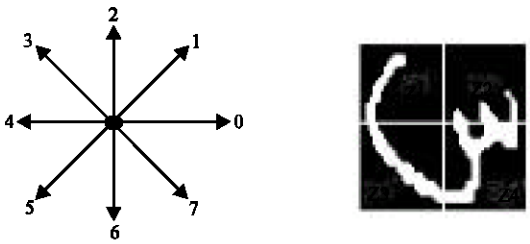

Fig. 11: Principle of the coding of freeman

Used features vector: It is the features vector used to characterize the image of character and with which, we will nourish the recognition module. For every character image this vector is constituted of the first six values of the centred moments for every projection sequence (that is to say twenty four parameters) more the sixty four parameters of distribution more sixty parameters of the Barr features more the thirty two parametres of the coding according to the Freeman directions.

Character recognition

Recognition of the secondary part: The secondary associated to the Arabic characters that we took in consideration are ('. ', '.. ', ':. '). Our system starts with the detection of the presence of the secondary part, to isolate this part if it exists, to detect its position in relation to the primary part (underneath or in over) and finally to recognize it by using the number of pixels.

Recognition of the primary part: The handwritten character recognition is a problem for which a recognition model must necessarily take in account an important number of variabilities, 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.

The input data: The database consists of 2800 binary images. These images represent all classes possible of the isolated arabic character with variable sizes and variable thickness and with 100 samples for every class. This database is divided to two sets, 70% for training the neural network and 30% for testing it.

Neural network parameters: The input layer nodes number is equal to the size of the used features vector

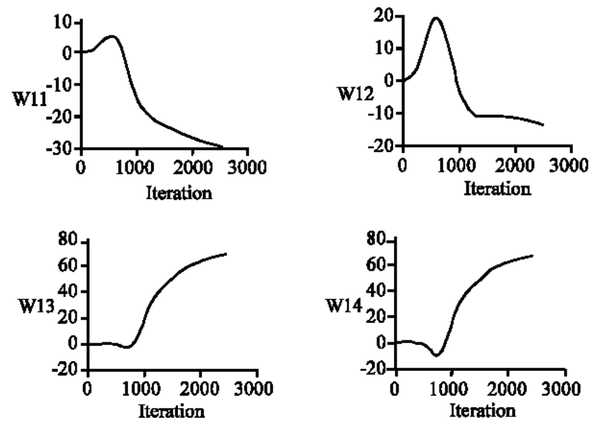

Fig. 12: The variation of the four first values of the connection weights between the hidden layer and the input layer

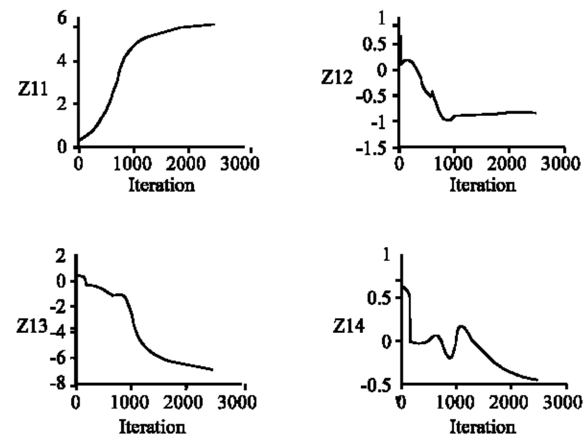

Fig. 13: The variation of the four first values of the connection weights between the hidden layer and the output layer

(N_IL=180), the output layer nodes number is equal to the classes number to recognize (N_OL=28), for the hidden layers, we used a single hidden layer with 86 nodes fixed by groping (N_HL=86). The initial connection weights are in the range [-1, 1].

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 η and momentum μ. 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=0.45 and μ=0.8. During the learning phase the neural network learns by example and the connection weights are updated in an iterative manner (Fig. 12 and 13). The training process for the network is stopped only when the sum of squared error falls below 0.001.

RESULTS

The neural network performances are measured on the entire database (training or learning set and testing set). During this phase, we present the digit 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:

