Recognition of the Arabic Handwritten Words of the Algerian Departments

Salim Ouchtati, Mohammed Redjimi, and Mouldi Bedda

***Abstract*—In this work we present 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 parameters of distribution, the moments centered of the different projections and the Barr features. Let's note that these methods are applied on segments gotten after the division of the binary image of the word in six segments. The classification is achieved by a multi layers perceptron. Detailed experiments are carried and satisfactory recognition results are reported.**

***Index Terms*—Optical characters recognition, neural networks, barr features, 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 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], [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. As a result of intensive research and development efforts, systems are available for English language [8]-[10], Chinese language [11], Arabic language [12], [13] and handwritten numerals [14]. 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. 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, omnifont machine printed text, and unconstrained handwritten text. A wide variety of techniques are used to perform handwriting recognition.

Manuscript received September 2, 2013; revised October 20, 2013. This work was supported in part by the Skikda Electronic Laboratory, skikda university, Algeria and the Electrical Engineering Department El Jouf University, Arabie Saoudite.

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: ouchtatisalim@ yahoo.fr).

Mohammed Redjimi is with the Computer Science Department, Faculty of Science, Skikda University, Algeria (e-mail: redjimimed@yahoo.fr).

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

II. The Different Parts of the Regonition System

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 the Fig. 1), this system is divided in three phases:

- Acquisition and preprocessing
- Feature 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 (**أدرار، الشلف، الأغواط، أم البواقي، باتنة، بجاية، بسكرة، بشار، البليدة، البويرة، تمنراست، تبسة، تلمسان، تيارت، تيزي وزو، الجزائر، الجلفة، جيجل، سطيف، سعيدة، سكيكدة، سيدي بلعباس، عنابة، قالمة، قسنطينة، المدية، مستغانم، المسيلة، معسكر، ورقلة، وهران، البيض، اليزي، برج بوعريريج، بومرداس، الطارف، تندوف، تسيمسيلت، الوادي، خنشلة، سوق أهراس، تيبازة، ميلة، عين الدفلى، النعامة، عين تموشنت، غرداية، غيليزان)** with variable sizes and variable thickness, and with 10000 samples for every class. The Fig. 2 shows some samples of the used database.

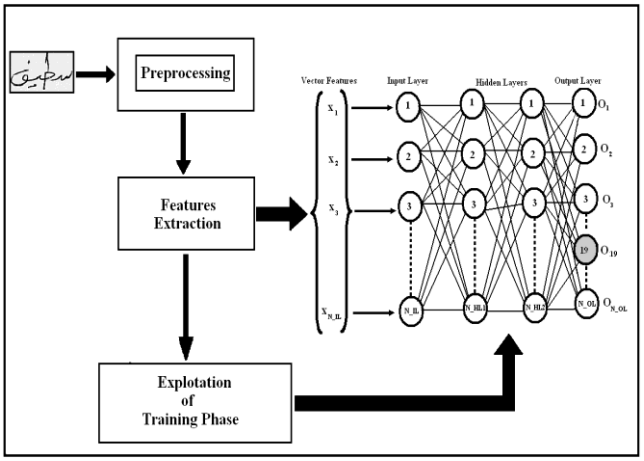

Fig. 1. General schema of our Arabic handwritten words recognition system.

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 (see Fig. 3):

- The binarisation: 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

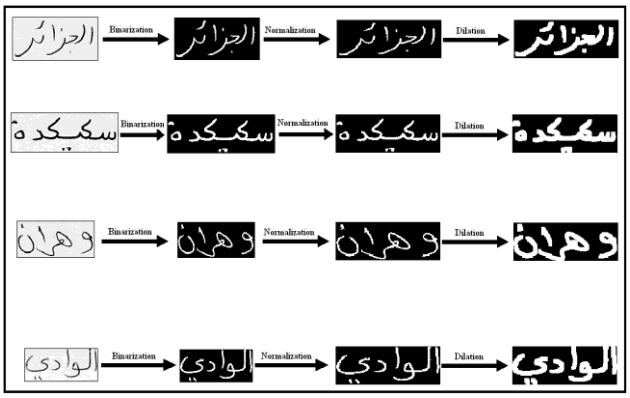

Fig. 3. Example of the different preprocessing operations used for our system.

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 [15], 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 [16]-[18]. 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 Distribution sequence

In this case we start with dividing the word image in six segments of size 64×32 pixel (see Fig. 4) 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.

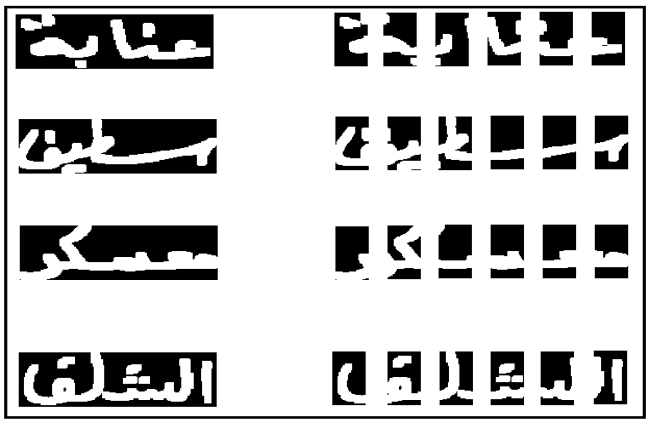

Fig. 4. : Example of the segmentation in six segments of some Arabic handwritten words of some Algerian departments.

For our application, every segment is divided in 32 zones and every zone is of size 8×8 pixel, and the values of the distribution sequence (see Fig. 5) are defined by:

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

With*:*

- x_i *:* is the *i*th value of the distribution sequence.
- N_i :is a number of the object pixels in the *i*th zone.
- N: is a total pixels number in the *i*th zone.

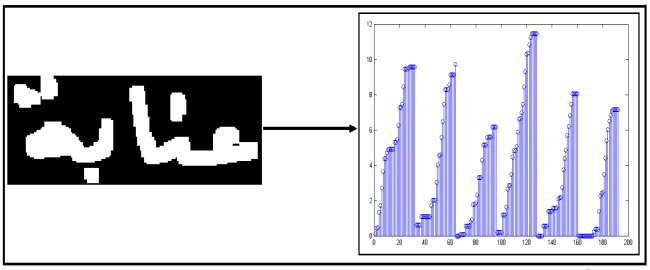

Fig. 5. The distribution sequence of the Arabic handwritten word "عنابة".

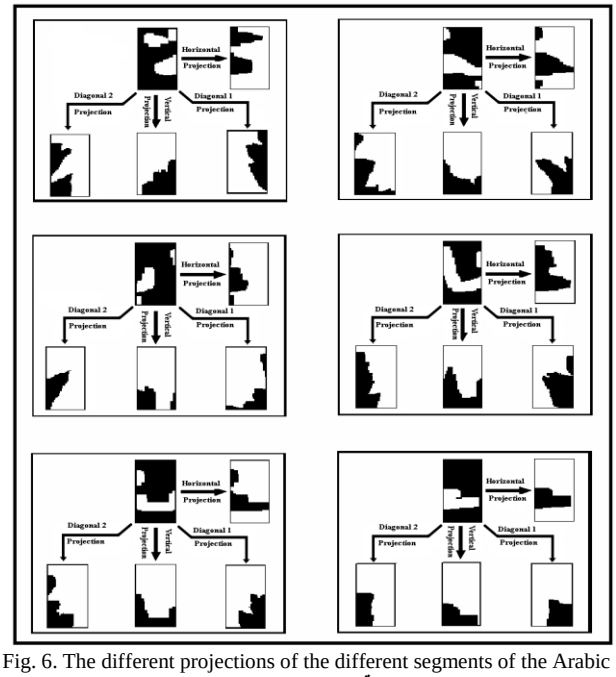

Fig. 6. The different projections of the different segments of the Arabic handwritten word "عنابة".

2) The centerd 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's 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). The Fig. 6 shows the different projections of the different segments of the Arabic handwritten word **"عنابة"**

The discrimination parameters are the centred moments of the different projections of the segments gotten by the following formulas (see Fig. 7)

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

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

With*:*

- x_i *:* is the *i*th value of the distribution sequence.
- N_i :is a number of the object pixels in the *i*th zone.
- N: is a total pixels number in the *i*th zone.
- k: the order of the moment

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

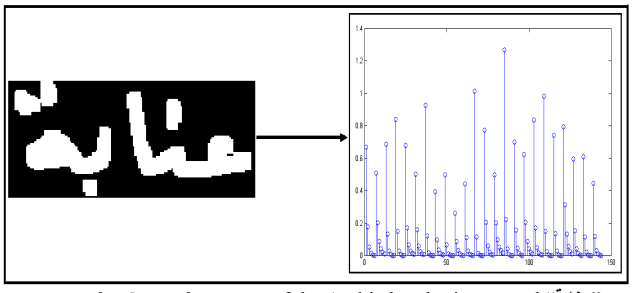

Fig. 7. The Centred moment of the Arabic handwritten word "عنابة".

3) Barr features

The Barr features have been used with success in several works [19], [20], they are calculated on the binary images. Firstly, we start with dividing the word image in six segments of size 64×32 pixel, and for every segment, four images parameters are generated (see Fig. 8), 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 $[(h/3)\times(w/2)]$ 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 for every segment is $15\times 4=60$, and the dimension of the feature vector for the whole word image is $60\times 6=360$ (see Fig. 9). 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 $[(h/3)\times(w/2)]$ 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:

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

C. Word Recognition

The handwritten word 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.