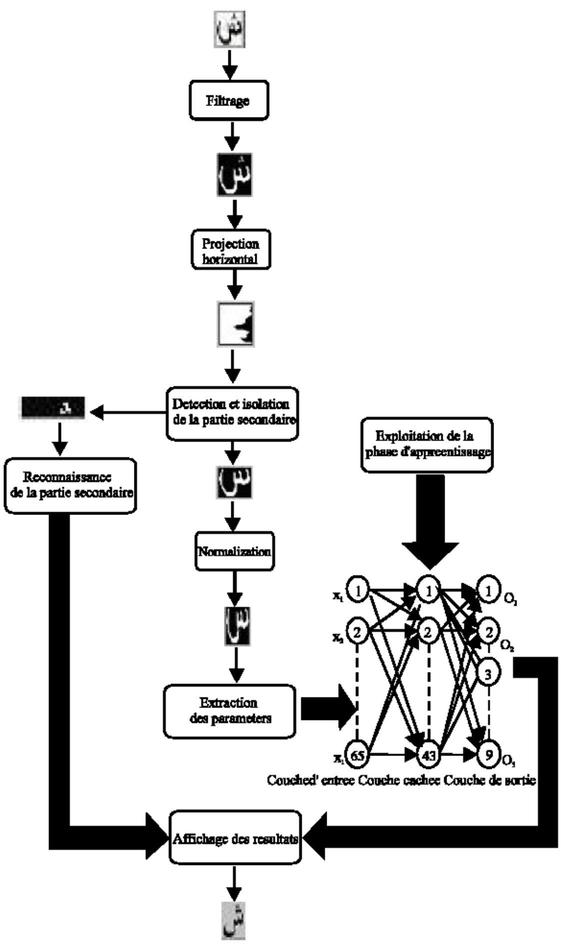

Fig. 14: Detailed schema of the developed system

- Recognized character: the system arrives to associate one and only one prototype to the digit to recognize. Ambiguous character: the system proposes several prototypes to the digit to recognize.
- Rejected character: the system doesn't take any decision of classification.

Table 1: Results and different rates

	R-R (%)	A-R (%)	J-R (%)	NR-R (%)
Training set	98.215	0.535	0.785	0.464
Testing set	95.178	1.107	0.428	3.285

- Non recognized character: 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 the Table 1:

With:

- R-R: Recognizer rate.
- A-R: ambiguity rate.
- J-R: Reject rate.
- NR-R: Non recognizer rate.

A detailed schema of the developed system is given by Fig. 14.

CONCLUSION

The recognition of the isolated handwritten Arabic characters is a problem for which a model of recognition must necessarily take in account an important number of variabilities and constraints due at the variation of the character shape of the same class (variation of the writing styles, use of different writing instruments, variation of writing of a writer to another.. etc) and to the problems of the detection and isolation of the secondary part. In our work, we presented an off line system for the recognition of the isolated handwritten Arabic characters 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 digits by the following methods: the centerd moments of the projections sequences, distribution parameters, the Barr features and Coding according the directions of Freeman. The gotten results are very encouraging and promoters; however we foresee the following evolution possibilities:

- To widen the database by taking in account a bigger number of writers and writing instruments.
- To consider other classification methods.
- Use of the algorithms capable to control the ambiguity, reject and non recognizer rates by adjusting the reject and ambiguity rates by use of suitable doorsteps.
- Use of other features extraction methods.

- Use of the post-processing techniques to improve the system performances.

REFERENCES

1. Mantas, J., 1986. An overview of character recognition methodologies. Pattern Recognition, 19: 425-430.
2. Mori, S., C.Y. Suen and K. Yamamoto, 1992. Historical review of OCR research and development. proceedings of the IEEE, 80: 1029-1058.
3. Koerich, A.L., R. Sabourin, C.Y. Suen and A. El-Yacoubi, 2000. A Syntax Directed Level Building Algorithm for Large Vocabulary Handwritten Word Recognition. In 4th Inte. Workshop on Document Analysis Systems (DAS) , Rio de Janeiro, Brazil.
4. Oliveira, L.S., R. Sabourin, F. Bortolozzi and C.Y. Suen, 2001. 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, pp: 389-394.
5. Filatov, A., N. Nikitin, A. Volgunin and P. Zelinsky, 1998. The Address Script TM recognition system for handwritten envelopes. In International association for pattern recognition workshop on document analysis systems (DAS'98), Nagano, Japan, pp: 157-171.
6. El-Yacoubi, A., 1996. Modélisation Markovienne de L'´écriture manuscrite application `a la reconnaissance des adresses postales. PhD thesis, Université de Rennes 1, Rennes, France.
7. Hu, J., M.K. Brown and W. Turin, 1996. HMM based on-line handwriting recognition. IEEE Transactions on Pattern analysis and machine intelligence, 18: 1039-1045.
8. Kim, G. and V. Govindaraju, 1997. A lexicon driven approach to handwritten word recognition for real-time applications", IEEE Transactions on pattern analysis and machine intelligence, 19: 366-379.
9. Buse, R., Z.Q. Liu and T. Caelli, 1997. A structural and relational approach to handwritten word recognition. IEEE Trans. Systems, Man and cybernetics, part-b, 27: 847-861.
10. Liu, K., Y.S. Huang and C.Y. Suen, 1999. Identification of fork points on the skeletons of handwritten Chinese characters. IEEE Transactions on pattern analysis and machine intelligence, 21: 1095-1100.
11. Amin, A., 1998. Off-line Arabic character recognition-the state of the art [review]. Pattern recognition, 31: 517-530.
12. Cai, J. and Z.Q. Liu, 1999. Integration of structural and statistical information for unconstrained handwritten numeral recognition. IEEE Transactions on pattern analysis and machine intelligence, 21: 263-270.
13. Ouchtati, S., M. Ramdani and M. Bedda, 2002. Un Réseau de Neurones Multicouches Pour la Reconnaissance Hors-Ligne des Caractères Manuscrits Arabes. Revue Sciences et technologie université de constantine, 17: 99-105.
14. Ouchtati, S., M. Redjimi, M. Bedda and F. Bouchareb, 2006. A New off Line System for Handwritten Digits Recognition. Asian J. Inform. Tech., (AJIT), 5: 620-626.