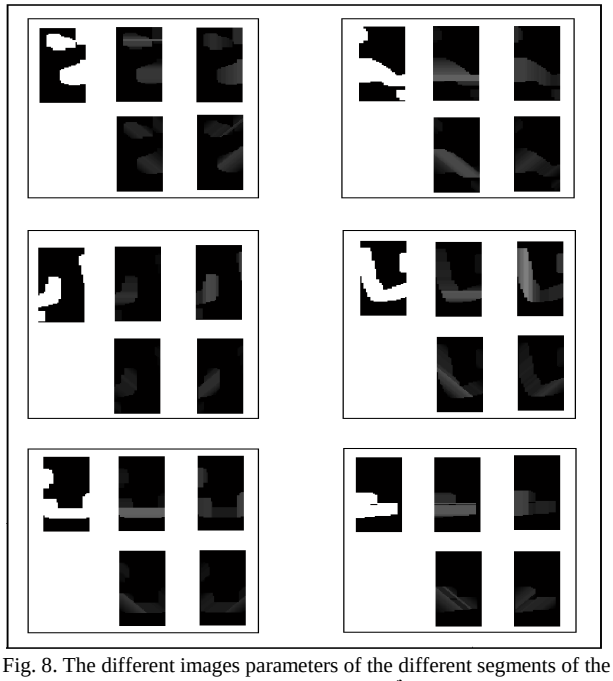

Fig. 8. The different images parameters of the different segments of the Arabic handwritten word "عنابة".

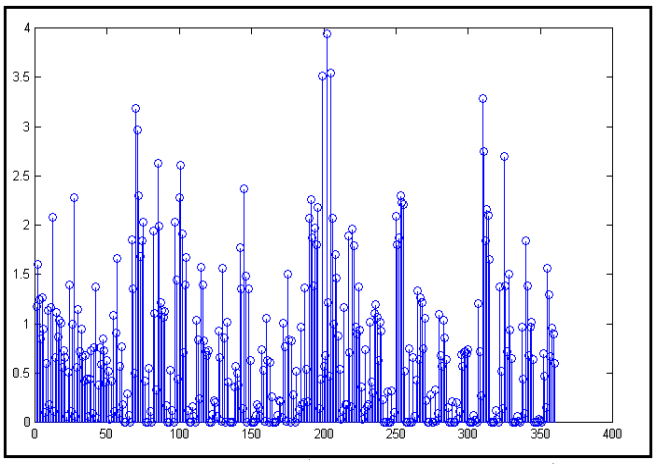

Fig. 9. The Barr features of the Arabic handwritten word "عنابة".

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 the hundred four twenty twelve parameters of distribution (192), hundred forty four parameters (144) of the moments of projections (the first six values of the centred moments of the four projections of the six segments) and the three hundred sixty (360) Barr-parameters and

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 department 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=696) 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 320 nodes for the first hidden layer and 157 for the second. The number of the hidden nodes is fixed by groping (N_HL1=320, 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 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=0.35 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 doesn't 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 the Table I:

TABLE I: RESULTS AND DIFFERENT RATES

R_R(%)	A_R(%)	J_R(%)	NR_R(%)
98.816	**0.255**	**0.482**	**0.445**

With:

- 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 Barr features applied on the whole binary image of the word, the parameters of distribution and the moments centered of the different projections. Let's note that these last two methods are applied on segments gotten after the division of the binary image of the word in six segments. The classification is achieved by an multi layers perceptron

The gotten results are very encouraging and promoters; however we foresee the following evolution possibilities:

- To widen the database by taking in account an important number of writers and writing instruments.
- To consider other classification methods.
- To use other algorithms capable to control the ambiguity, reject and non recognizer rates by adjusting the reject and ambiguity rates by use of suitable doorsteps.
- To use other features extraction methods.
- Use of the post-processing techniques to improve the system performances.

REFERENCES

[1] J. Mantas, "An overview of character recognition methodologies," *Pattern Recognition*, vol. 19, no. 6, pp. 425–430, 1986.

[2] S. Mori, C. Y. Suen, and K. Yamamoto, "Historical review of OCR research and development", in *Proc. the IEEE*, vol. 80, no. 7, 1992, pp. 1029-1058.

[3] A. L. Koerich, R. Sabourin, C. Y. Suen, and A. El-Yacoubi, "a syntax directed level building algorithm for large vocabulary handwritten word recognition," in *Proc. 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," in *Proc. 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)*, 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 *Proc. 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," Ph.D. 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, pp. 1039-1045, October 1996.

[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, pp. 366-379, April 1997.

[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, pp. 847-861, October 1997.

[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, pp. 1095-1100, October 1999.

[12] Amin, "Off-line Arabic character recognition – the state of the art [review]," *Pattern Recognition*, vol. 31, no. 5, pp. 517-530, 1998.

[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)*, vol. 5, no. 8, pp. 912-918, 2006.

[14] 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, pp. 263-270, March 1999.

[15] O. D. Trier, A. K. Jain, and T. Taxt, "Feature extraction methods for character recognition – a survey," *Pattern recognition*, vol. 29, no. 4, pp. 641-662, 1996.

[16] M. Bedda, M. Ramdani, and S. Ouchtati., "Sur le choix d'une représentation des caractères manuscrits arabes," in *Proc. du 2ème Conférence Internationale Signaux, Systèmes, et Automatique SSA2'99,* Université de Blida, Algérie, 10-12 Mai 1999, pp. 73-84.

[17] 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, école de technologie supérieure, université du Québec Canada, Janvier 2003.

[18] 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.

[19] S. Ouchtati, M. Bedda, and A. Lachouri "Segmentation and recognition of handwritten numeric chains," *Journal of Computer Science (JCS)*, vol. 3, no. 4, pp. 242-248, 2007.

[20] S. Ouchtati, M. Redjimi, M. Bedda, and F. Bouchareb, "A New off Line System for Handwritten Digits Recognition," *Asian Journal of Information Technology (AJIT)*, vol. 5, no. 6, 2006, pp. 620-626.

Salim Ouchtati was born on July 8, 1970 in Azzaba W Skikda Algeria. He received his B. Eng. and M.Sc. degrees in Electronic from Annaba University, Algeria in 1994 and 1999 respectively. He obtained his Doctorate diploma in 2007 in Automatic and his HDR diploma (Habilitation to Direct the Research) in Electronic in 2010 from Annaba University.

He is currently an associate professor at Skikda university, member of the Skikda Electronic Laboratory, responsible course of Industrial Control, responsible of the research project with the title: Selection and Weighting of Discrimination Parameters in a Recognizing Handwritten Digits System and member of the project: Realization of a System for Reading of the Algerians Postal Checks. He has a several scientific works published in many international journals such as: "Segmentation and Recognition of Handwritten Numeric Chains" (Journal of Computer Science), "A New off Line System for Handwritten Digits Recognition" (Asian Journal of Information Technology), "An off line System for the Handwritten Numeric Chains Recognition" (International Journal of Soft Computing). His researches area focused mainly on the Handwritten Recognition, Artificial Intelligence and Image Processing.

Dr. Salim Ouchtati was a member of the scientific committee of the Electrical Engineering Department in Skikda University from 2005-2010, member of the scientific committee of the Technology Faculty in the Skikda University from 2005-2008, he was a reviewer in several international conferences such as: First International Conference on 'Networked Digital Technologies(NDT 2009) Ostrava, The Czech Republic, July 29 - 31, 2009 and The Second International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2009), August 04-09 2009 London Metropolitain University, UK, member of the Scientific Committee in several international conference

Mohammed Redjimi was born on June 3, 1956 in Rejattas, Skikda, he received the Diploma of Docteur - ingenieur in computer science from the University of LILLE 1 – FRANCE (1984), and his HDR diploma (Habilitation to Direct the Research) in Computer Science in 2007 from Annaba University.

He is currently an associate professor of computer Science at Skikda university, responsible of the teamresearch "Modeling and simulation of complex processes" at Skikda Computer Science and Communication Laboratory, responsible of the research project with the title: "Cooperative approaches for images segmentation", responsible of the research project with the title: "Bayesian approach in computer vision", he has a several scientific works published in many international journals such as: "An Adaptative Multi-agent System Approach for Image Segmentation" (International Journal of Computer Applications), " An image encryption approach using stream ciphers based on nonlinear filter generator " (Theoretical and Applied Information Technology). His main research interests include modeling and simulation systems mainly by using multi-agents systems, image recognition and computing hardware and software systems.

Dr Mohammed Redjimi was a review and member of the Scientific Committee in several international conferences such as: First Conference on Theoretical and Applicative Aspects of Computer Science – Skikda University-, Fractional Order Systems and Applications (SOFA 2010) – Skikda University

Mouldi Bedda was born on October 3, 1956 in El_Oued Algeria. He received bachelor's degree from the university of Haouari Boumedienne Algers, Algeria in 1981, the M.Sc degree from the university of Languedoc Montpelier French in1982, and the Ph.D. degree in electrical engineering from the university Nancy 2, Nancy, French in 1985.

He was assistant professor from 1985- 1990, associate professor from 1990-2004 and professor from 2004-2006 at the university of Annaba Algeria, from 2006 to date full professor at the college of engineering at aljouf university KSA. He has a several scientific works published in many international.

Professor Mouldi Bedda was the director of Automatic and Signals Laboratory from 2001 -2006. His researches interests are: DSP, Speech Processing, OCR, Artificial Intelligence, and Biomedical.